UNIVERSITY OF CALIFORNIA

Los Angeles

Advancing Regional Seismic Impact Assessment using Building Strong Motion Data with Statistical and Causal Inference

A dissertation submitted in partial satisfaction

of the requirements for the degree Doctor of Philosophy in Civil Engineering

by

Eusef Abdelmalek-Lee

2024

ABSTRACT OF THE DISSERTATION

Advancing Regional Seismic Impact Assessment using Building Strong Motion Data with Statistical and Causal Inference

by

Eusef Abdelmalek-Lee
Doctor of Philosophy in Civil Engineering
University of California, Los Angeles, 2024
Professor Henry Burton, Chair

Strong motion building response data has been shown to be useful for supporting decision-making in the post-earthquake environment. However, most of the prior research in this area has focused on individual buildings. Moreover the few studies that have addressed regional impact assessment using strong motion data have only focused on structural response reconstruction. In other words, stakeholder-driven performance metrics such as repair costs from earthquake-induced damage have not been addressed. Another major challenge with prior research stems from the mostly low-to-moderate response demands that are present in the measurements from prior earthquakes. This research seeks to advance regional seismic impact assessment using strong motion data by addressing the aforementioned limitations and other gaps in the research. The first major issue that is addressed is the lack of a comprehensive and easily accessing database of building strong motion response recordings. To this end, a relational database of strong motion responses from buildings across California subjected to historical earthquakes is addressed. In addition to the response measurements, various characteristics of the buildings and seismic events are included in the database to support predictive modeling. As a compliment to the database, a python-based tool that enables ease of access and streamlined queries is also developed. Because all of the buildings in the constructed database are only partially instrumented, within a given structure, only

a subset of floors have response measurements. Moreover, within a given inventory, only a subset of the buildings will be instrumented. To deal this challenge, a structural response reconstruction (SRR) model is developed for use in a regional impact assessment context. This model takes in response measurements in partially instrumented buildings as well as structural and seismic event parameters and estimates the full set of responses for all the building in the inventory. A dual modeling approach is utilized, which initiates with using kriging to estimate the peak ground acceleration at the location of the uninstrumented buildings. Then, the extreme gradient boosting algorithm is used to reconstruct the EDPs across all buildings in the inventory. This study raised questions about which ground motion intensity measures (IMs) are most suitable for use as inputs in SRR models. To further probe this question, the strong motion data from the database was used to evaluate the effectiveness of several IMs. An important distinction from prior IM evaluation studies is the use of strong motion response data. In other words, most prior studies on assessing IM effectiveness used data from response history analysis.

Because the input data comprised mostly linear elastic (or near-elastic responses), the applicability of the SRR model was limited to use in low-to-moderate-sized earthquakes. To increase the generalizability of the SRR model, the dataset of measured EDPs was augmented with responses obtained from nonlinear response history analyses (NRHAs). Through this augmentation, the range of structural responses was increased, particularly in terms of nonlinear responses. To facilitate this data augmentation, an end-to-end framework was developed to automate the generation of simulated structural responses in an inventory that includes multiple lateral force resisting systems (LFRS). Steel moment frames, woodframe shear walls, reinforced concrete shear walls and steel braced frames are the lateral systems that were included in the simulated response development. The benefits of the augmented dataset was demonstrated by comparing the performance of SRR models developed using (i) only the measured responses and (ii) the combined simulated and measured responses. Lastly, the SRR model based on the combined dataset was used to perform a regional impact assessment of a building inventory using the earthquake-induced repair costs as the performance metric of interest.

The dissertation of Eusef Abdelmalek-Lee is approved.

Scott Joseph Brandenberg

Enrique Andres Lopez Droguett

Thomas A. Sabol

Henry Burton, Committee Chair

University of California, Los Angeles

2024

Table of Contents

List of Figures

List of Tables

Vita

2014 – 2018 B.S. in Civil Engineering, University of Florida (UF), Gainesville, Florida

2019 – 2020 M.S. in Civil Engineering, University of California Los Angeles (UCLA), Los Angeles, California

2020 – 2024 PhD Candidate in Civil Engineering, University of California Los Angeles (UCLA), Los Angeles, California

CHAPTER 1

Introduction

1.1 Background and Motivation

Strong motion building response data has been shown to be useful for near real time seismic impact assessment in the post-earthquake environment. However, the issue of partial instrumentation is a major challenge in utilizing these models to support real decision-making. Specifically, it is often the case that only a subset of the floors in an individual building and a fraction of buildings within an inventory is instrumented. This challenge can be addressed using structural response reconstruction (SRR) whereby the ”missing” responses are generated using a predictive model that utilizes, in part, the measured responses. While post-earthquake SRR has been addressed in prior studies, there are several outstanding issues such as the lack of well-organized strong motion response data sets and uncertainty quantification that have not been adequately addressed. The performance-based earthquake engineering (PBEE) paradigm has been shown to be useful for post-event evaluations informed by strong motion measurements. However, most of the work in this area has focused on individual buildings. Another challenge in this regard is the fact that the overwhelming majority of response measurements from historical earthquakes contain linear elastic responses. This significantly limits the usefulness of the SRR models developed because the domain of application is restricted to buildings that have experienced minimal damage.

The goal of this dissertation is to advance the state-of-the-art in using strong motion data for near real time seismic impact assessments in the post-earthquake environment. To address the issue of the lack of well-curated and easily accessible data sets, a relational database is developed, comprising historical data from 216 buildings subjected to historical California earthquakes over a period of 36 years. This data is used to develop a dual SRR model that combines kriging with the extreme gradient boosting algorithm. A key departure from prior SRR methods is the ”event-agnostic” nature of the model whose performance is not constrained by the number of recordings in the event of interest. The work will also consider regional seismic impact using performance metrics that are useful for stakeholders and overall decision making including economic losses. The model will propagate and quantify the uncertainty in different stages of the assessment including the SRR and loss assessment. Within the PBEE paradigm, the relative effectiveness of different ground motion intensity measures (IMs) in estimating responses is a key issue. This issue has been well-studied in simulation-based studies. In this dissertation, building strong motion recordings are, for the first time, used to assess the relative efficacy of different ground motion intensity measures. The culmination of this work concludes with a hybrid-data machine learning model, combining the benefits of both measured and simulated building response data to develop a multi-LFRS seismic response reconstruction and risk framework.

1.2 A Review of Relevant Literature

This section provides a brief review of the relevant literature in the context of the various sub-topics that are addressed in this dissertation. The objective of the review is to discuss pertinent prior studies, identify key areas for improvement, and set the stage for the research gaps that this dissertation will address. Relevant topics include structural health monitoring technologies and its role in structural response reconstruction modeling, the performance based earthquake engineering (PBEE) paradigm, data-driven predictive modeling, and other foundational studies. Subsequent chapters further explore literature relevant to the study being introduced where necessary.

1.2.1 Advancements and Techniques in Structural Health Monitoring (SHM)

Structural Health Monitoring (SHM) technology is an important aspect of new building design and analysis. It aids in more rapid and precise post-event damage assessments, can inform the sequencing and timing of on-site inspections, and allows for the advancement of structural modeling. SHM has been researched and improved greatly since its inception. The original sensors were much larger, hardwired, and costly due to the required installation labor. One estimate of the expected cost of installing a single wired sensor in a tall building exceeded $5000 [73]. There is also the distinction between passive and active sensors, where passive sensors merely record the states and excitations of structures, providing engineers and researchers with measurement data. Active sensors have the ability to excite the buildings themselves and record the response to that excitation [69]. Regardless of the type, size, or installation costs, these sensors were developed with the common goal of monitoring and improving structural health. Buildings are typically designed with an intended minimum "service life". To guard against the effects of both extreme (e.g., earthquakes) and slow onset (e.g., material degradation due to aging) events, the ability to monitor and maintain structural integrity through a building's service life, is of utmost importance. SHM enables a dynamic maintenance approach that is based on the current condition rather than the design life, potentially prolonging the structural longevity [48]. The trade-off is the need for advanced sensing technologies and analytical methods for measurement and performance assessment. However, like any other technology, SHM is not without its drawbacks. In most contexts, only a subset of the buildings within an inventory are equipped with sensors. This poses the challenge of using only the data collected from instrumented buildings to estimate the post-event impact to an entire inventory. One way to alleviate this challenge is to further expand the use of SHM technology. However, equipping an entire building portfolio with sensors is typically not economically feasible, so other methods must be employed to account for this information gap.

An important aspect of SHM is the post-processing of the data into models that can be used to make meaningful observations and predictions. Response reconstruction modeling serves as a way of understanding, predicting, and assessing structural responses to natural hazards which can also help fill the information gap left by partial instrumentation. Advancements in SHM have aided in the development of these types of models. Much of the more recent literature seeking to improve SHM methodologies use machine learning algorithms to make predictions and conduct damage assessments [69]. Some of these methodologies are limited to global damage categorizations, lacking the ability to identify and assess localized damage [1]. In [66], this issue is addressed, and a model is developed using neural networks that can characterize damage at both the local and global scales. However, simulated data from a finite element model was used rather than real measured data that would be provided by sensor-equipped structures. Similarly, [128] developed vibration-based models that can detect both global and local damage using finite element analyses paired with experimental data. These model-dependent damage assessments are limited by the ability of the finite element analyses to replicate the response of a real structure. Alternatively, [1] used convolutional neural networks for damage detection using measured data for a building at the University of British Columbia. The study used nine different structural damage conditions to test the algorithm, which was trained only with either damaged or undamaged states. The algorithm successfully assigned global (building-level) damage scores to each case but was unable to detect local damage. Another recent study coupled cumulative absolute velocity features with different machine learning (ML) algorithms to assess damage [83]. Global structural damage was classified with an accuracy of 84 % on the low end and 97% on the high end. The model also predicted damage severity and location with high accuracy (97% and 95%, respectively). The authors of this study noted that the performance of the models could be significantly improved with auxiliary data, thus highlighting the need for a well-established database of recorded building structural responses. Such a database can support the future development of robust models to assess both local and global damage in individual structures as well as the cumulative impact to building inventories.

1.2.2 The Role of Performance Based Earthquake Engineering (PBEE)

With advancements in structural modeling capabilities, seismic risk frameworks, such as the Pacific Earthquake Engineering Research (PEER) Center's Performance Based Earthquake Engineering (PBEE) [81] have been developed, utilizing nonlinear response history analyses (NRHA) to estimate building performance. PBEE enhances seismic risk decision-making by focusing on performance metrics like financial losses, downtime, and casualty risks to guide mitigation decisions. These principles are central to the PEER center's efforts to develop a comprehensive PBEE framework. To summarize, PBEE implements a probabilistic hazard approach, generating a seismic risk profile based on site conditions and location. Fragility models that estimate the probability of damage as a function of an intensity measure are used to estimate probable damage and losses. The predicted outcome from this procedure is described as a function of an intensity measure (IM), engineering demand parameters (EDP), damage measures (DM), and decision variables (DV). IMs describe the hazard level (i.e. ground motion when considering seismic hazards). EDPs are parameters of the structure response, such as the accelerations and drifts at various floors/stories in a building. DMs relate the responses from the EDPs to levels of damage to building components. DVs then relate these damage measures to important metrics for risk assessment such as economic losses and life safety. The risk prediction procedure can be described by the following triple integral:

$$\lambda(DV) = \int\int\int G(DV|DM)dG(DM|EDP)dG(EDP|IM)d\lambda(IM) \tag{1.1}$$

In this equation, $\lambda(DV)$ is a probabilistic characterization of the loss exceedance. For example, one might want to determine the probability of the average loss exceeding a certain fraction of the replacement cost for a building. $G(\cdot|\cdot)$ is a conditional probability of exceeding one variable conditioned on the value of another. For example $G(EDP|IM)$ would represent the probability of exceeding some response demand level, given a specific ground motion intensity value. This framework has had major implications on the field. For example, studies have been conducted that utilize the framework for optimizing automated building design and analysis [53, 36, 5, 94], potentially reducing the time and cost for building design. These automated programs also led to the development of comprehensive risk-focused databases, that include code-conforming designs [45], as well as building responses to suites of ground motions [78], providing repositories to support more efficient studies. Traditionally, PBEE is employed based on simulated responses obtained from NHRA of structural models, with few studies using instrumentation data [92, 60, 34]. This highlights an area for further exploration within PBEE. [34] directly addresses this as an area with potential. The study explored the benefits of including instrumentation data in the PBEE framework, evaluating how varying levels of instrumentation affect the accuracy of structural response data during seismic events. The results showed that, unsurprisingly, increasing the number of building instruments significantly reduces prediction errors, with substantial improvements even when only a few floors are instrumented. However, the arrangement of the sensors plays a critical role in prediction accuracy. The location of sensors in buildings across a region, for example, is not guaranteed. The findings offer practical guidance for building owners, informing optimal instrumentation placement. This study was also limited in scope, applying the methodology only to two 7-story buildings

1.2.3 Hybrid Data and Structural Response Reconstruction

As mentioned earlier, a common theme amongst many of the prior response reconstruction studies is that they often depend on one data type (measured or simulated). Comparatively, much less work has been done on the use of hybrid data (i.e. combining instrumentation data with simulated data) to ”fill the gaps” in the data distribution. The existing literature implementing this type of hybrid data is limited, and those that do often focus on reconstructing seismic excitations [127, 124], as opposed to structural responses and performance. For example, [127] presents a hybrid identification method to estimate structural parameters and ground motions in multi-story buildings. Shake table tests on a scaled building model validated the method, with acceleration data used to calculate velocity and displacement. The method identified stiffness and damping parameters and reconstructed seismic inputs. [124] introduced a method that combines sparsity-promoting transforms and rank reduction techniques. This hybrid approach improves signal-to-noise ratio (S/N) and data recovery efficiency. Tested on both synthetic and real seismic data, it outperformed existing methods in reconstruction and de-noising. Both of these studies employed some form of hybrid data modeling, but focused on reconstructing seismic excitations. One study that does focus on hybrid response reconstruction, published earlier this year, developed a two-step hybrid method that combines model-based and data-driven approaches to reconstruct the seismic response of non-instrumented floors in buildings [52]. The approach first calibrates a simplified shear-flexural beam model using data from instrumented floors. Then, Gaussian Process (GP) regression is applied to model the residual differences between measured data and model predictions, enabling the estimation of seismic responses at non-instrumented floors with quantified uncertainty. This study provided valuable insights on reconstructing responses at uninstrumented floors, but focused on a single building. The model performance was also largely dependent on sensor placement location and was limited by the simplified beam model and potential numerical convergence challenges that may arise when using GP regression.

1.2.4 Foundational Works

The proposed study aims to address some of the gaps in the state-of-the-art in using building strong motion response data to inform decision-making in the post-earthquake environment. An approach for developing structural response reconstruction models at the inventory level is proposed. The relative effectiveness of different ground intensity measures used in these predictive models is also examined. Next, data retrieved from the California Strong Motion Instrumentation Program (CSMIP) [35] are combined with simulations based on a comprehensive seismic design and analysis (SDA) platform, consisting of a conglomerate of previously developed SDA programs: AutoSDA ([53]) for steel moment frame structures, woodSDA [36], and RCWallSDA. Simulated response data generated from this platform is combined with the instrumentation data, creating a more extensive data distribution to be used as inputs into the predictive models. Using any kind of model introduces a level of uncertainty in the resulting inferences and predictions. So, quantifying this uncertainty will be important as it provides the level of confidence associated with the predictions. It also enables a comparative assessment of the models developed using a single data type (measured vs simulated) to those developed using hybrid data.

A portion of this work, specifically chapter 3, is similar in spirit to the work of [114], utilizing most of the same data, excluding recordings that were taken post-publication. The study developed a cross-building response reconstruction (CBRR) model that combines a structural response prediction model (SRPM) with the Kriging algorithm. The SRPM was used to estimate the median response demands and kriging was used to interpolate the residuals. Embedded in the SRPM was a ground motion model (GMM) that provided the benefit of a large data set of historical records. The use of the GMM and kriging on the residuals differs from the implementation outlined in this proposal. Other major differences between what this and the [114] study are as follows:

- **Hybrid Data**: A key limitation to the use of instrumentation data alone is that many buildings are not equipped with sensors. The typical solution is the development of predictive models, such as one study performed in this project, but a less-recognized approach is combining recorded data with simulated data to create a more comprehensive data repository.

- **Causal Inference**: A portion of this study focuses on using causal inference methodologies to assess and compare the effectiveness of ground motion measures on building responses.

- **Uncertainty Quantification**: In order to foster confidence in the models developed, part of the study will consider the various sources of uncertainty that contribute to the results.

- **Loss Assessment**: Rapid response reconstruction modeling is only one aspect of impact assessment. The ability to assess potential losses can aid in directing resources to areas of most significant need.

The goal is to develop a comprehensive framework for rapid post-event impact.The study builds upon the foundations of previous work, implementing novel ideas to advance regional seismic impact modeling.

1.3 Objectives

The proposed study seeks to improve previous methodologies, as well as implement new ones, to advance regional impact assessment using building strong motion response measurements. To do so, the primary research tasks are outlined as follows:

1. Develop a comprehensive and well-organized database of strong motion instrumentation data that is easily accessible.

2. Implement novel modeling techniques for predicting regional building responses at uninstrumented locations using only the instrumentation data, seeking to improve single-data type predictions.

3. Evaluate the effectiveness of different ground motion parameters for predictive modeling using strong motion building response data.

4. Create a computational platform that can generate simulated responses for multiple types of lateral force resisting systems including steel moment frames and braced frames, reinforced concrete shear walls and woodframe shear walls. The simulated data generated using this platform will be combined with the strong motion recordings to create the hybrid data set.

5. Use the hybrid structural response data (i.e., measured data from (1) and simulated data from (4) to develop a structural response reconstruction model and use the predicted responses to perform earthquake loss assessment at the building inventory level. The various sources of uncertainty are quantified and propagated and used to inform the level of confidence in the resulting conclusions.

1.4 Organization

The main study is comprised of five chapters, two of which are based on published journal articles.

In chapter 2, a summary of the development of a relational database in MySQL which aims to serve as an aid to engineers and researchers is provided. The database is hosted on DesignSafe and includes data from a wide range of building heights, construction types, and materials, aiding in modeling and analysis. It supports dynamic updating, eliminating tedious data collection and providing an efficient way to manage the information. The relational structure of the database serves as an efficient medium for storing data and retrieving specific information. The data collected for this database is integral to the research initiatives outlined in the following chapters.

Chapter 3 presents a machine learning model utilizing the database to make predictions on maximum building response parameters when subjected to seismic loading. It utilizes a dual-phase modeling technique that takes advantage of spatial relationships in the data to improve model performance. This study builds upon the work of [114], taking an event-agnostic approach to create a more generalized model formulation.

The goal of chapter 4 is to evaluate the effectiveness of various ground intensity measures (IMs) that could potentially be used as inputs into the response prediction models chapter 3. The commonly used sufficiency and efficiency are used as metrics for evaluating the performance of the IMs. Additionally, an evaluation approach based on causal inference is also introduced.

Chapter 5 summarizes the development of an Automated Seismic Design and Analysis (AutoSDA) platform that is used to generate the simulated responses used in the hybrid data employed in the final chapter. The program unifies three previously developed SDA programs (Steel-SDA [53], Wood-SDA [36], and RCWall-SDA [5]. In addition, it develops a module to conduct loss assessments on the buildings designed and analyzed in the SDA modules using the performance based earthquake engineering (PBEE) methodology.

Chapter 6 combines data and methodologies from previous chapters to develop a structural response reconstruction (SRR) model. Models designed and analyzed using the program developed in chapter 5 are combined with the data from chapter 2 to train a machine learning model similar to that of chapter 3. Building response parameters outlined in chapter 4 are integral to the reconstruction. The model improved full-profile response predictions of measured data by augmenting it with the simulated response data. It serves as a way to alleviate the persistent problem of partial instrumentation (both within a single building and across a building portfolio) in response reconstruction. This chapter (and study) culminates with a regional assessment of the earthquake induced economic losses (or repair costs) that relies on the aforementioned response reconstruction model.

Finally, in chapter 7, the key findings of the dissertation are summarized, highlighting the potential impact of the outlined studies, the limitations, and future work.

CHAPTER 2

Relational Database for Building Strong Motion Recordings used for Seismic Impact Assessments

2.1 Introduction

A robust set of historical data from instrumented buildings impacted by earthquakes is crucial to system-identification, predictive modeling, and risk mitigation efforts. Such models can be used for “forward-prediction” in the case of likely future events or reconstructing responses for an event that has already occurred. These models also facilitate rapid impact assessment on a regional scale, serving as complement to in-person inspections which often take much longer to complete [40]. The field of Structural Health Monitoring (SHM) was established with the goal of creating technologies and tools for collecting and synthesizing response measurements from instrumented structures. SHM has enabled faster identification of structural deficiencies [10] and more accurate modeling capabilities. Compared to building-specific problems, the use of SHM data for inventory-based seismic impact assessments has received much less attention in the research literature. This is partly due to the absence of a single, well-curated and easily accessible repository of building strong motion data.

A review of the SHM literature found many developments to improve both the functionality and ease of installation and maintenance of sensors. [73] discusses different types of sensors, including active and passive. The distinction between the two is that, unlike passive sensors which only record the response, those that are active have the added ability to excite the structure. Other studies have documented changes in the cost of installation and maintenance with the development of more advanced technology, which allows the sensor and the data acquisition unit to be separately located. These advancements have led to smaller sensors that are easier to access [69, 114, 40, 116]. While older sensors were installed by connecting them to coaxial wires located throughout the instrumented building, wireless sensors are becoming the standard, greatly reducing the cost of installation and maintenance. With these improvements in SHM, it is becoming more accessible and feasible to be included in construction projects.

The engineering demand parameters ($EDPs$) used in performance-based assessments can be calculated from sensor data [81]. In [113], spatial correlation of simulated $EDPs$ was leveraged and the kriging interpolation algorithm was used to estimate peak responses for a given building and set of event parameters. A later study by the same authors [114] used strong motion data to develop a generalized cross-building response reconstruction model. The formulation followed that of typical ground motion models ($GMMs$) and kriging was applied to the residuals of a structural response prediction equation.

The purpose of this paper is to develop a relational database that is easily accessible and facilitates response visualization, modeling, and performance assessment. The database is designed for ease of expansion as new data becomes available. The database has been developed using MySQL [84] [3] and made available on the DesignSafe platform [93] using SQLite. It is paired with a Jupyter Notebook tool developed using the Python language. The tool allows the user to visualize specific data within the database without the need for prior knowledge in SQL or querying. The scripts are compartmentalized based on the type of information to be accessed, and is organized in a way that requires minimal user input and manipulation. The following sections will summarize the data included in the database and how the sensor data is collected. The organization structure of the database is also described, including examples that demonstrate how to visualize the tabular layout of the different data types.

2.2 Summary of the Data

The relational database consists of data collected from the California Strong Motion Instrumentation Program ($CSMIP$) [35]. Most of the data was used in the study by [114], however, additional recordings from 10 recent events have been included as part of the current database. The set includes sensor data from instrumented buildings that recorded events from 1984 to 2020. Apart from a few exceptions, only events with $\mathbf{M} > 4$ and more than 5 recordings are included. Table 2.1 shows the type of data collected for each event and Table 3.1 shows the data collected for each building for a given event. Note that only a small subset of the buildings in the database is included in Table 3.1 for illustration purposes. The latitude and longitude information is important for calculating the Joyner-Boore distance for each building, which is often used as a feature in training predictive models. The Joyner-Boore distance is defined as the closest distance to the fault surface projection [64]. The fault types were collected from the US Geological Survey [120]. For the building parameters, time-series responses (acceleration, velocity, and displacement histories) for the vertical and two horizontal directions were collected, and the channel numbers that correspond to the sensor layout plans (e.g., Figure 3.4) are included. The floor level where the sensor is located and its height relative to the ground is also recorded.

Shown in Figure 3.1 is the distribution of the number of stories for the buildings and event magnitudes that are included in the data set. These distributions highlight the lack of recordings for buildings with more than 20 stories, which has implications to the generalizability of the models developed using these responses. Figure 2.2 and Figure 2.3 show the recorded $EDPs$ (calculated using the Python tool) for each building, which are grouped based on the magnitude of the earthquake. Note that peak story drift ratios ($PSDR$) (Figure 2.2) and peak floor acceleration (PFA) Figure 2.3 are used to denote the peak response corresponding to each story and floor respectively. We will use $maxPSDR$ and $maxPFA$ to denote the maximum $PSDR$ and PFA over all stories and floors, respectively.The individual story drift ratios are calculated by taking the difference in the horizontal displacement between the two floors that make up a story and dividing by the story height. The floor

Table 2.1: Summary of all events included in the database

Event Name	**Depth (km)**	**Magnitude**	**Date**	**Latitude**	**Longitude**	**Fault Type**	**No. of Records**
Morgan Hill	9	6.2	4/24/1984	37.32	-121.68	Strike Slip	7
Mount Lewis	6	5.8	3/31/1986	37.466	-121.691	Strike Slip	3
Whittier	9.5	6.1	10/1/1987	34.06	-118.07	Strike Slip	16
Loma Prieta	18	7	10/17/1989	37.04	-121.88	Strike Slip	29
Sierra Madre	12	5.8	6/11/1991	34.26	-118.00	Reverse	8
Big Bear	1	6.5	6/28/1992	34.20	-116.83	Strike Slip	5
Landers	1.1	7.3	6/28/1992	34.217	-116.433	Strike Slip	23
Northridge	19	6.4	1/17/1994	34.2057	-118.5539	Reverse	35
Bolinas	7	5	8/17/1999	37.91	-122.69	Strike Slip	12
Gilroy	7.6	4.9	5/13/2002	36.97	-121.6	Strike Slip	23
Big Bear C	12.7	5.4	2/22/2003	34.31	-116.85	Strike Slip	6
Simi Valley	5.1	3.7	10/29/2003	34.29	-118.75	Reverse	1
San Simeo	4.7	6.5	12/22/2003	35.71	-121.1	Transform	6
Parkfield	7.9	6	9/28/2004	35.81	-120.37	Strike Slip	3
Anza05	13.1	5.2	6/12/2005	33.53	-116.57	Strike Slip	13
Alum Rock	9.2	5.4	10/30/2007	37.432	-121.776	Strike Slip	43
Chino Hill	13.6	5.4	7/29/2008	33.95	-117.77	Strike Slip	56
Calexico	10	7.2	4/4/2010	32.26	-115.29	Strike Slip	40
Borrego	14	5.4	7/7/2010	33.42	-116.49	Strike Slip	50
Berkeley11	8	4	10/20/2011	37.86	-122.25	Strike Slip	22
Anza13	13.1	4.7	3/11/2013	33.50	-116.46	Strike Slip	5
Encino	9.9	4.4	3/17/2014	34.13	-118.49	Strike Slip	22
South Napa	11.3	6	8/24/2014	38.2155	-122.3117	Strike Slip	45
Borrego	12.3	5.2	6/10/2016	33.43	-116.44	Strike Slip	21
Berkeley18	12.3	4.4	1/4/2018	37.8552	-122.2568	Strike Slip	31
Trabuco	11.2	4	1/25/2018	33.741	-117.4912	Strike Slip	5
Cabazon	12.9	4.5	5/8/2018	34.016	-116.7798	Reverse	19
Laverne	5.5	4.4	8/28/2018	34.1363	-117.7747	Reverse	6
Ridgecrest	8	7.1	7/5/2019	35.7695	-117.5993	Strike Slip	67
Pleasant Hill	14	4.5	10/14/2019	37.938	-122.057	Strike Slip	21
Petrolia	3.2	5.8	3/8/2020	40.3917	-125.0937	Strike Slip	2
Anza	10.3	4.9	4/3/2020	33.4932	-116.505	Strike Slip	9
Searles Valley	8.4	5.5	6/3/2020	35.6148	-117.4282	Strike Slip	7
Pacoima	8.8	4.2	7/30/2020	34.3017	-118.4383	Reverse	11
South El Monte	16.9	4.5	9/18/2020	34.0385	-118.0803	Blind Thrust	28

Table 2.2: Summary of information for a sample set of buildings in the database

Building ID	Latitude	Longitude	No. Channels	Story Above	Story Below	X Channel	Y Channel	Z Channel	Story Height (cm)	Instrumentation Date
13589	33.6243	-117.9304	18	11	0	3;6;9;13	4;7;10;14	0;3;6;11	0;1229.36;2418.08;4478.02	1990
13620	33.979	-117.373	12	2	0	2;4;8	3;5;10	0;1;2	0;365.76;838.2	1991
14578	33.7567	-118.0851	31	8	1	6;8;25;11;14	7;10;18;12;16	-1;0;1;5;8	-487.68;0;441.96;2209.8;3535.68	1989
14606	33.9749	-118.0366	12	8	0	3;6;9	4;7;10	0;4;8	0;1158.24;2316.48	1991
14654	33.9206	-118.3914	16	14	0	5;8;10;13	4;7;9;11	0;3;8;14	0;1310.64;3291.84;5730.24	1993
23285	34.1822	-117.3242	10	5	1	3;9;5	1;10;6	-1;2;5	-396.24;853.44;2042.16	1976
23287	34.0655	-117.2802	9	6	0	3;5;8	1;4;7	0;2;6	0;530.86;1557.02	1976
23497	34.1044	-117.5739	16	4	1	16;15;14	12;9;6	-1;0;4	-426.72;0;1828.8	1985
23540	34.0721	-117.3359	13	1	0	13;8	6;4	0;1	0;883.92	1988
23634	34.1318	-117.3219	12	5	0	2;5;9	3;7;11	0;2;5	0;975.36;2103.12	1992

accelerations are directly measured by the sensors.

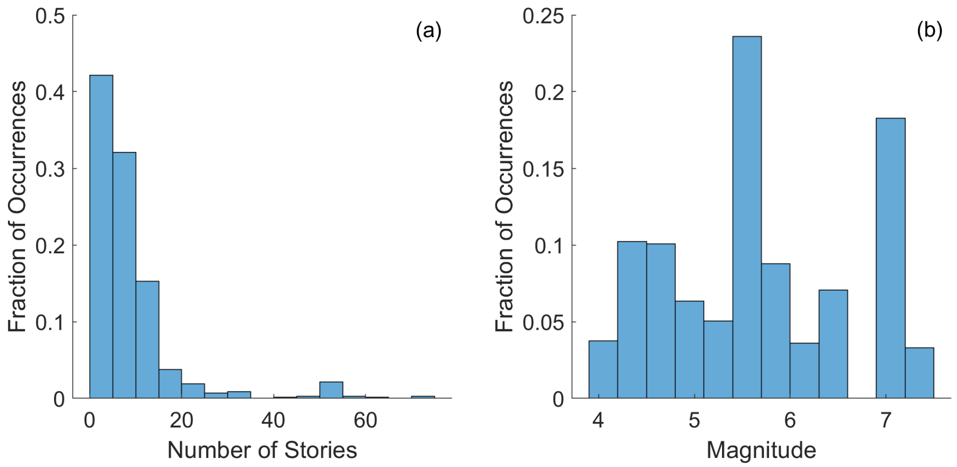


Figure 2.1: Histograms showing empirical distribution of (a) the number of stories and (b) earthquake magnitude

Using their locations, a map of all buildings that have recordings for an event can be rendered, as shown in Figure 3.2, which is associated with the 2019 Ridgecrest earthquake. The epicenter is marked with a red star and building locations are marked with green crosses. The $EDPs$ by floor are also visualized in Figure 3.3, which shows $PSDR$ and PFA profile over the building height.

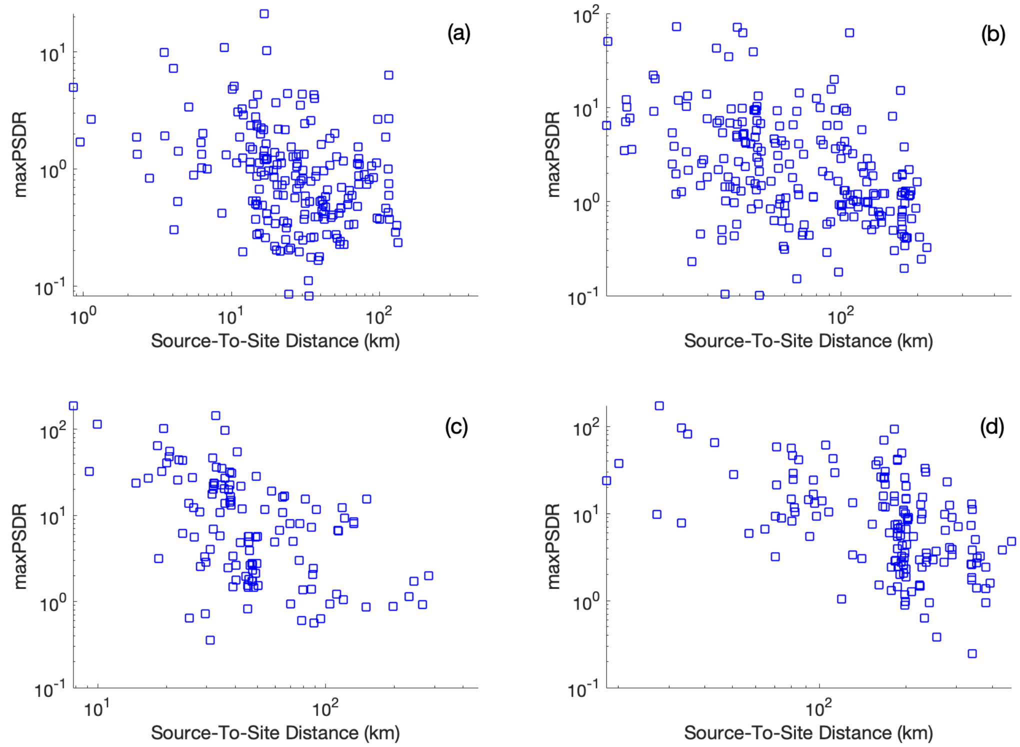


Figure 2.2: Maximum (over the building height) peak story drift ratio (maxPSDR) for all buildings disaggregated based on event magnitude: (a) $\mathbf{M} < 5.0$, (b) $5.0 \leq \mathbf{M} < 6.0$, (c) $6.0 \leq \mathbf{M} < 7.0$ and (d) $\mathbf{M} \geq 7.0$

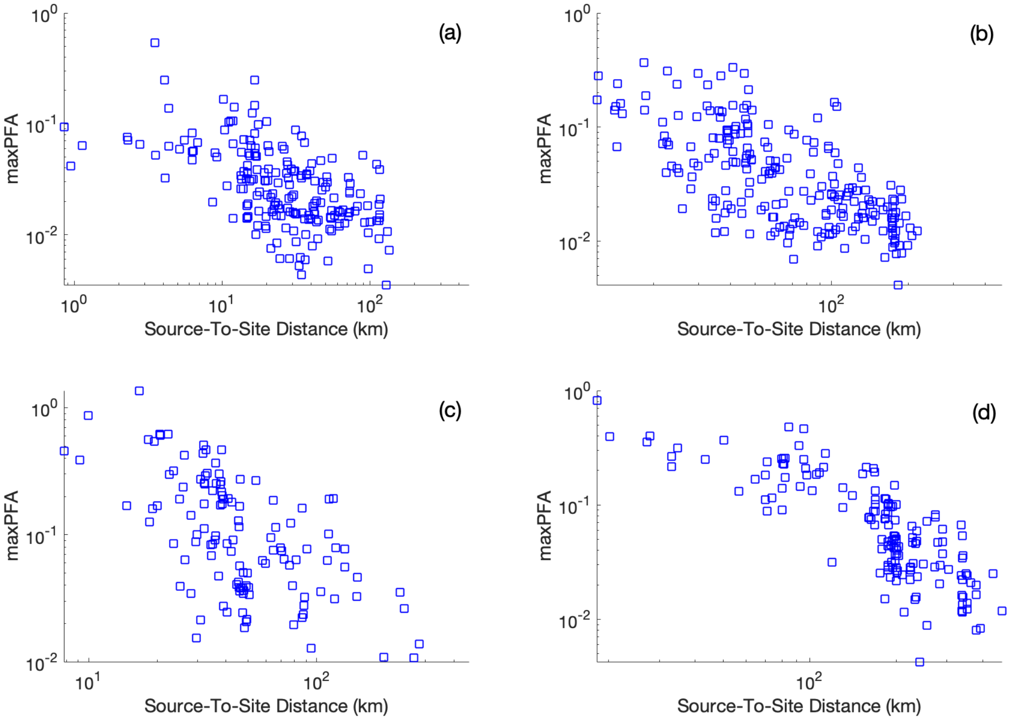


Figure 2.3: Maximum (over the building height) peak floor acceleration (maxPFA) for all buildings dissagrregated based on event magnitude: (a) $\mathbf{M} < 5.0$, (b) $5.0 \leq \mathbf{M} < 6.0$, (c) $6.0 \leq \mathbf{M} < 7.0$ and (d) $\mathbf{M} \geq 7.0$

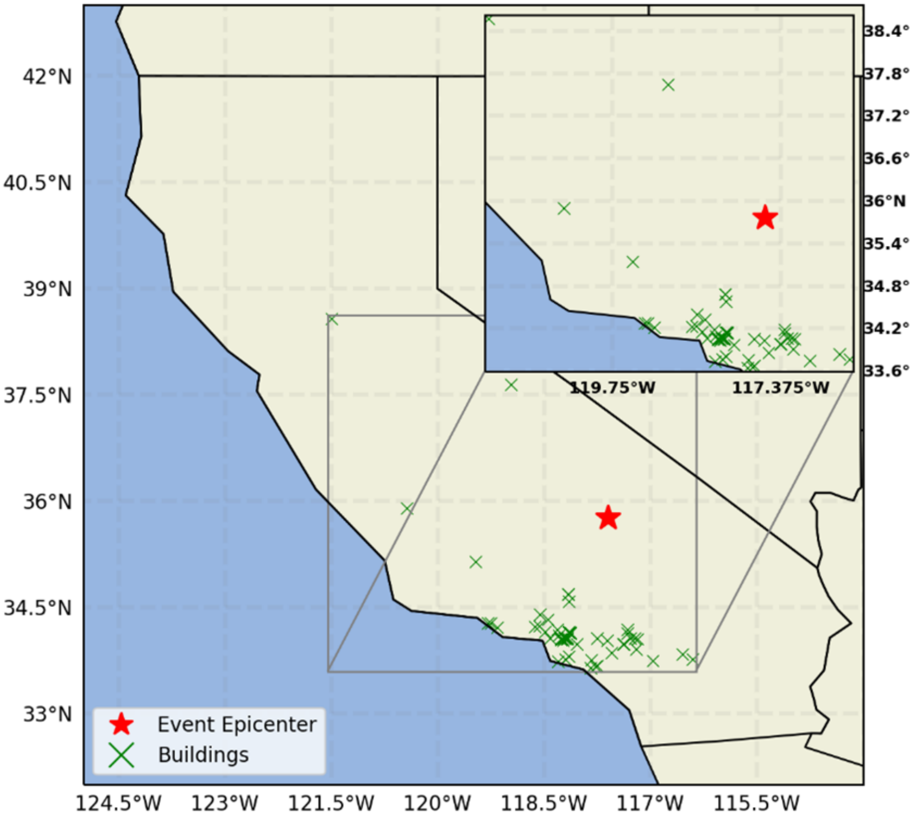


Figure 2.4: Map showing the location of buildings with recordings from the **M** 7.1 Ridgecrest earthquake

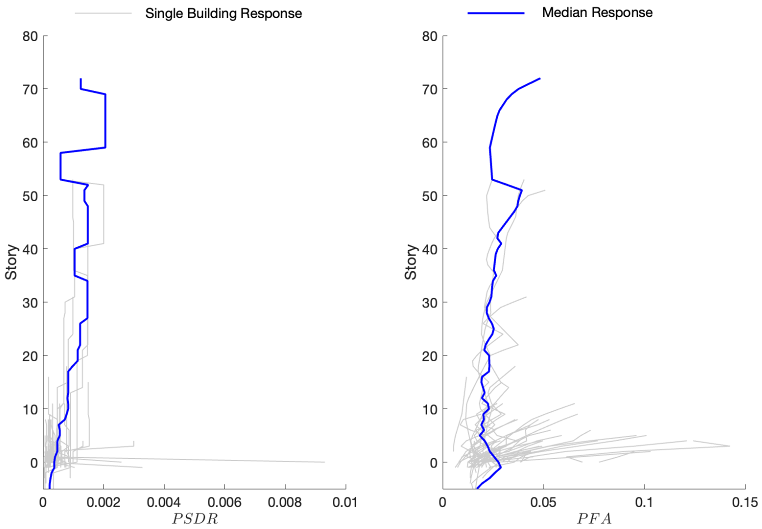


Figure 2.5: Building response profiles for the **M** 7.1 Ridgecrest earthquake: (a) PSDR and (b) PFA

Figure 3.4 and Figure 3.5 show information that are specific to each building. In Figure 3.4, a partial sensor layout plan for building 23285 (CSMIP ID) is shown. It shows the exact location of the sensors in the building and the associated direction of motion. The time-series roof displacement response for the same building is visualized in Figure 3.5. Figure 3.6 is similar to Figure 3.3, except that it shows the PSDR and PFA at each story/floor for building 23285 during the corresponding event (in this case, Ridgecrest). The Python tool provides additional animated visualizations that show the real-time displacement response of each building.

Figure 3.5 only shows the displacement time-series data for a single building. However, as noted earlier, the corresponding acceleration and velocity responses are also included in the database. From this raw data, the $EDPs$ can be calculated, which is also performed and visualized in the Python tool. The lateral force resisting system ($LFRS$) of each building is also indicated in the database (Table A.1). Many of the buildings contain mixed $LFRS$ types, which is common. For example, it is not uncommon for a building to be constructed with a concrete shear wall and steel moment frames. This information is important because it can be used as a feature in predictive models. By including the $LFRS$ as a feature, the strength limits of different materials and system configurations are considered.

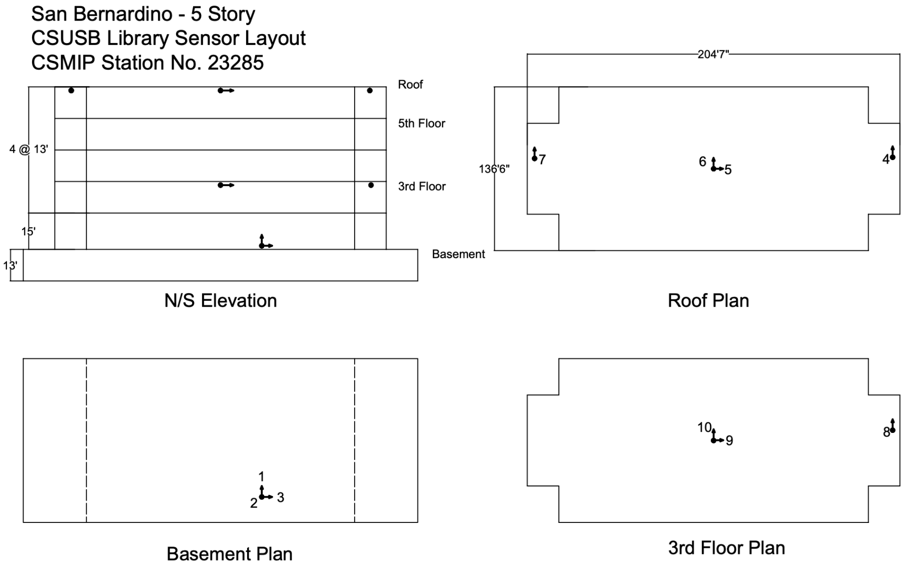


Figure 2.6: Building 23285 partial sensor layout

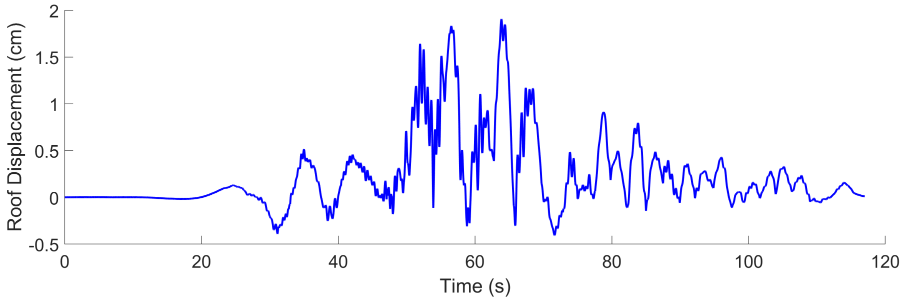


Figure 2.7: Building 23285 roof displacement time series for the **M** 7.1 Ridgecrest earthquake

2.3 Relational Database

2.3.1 Overview of Relational Databases

Relational databases are digital platforms that organize data in a tabular form based on relationships in the data [84]. This allows the data to be easily queried and users can extract specific information. Relational databases can easily be modified without affecting the original organization or data. The database presented in this paper was developed using $MySQL$

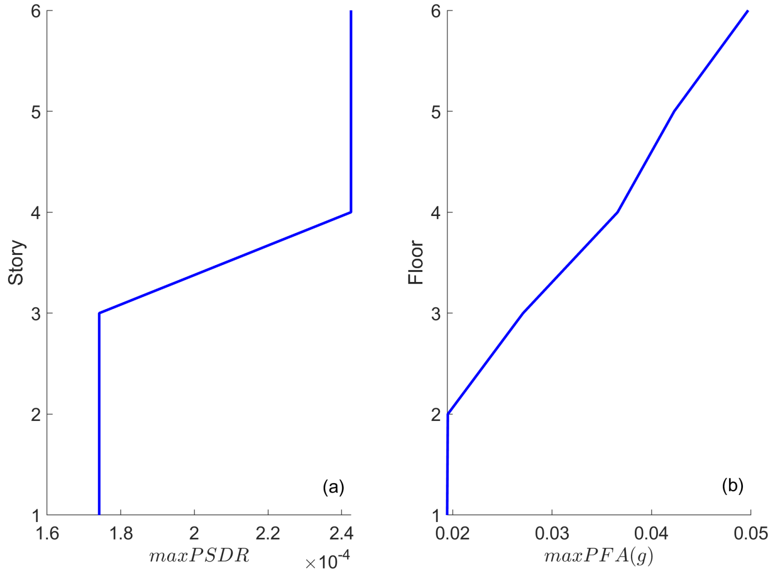


Figure 2.8: Building 23285 (a) PSDR and (b) PFA recorded during the **M** 7.1 Ridgecrest earthquake

and can be dynamically updated as more events occur, and additional data is generated. A relational database containing organized information from a large set of earthquakes and instrumented buildings would, for example, allow for easily accessible training data for ML models. The relational database introduced in this paper aims to serve as a medium for public use. Within the earthquake engineering community, relational databases have been used to organize data related to the seismic performance of buildings [47], ground motion liquefaction [17], post-earthquake damage and recovery of buildings [88] and subduction ground motions [76].

2.3.2 Schema

Shown in Figure 3.7 is the schema for the relational database, which provides a graphical visualization of how it is organized. In other words, the schema defines the structure of the relational database and has three main components. The primary key identifies the type of information in the table. This could correspond to the title of the table if the data were to be organized in Excel. The foreign key is a field within the table that links to the primary key of another table. This is important because it eliminates the need for duplicate information in the database. Attributes are the descriptors of the data points i.e., specific properties or values associated with the data. These can be in the form of integers, dates, strings, etc.

The schema in Figure 3.7 shows the tables that are included in the database as well as their relationships to one another and Figure B.1 shows the Schema implementation in *MySQL*. There are 3 categories of data: earthquake properties, building properties, and response data. These categories contain tables that summarize information about each of the primary keys. Tables A.1 - A.5 provide more detailed information about the data displayed in the schema. The earthquake properties table (Table A.2) contains the earthquake name, the date that it occurred, latitude and longitude of the epicenter, magnitude, type of fault, focal depth, and the number of buildings that recorded data for the event. The earthquake properties table does not include any foreign key relationships, but is referenced by other tables.

In the building properties section, there are the site information and building information tables. The building information table contains details that are specific to each building, including the IDs and locations of the sensors. The table has a foreign key to the site information table. The site information table provides the geography of the building location.

The response information category is where the actual data that was collected by the sensors is stored. The time-series data is stored in a single cell of the table as a comma-separated MEDIUMTEXT type object. This method was chosen as it simplifies the structure of the database, while ensuring ease of use. The earthquake-building pair table is a junction table that is used to alleviate the many-to-many relationship between earthquake and CSMIP building IDs. Each building is not limited to recording data from only one event and each event has recordings from many different buildings. The table creates a unique ID for each earthquake-building pair, which makes accessing the specific data more intuitive.

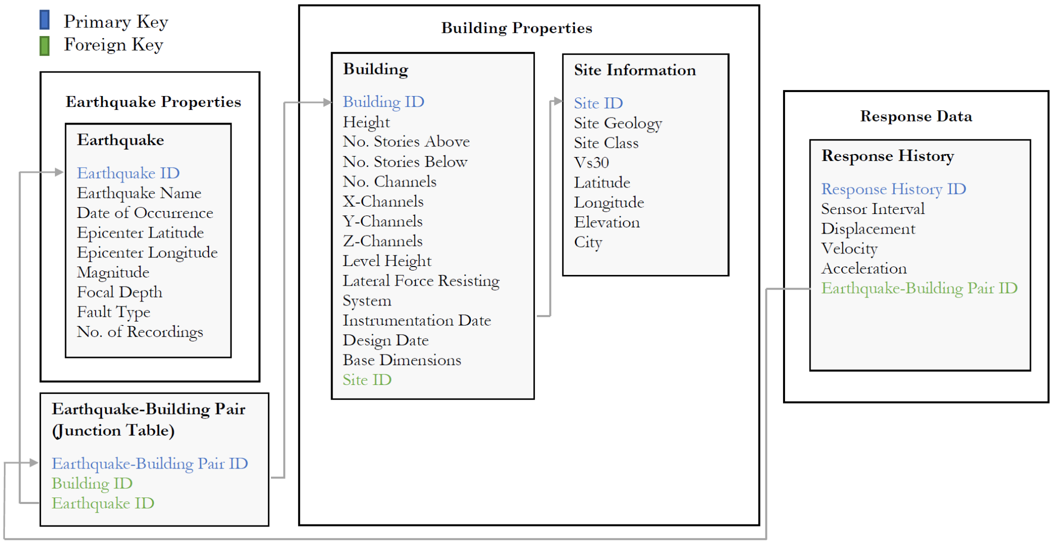


Figure 2.9: Schema for the relational database

2.3.3 Description of Tables and Relationships

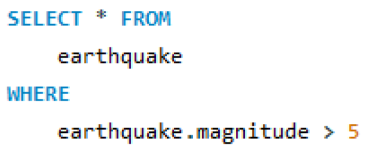
Tables A.1 to A.5 visualize the tabular data organization within the database. Similar to the schema, in each table, the primary keys are highlighted blue, and the foreign keys are highlighted in green. There are 6 different data types used in the database: INT, DECIMAL, VARCHAR, DATE, YEAR and MEDIUMTEXT. The INT and DECIMAL types are used to display numbers in either integer or decimal format, respectively. VARCHAR is used for strings that may contain up to a specified number of characters (usually 255). DATE displays the date in the MM/DD/YYYY format. The MEDIUMTEXT is used to store the time series data in a single cell as a comma separated list for each sensor in each building. The example column in each table corresponds to the Ridgecrest_23285 earthquake-building pair.

Table 3.1 shows how the information in each building is defined and organized in the database. It has a foreign key to the site information (Table A.4). It should be noted that the Channels attributes are defined by the channel ID and the story at which the channel is located in parenthesis. There are two non-numeric story level identifiers used in the database (P and F). These correspond to Penthouse and Foundation levels, which are not included in the number of stories in a building, but may have sensors in these locations. The site information contains details of the geography of the building site such as the location, geology, site class, average shear wave velocity in the top 30 meters of soil (Vs30), and elevation. Note that Vs30 is provided as VARCHAR to allow for distinction between measured and inferred Vs30 values. This table does not have foreign keys to other tables because the attributes are specific to each building. Table A.2 shows the earthquake table, displaying information about each event. This table is only referenced by the junction table shown in Table A.3, but does not reference other tables. The junction table creates a unique ID for each earthquake-building pair which is used to extract response history data. This avoids the inadvertent retrieval of incorrect data. For example, if a building is associated with multiple events, but the response data for one earthquake is needed, without this junction table, it would be easy to accidentally obtain the data for one of the other events recorded by that building. The response history table uses the earthquake-building pair ID as well as the Response History ID to identify each recording. It contains the time-dependent displacement, velocity, and acceleration data for each sensor.

2.3.4 Example Queries

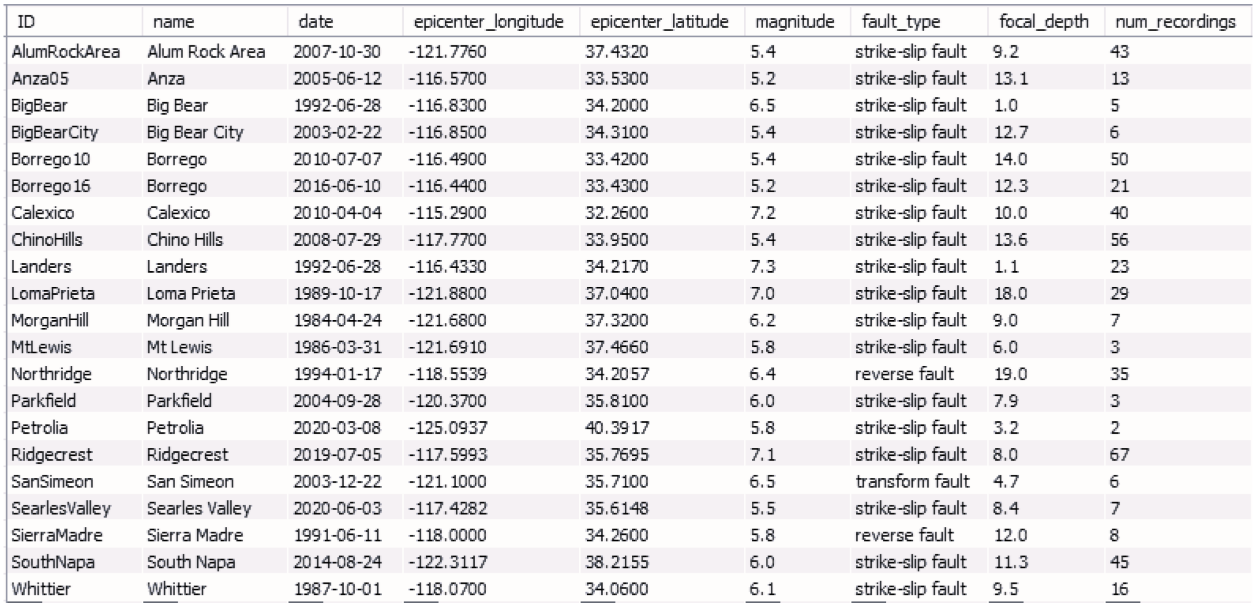
This section demonstrates how to perform queries in *MySQL* to extract specific information from the database. Five queries are presented, starting with a simple example, and progressively becoming more complex. The goal is to show how information can be extracted without knowing the exact structure of the database i.e., labels, join tables, etc. Figure 3.9a shows a simple query, selecting all columns and entries from the earthquake table for **M** ¿ 5. The '*' symbol after 'SELECT' is the syntax for selecting all columns/entries. The 'WHERE' command is used to specify the data to be selected, in this case, selecting rows in the table where the value in the magnitude column is greater than 5. Figure 3.9b shows the table that is output from this query.

If all columns from a table are not needed, a specific subset within the table can also be selected, as shown in Figure 3.10. The ID, date, magnitude, and fault_type columns are



```
SELECT * FROM
    earthquake
WHERE
    earthquake.magnitude > 5
```

(a)

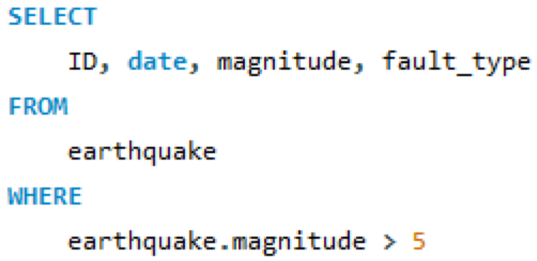


ID	name	date	epicenter_longitude	epicenter_latitude	magnitude	fault_type	focal_depth	num_recordings
AlumRockArea	Alum Rock Area	2007-10-30	-121.7760	37.4320	5.4	strike-slip fault	9.2	43
Anza05	Anza	2005-06-12	-116.5700	33.5300	5.2	strike-slip fault	13.1	13
BigBear	Big Bear	1992-06-28	-116.8300	34.2000	6.5	strike-slip fault	1.0	5
BigBearCity	Big Bear City	2003-02-22	-116.8500	34.3100	5.4	strike-slip fault	12.7	6
Borrego10	Borrego	2010-07-07	-116.4900	33.4200	5.4	strike-slip fault	14.0	50
Borrego16	Borrego	2016-06-10	-116.4400	33.4300	5.2	strike-slip fault	12.3	21
Calexico	Calexico	2010-04-04	-115.2900	32.2600	7.2	strike-slip fault	10.0	40
ChinoHills	Chino Hills	2008-07-29	-117.7700	33.9500	5.4	strike-slip fault	13.6	56
Landers	Landers	1992-06-28	-116.4330	34.2170	7.3	strike-slip fault	1.1	23
LomaPrieta	Loma Prieta	1989-10-17	-121.8800	37.0400	7.0	strike-slip fault	18.0	29
MorganHill	Morgan Hill	1984-04-24	-121.6800	37.3200	6.2	strike-slip fault	9.0	7
MtLewis	Mt Lewis	1986-03-31	-121.6910	37.4660	5.8	strike-slip fault	6.0	3
Northridge	Northridge	1994-01-17	-118.5539	34.2057	6.4	reverse fault	19.0	35
Parkfield	Parkfield	2004-09-28	-120.3700	35.8100	6.0	strike-slip fault	7.9	3
Petrolia	Petrolia	2020-03-08	-125.0937	40.3917	5.8	strike-slip fault	3.2	2
Ridgecrest	Ridgecrest	2019-07-05	-117.5993	35.7695	7.1	strike-slip fault	8.0	67
SanSimeon	San Simeon	2003-12-22	-121.1000	35.7100	6.5	transform fault	4.7	6
SearlesValley	Searles Valley	2020-06-03	-117.4282	35.6148	5.5	strike-slip fault	8.4	7
SierraMadre	Sierra Madre	1991-06-11	-118.0000	34.2600	5.8	reverse fault	12.0	8
SouthNapa	South Napa	2014-08-24	-122.3117	38.2155	6.0	strike-slip fault	11.3	45
Whittier	Whittier	1987-10-01	-118.0700	34.0600	6.1	strike-slip fault	9.5	16

(b)

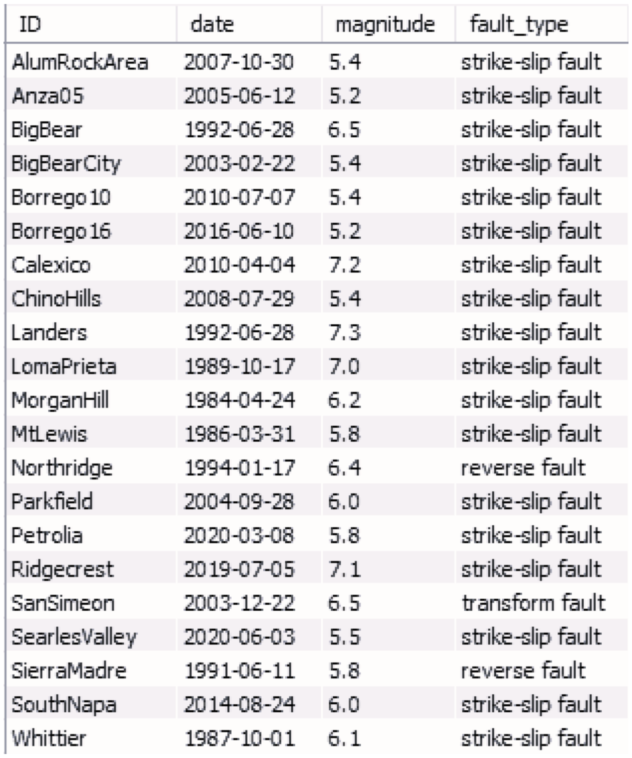
Figure 2.10: Query used to retrieve all attributes for earthquakes with $\mathbf{M} > 5$: (a) *MySQL* commands and (b) query results

selected from the earthquake table using this simple query.



```
SELECT
    ID, date, magnitude, fault_type
FROM
    earthquake
WHERE
    earthquake.magnitude > 5
```

(a)

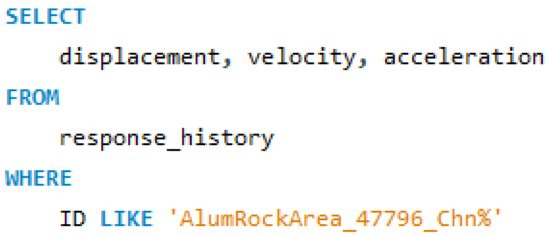


ID	date	magnitude	fault_type
AlumRockArea	2007-10-30	5.4	strike-slip fault
Anza05	2005-06-12	5.2	strike-slip fault
BigBear	1992-06-28	6.5	strike-slip fault
BigBearCity	2003-02-22	5.4	strike-slip fault
Borrego10	2010-07-07	5.4	strike-slip fault
Borrego16	2016-06-10	5.2	strike-slip fault
Calexico	2010-04-04	7.2	strike-slip fault
ChinoHills	2008-07-29	5.4	strike-slip fault
Landers	1992-06-28	7.3	strike-slip fault
LomaPrieta	1989-10-17	7.0	strike-slip fault
MorganHill	1984-04-24	6.2	strike-slip fault
MtLewis	1986-03-31	5.8	strike-slip fault
Northridge	1994-01-17	6.4	reverse fault
Parkfield	2004-09-28	6.0	strike-slip fault
Petrolia	2020-03-08	5.8	strike-slip fault
Ridgecrest	2019-07-05	7.1	strike-slip fault
SanSimeon	2003-12-22	6.5	transform fault
SearlesValley	2020-06-03	5.5	strike-slip fault
SierraMadre	1991-06-11	5.8	reverse fault
SouthNapa	2014-08-24	6.0	strike-slip fault
Whittier	1987-10-01	6.1	strike-slip fault

(b)

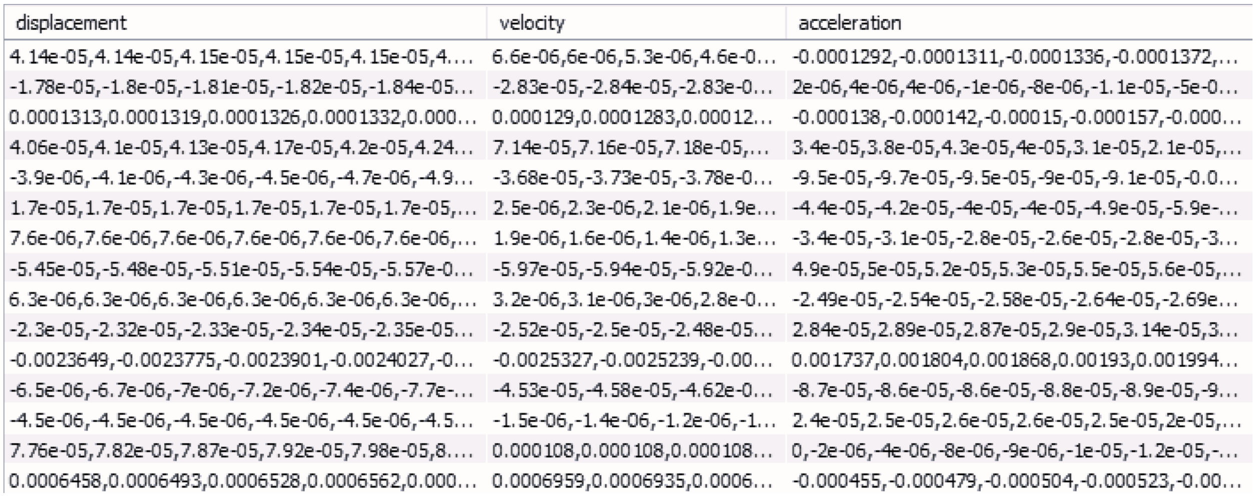
Figure 2.11: Query used to retrieve the ID, date, magnitude, and fault_type from earthquakes table **M** > 5: (a) MySQL commands and (b) query results

Another important feature is the ability to query data without fully specifying the name or label in the database. This is done using the LIKE command with the '%' symbol. For example, in Figure 3.11, the response data for building ID 47796 is queried. The channel number is not specified, but replaced with a % symbol. This means that the response data for every entry beginning with “AlumRockArea_47796_Chn” will be selected. This will select all channels in this building since the response data all have the same label, except for the channel number. The output is shown in Figure 3.11b. The LIKE command is useful not only when multiple outputs are desired, but if the user is unsure of the exact label, but is aware of one that is similar.



```
SELECT
    displacement, velocity, acceleration
FROM
    response_history
WHERE
    ID LIKE 'AlumRockArea_47796_Chn%'
```

(a)

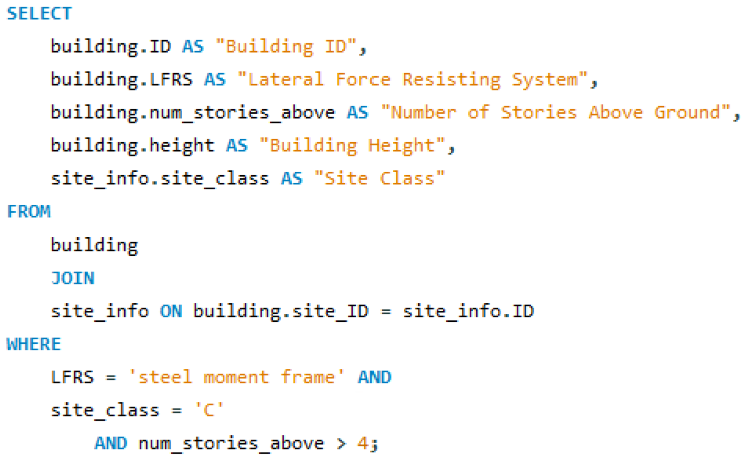


displacement	velocity	acceleration
4.14e-05,4.14e-05,4.15e-05,4.15e-05,4.15e-05,4....	6.6e-06,6e-06,5.3e-06,4.6e-0...	-0.0001292,-0.0001311,-0.0001336,-0.0001372,...
-1.78e-05,-1.8e-05,-1.81e-05,-1.82e-05,-1.84e-05...	-2.83e-05,-2.84e-05,-2.83e-0...	2e-06,4e-06,4e-06,-1e-06,-8e-06,-1.1e-05,-5e-0...
0.0001313,0.0001319,0.0001326,0.0001332,0.000...	0.000129,0.0001283,0.00012...	-0.000138,-0.000142,-0.00015,-0.000157,-0.000...
4.06e-05,4.1e-05,4.13e-05,4.17e-05,4.2e-05,4.24...	7.14e-05,7.16e-05,7.18e-05,...	3.4e-05,3.8e-05,4.3e-05,4e-05,3.1e-05,2.1e-05,...
-3.9e-06,-4.1e-06,-4.3e-06,-4.5e-06,-4.7e-06,-4.9...	-3.68e-05,-3.73e-05,-3.78e-0...	-9.5e-05,-9.7e-05,-9.5e-05,-9e-05,-9.1e-05,-0.0...
1.7e-05,1.7e-05,1.7e-05,1.7e-05,1.7e-05,1.7e-05,...	2.5e-06,2.3e-06,2.1e-06,1.9e...	-4.4e-05,-4.2e-05,-4e-05,-4e-05,-4.9e-05,-5.9e-...
7.6e-06,7.6e-06,7.6e-06,7.6e-06,7.6e-06,7.6e-06,...	1.9e-06,1.6e-06,1.4e-06,1.3e...	-3.4e-05,-3.1e-05,-2.8e-05,-2.6e-05,-2.8e-05,-3...
-5.45e-05,-5.48e-05,-5.51e-05,-5.54e-05,-5.57e-0...	-5.97e-05,-5.94e-05,-5.92e-0...	4.9e-05,5e-05,5.2e-05,5.3e-05,5.5e-05,5.6e-05,...
6.3e-06,6.3e-06,6.3e-06,6.3e-06,6.3e-06,6.3e-06,...	3.2e-06,3.1e-06,3e-06,2.8e-0...	-2.49e-05,-2.54e-05,-2.58e-05,-2.64e-05,-2.69e...
-2.3e-05,-2.32e-05,-2.33e-05,-2.34e-05,-2.35e-05...	-2.52e-05,-2.5e-05,-2.48e-05...	2.84e-05,2.89e-05,2.87e-05,2.9e-05,3.14e-05,3...
-0.0023649,-0.0023775,-0.0023901,-0.0024027,-0...	-0.0025327,-0.0025239,-0.00...	0.001737,0.001804,0.001868,0.00193,0.001994...
-6.5e-06,-6.7e-06,-7e-06,-7.2e-06,-7.4e-06,-7.7e-...	-4.53e-05,-4.58e-05,-4.62e-0...	-8.7e-05,-8.6e-05,-8.6e-05,-8.8e-05,-8.9e-05,-9...
-4.5e-06,-4.5e-06,-4.5e-06,-4.5e-06,-4.5e-06,-4.5...	-1.5e-06,-1.4e-06,-1.2e-06,-1...	2.4e-05,2.5e-05,2.6e-05,2.6e-05,2.5e-05,2e-05,...
7.76e-05,7.82e-05,7.87e-05,7.92e-05,7.98e-05,8....	0.000108,0.000108,0.000108...	0,-2e-06,-4e-06,-8e-06,-9e-06,-1e-05,-1.2e-05,-...
0.0006458,0.0006493,0.0006528,0.0006562,0.000...	0.0006959,0.0006935,0.0006...	-0.000455,-0.000479,-0.000504,-0.000523,-0.00...

(b)

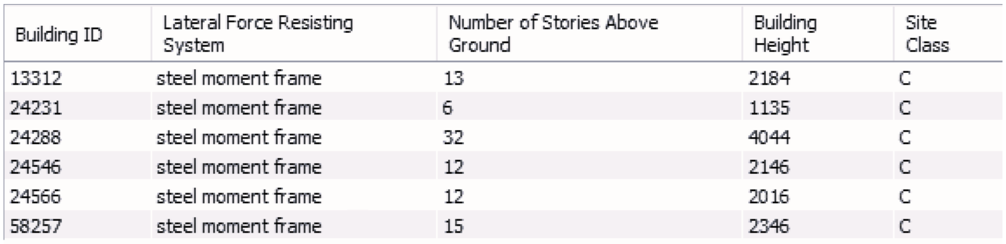
Figure 2.12: Query performed without specifying the exact database label: (a) *MySQL* commands and (b) query results

The JOIN command is particularly useful for simultaneously retrieving data from different tables. Figure 2.13 shows how to obtain data from both the building and site information tables by joining them using the relationship between the tables, i.e., the foreign key, in one table associated with the primary key in the other. In this case, the site ID is a foreign key in the building table and the primary key in the site information table. The AS command is useful for this type of query as it allows the user to modify the column heading. For example, “building.ID AS ‘Building ID’ will label the column as ‘Building ID’ instead of the default label in the database. Figure 2.13b shows the resulting output from the query.



```
SELECT
    building.ID AS "Building ID",
    building.LFRS AS "Lateral Force Resisting System",
    building.num_stories_above AS "Number of Stories Above Ground",
    building.height AS "Building Height",
    site_info.site_class AS "Site Class"
FROM
    building
    JOIN
    site_info ON building.site_ID = site_info.ID
WHERE
    LFRS = 'steel moment frame' AND
    site_class = 'C'
        AND num_stories_above > 4;
```

(a)



Building ID	Lateral Force Resisting System	Number of Stories Above Ground	Building Height	Site Class
13312	steel moment frame	13	2184	C
24231	steel moment frame	6	1135	C
24288	steel moment frame	32	4044	C
24546	steel moment frame	12	2146	C
24566	steel moment frame	12	2016	C
58257	steel moment frame	15	2346	C

(b)

Figure 2.13: Query used to simultaneously retrieve data from different tables: (a) MySQL commands and (b) query results

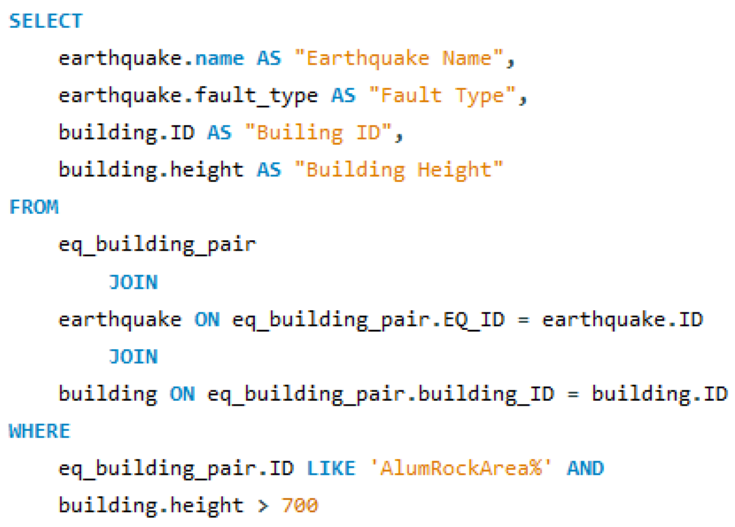
The final, and most involved example shown in Figure 2.14, illustrates how to query data using the junction that relates the earthquake and building tables. The JOIN command links the earthquake ID and building ID into the eq_building_pair IDs. The LIKE command is then used to specify the earthquake, “AlumRockArea”. This will select only the building

IDs associated with the AlumRockArea earthquake because of the defined eq_building_pair IDs. The associated output is shown in Figure 2.14b, where all buildings associated with AlumRockArea taller than 700 inches are displayed with the information specified in the SELECT command.

2.4 Applications and Potential Extensions of the Database

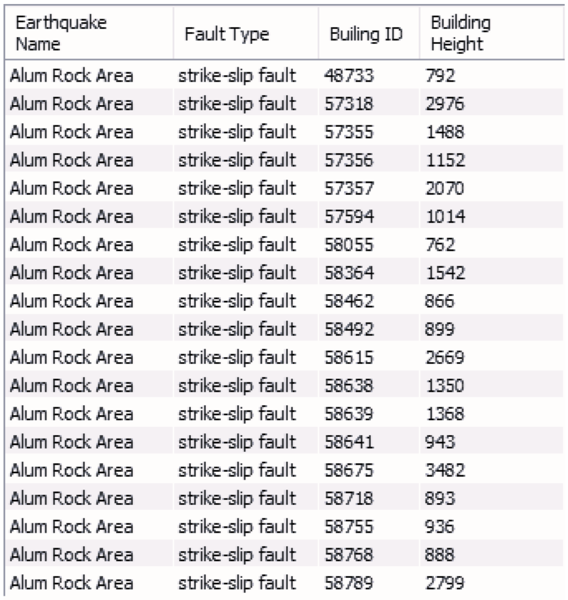
2.4.1 Summary of Prior Applications

[114] used a subset of the data in the database to develop a structural response reconstruction model for an inventory of buildings. A structural response prediction model ($SRPM$) was coupled with the kriging algorithm to estimate building responses for post-earthquake impact assessments. The dataset used for this study did not include the responses from several recent events (e.g., 2019 **M** 7.1 Ridgecrest, 2020 **M** 5.8 Petrolia). The specific aim of the paper was to create a generalized model that can perform post-earthquake prediction of responses in non-instrumented buildings using the measurements from those that are instrumented. The $SRPM$ is a modified version of the GMM developed by [64]. The model is comprised of a combination of ground motion and building response terms. The GMM part of the model is used to predict the median value of the ground motion intensity measure (IM). This prediction is then modified using structural features to predict the $PSDR$ and PFA. Using the responses from the $SRPM$, residuals are calculated as the difference between the predicted and measured $EDPs$. The Kriging algorithm is then used to spatially interpolate the residuals for the uninstrumented buildings. The final prediction is made by taking the median value predicted by the $SRPM$ model and adding the residual estimated from, Kriging. A limitation in this study was the use of a linear $SRPM$ for the initial predictions, as it does not capture the nonlinearity of responses that are expected from most buildings [114]. Another limitation was use of the estimated period based on ASCE 7-10 Equation 12.8-7 (instead of more precise approaches such as system identification) as a feature in the prediction model.



```
SELECT
    earthquake.name AS "Earthquake Name",
    earthquake.fault_type AS "Fault Type",
    building.ID AS "Builing ID",
    building.height AS "Building Height"
FROM
    eq_building_pair
        JOIN
    earthquake ON eq_building_pair.EQ_ID = earthquake.ID
        JOIN
    building ON eq_building_pair.building_ID = building.ID
WHERE
    eq_building_pair.ID LIKE 'AlumRockArea%' AND
    building.height > 700
```

(a)



Earthquake Name	Fault Type	Builing ID	Building Height
Alum Rock Area	strike-slip fault	48733	792
Alum Rock Area	strike-slip fault	57318	2976
Alum Rock Area	strike-slip fault	57355	1488
Alum Rock Area	strike-slip fault	57356	1152
Alum Rock Area	strike-slip fault	57357	2070
Alum Rock Area	strike-slip fault	57594	1014
Alum Rock Area	strike-slip fault	58055	762
Alum Rock Area	strike-slip fault	58364	1542
Alum Rock Area	strike-slip fault	58462	866
Alum Rock Area	strike-slip fault	58492	899
Alum Rock Area	strike-slip fault	58615	2669
Alum Rock Area	strike-slip fault	58638	1350
Alum Rock Area	strike-slip fault	58639	1368
Alum Rock Area	strike-slip fault	58641	943
Alum Rock Area	strike-slip fault	58675	3482
Alum Rock Area	strike-slip fault	58718	893
Alum Rock Area	strike-slip fault	58755	936
Alum Rock Area	strike-slip fault	58768	888
Alum Rock Area	strike-slip fault	58789	2799

(b)

Figure 2.14: Query performed using the junction table: (a) MySQL commands and (b) query results

Portions of the $CSMIP$ (i.e., the original source) data have been used in several other prior studies. The data has been used to estimate damping ratios in different types of building structures [37, 70, 65]. Several system identification studies have also been performed using the $CSMIP$ data. [68, 44, 91]. More recent studies have applied ML algorithms to the building strong motion data to develop response prediction models [126, 125, 67].

2.4.2 Possible Future Applications

The previously mentioned studies (in this section and the Introduction) all have limitations, many related to the type of data used to develop the models as well as the size of the data set itself. Some studies improved upon others in one respect (e.g., using measured data rather than simulated data), but fell short in others (e.g., inability to assess localized damage). One study by [83] was able to address both of these issues, but was limited by a small set of measured data. A common feature of several recent studies is the use of ML to develop the damage models. The limitation of a small data set can have a large impact on the accuracy of ML models, which are dependent on high-quality training data. Most of the prior damage detection studies focused on a single building, but a useful application of SHM technology is the ability to assess the impact to an entire inventory of structures after an event. This would allow for rapid damage assessments, better pre-allocation of resources for likely natural hazard events, and more efficient regional recovery efforts. The database developed in this paper therefore contributes to the development of models that can, with reasonable accuracy, assess the damage to a portfolio of building structures. Much of the existing seismic SHM literature focuses on structural responses in the horizontal direction with comparatively much less studies considering vertical responses. Having a repository of vertical response histories, such as in the developed database, is likely to spur much-needed future research on this topic.

In addition to providing a robust collection of earthquake response data for rapid damage assessments, the database can serve as a template for those that consider other types of hazards (e.g., windstorms and floods) and infrastructure (e.g. bridges). Developing a multi-hazard response database would allow for generalized assessments of building performance damage modeling. Additionally, the database developed in this paper is exclusive to building response data, but expanding it (or creating new ones) with multiple infrastructure types could further improve impact estimation and recovery efforts. The current database can also be expanded to consider other $LFRS$ types. Much of the current literature is focused on predicting responses for a single type of $LFRS$. Expanding the current database to include more $LFRS$ and construction material types would increase the generalizability of data-driven predictive models.

2.5 Summary and Conclusion

A database of historical earthquake response records from instrumented buildings in California is developed in $MySQL$ and hosted on DesignSafe. It includes data from a wide range of building heights, construction types, and materials, to support modeling and analyses. A total of 216 buildings linked to 35 historical events with magnitude 4 or higher are recorded. The buildings range from single-story to more than 70 stories. The time-series responses (i.e., accelerations, velocities and displacements) are organized into MEDIUMTEXT objects in the database. This type of data structure allows for simple extraction of the entire response history for a given event-building pair. A python-based tool is included with the database to facilitate ease of extraction, manipulation, and visualization of the data.

Building response data from structural health monitoring technology has driven innovation in natural hazard engineering. However, most of the prior studies on the use of building seismic strong motion data have focused on individual buildings and specific structure-types. Mutli-event and multi-building (structural system) type databases such as the one developed in the current study can increase the feasibility and generalizability of data-driven models focused on regional impact assessment. The database has been constructed to facilitate dynamic updating, which eliminates the need for tedious data collection efforts and provides an efficient and intuitive medium for curating information. Future efforts should focus on expanding the current database to increase the diversity of events and building construction types that are covered. The database can also serve as a template for others that incorporate other types of hazards (e.g., hurricanes) and infrastructure types (e.g., bridges). Ultimately, a single database that considers different types of hazards and infrastructure would facilitate a unified approach to multi-hazard risk and resilience assessment of communities.

CHAPTER 3

A Dual Kriging-XGBoost Model for Reconstructing Building Seismic Responses Using Strong Motion Data

3.1 Introduction

In the post-earthquake environment, Structural Response Reconstruction (SRR) models aim to estimate the seismic demands in buildings/locations that are not equipped with sensors using the response measurements from instrumented buildings/locations. To date, most of the methods that can be found in the literature have been focused on interpolating measured responses within a single building [85, 75, 58, 121]. Meaning that demands are reconstructed in those parts of the building that don't have sensors using the measurements from the instrumented locations. However, to support rapid post-earthquake assessment at the inventory scale, there is also a need for cross-building reconstruction, where responses measured in an instrumented building inform the demands in those without sensors. The initial studies in this area used responses from numerical simulations [113, 111].

The current study aims to develop a cross-building response reconstruction (CBRR) model that integrates kriging and the extreme gradient boosting algorithm [26] (XGBoost) for reconstructing seismic responses across an inventory of buildings. Using the dataset developed in chapter 2, the dual model is constructed in two phases. In the first phase, kriging is used to interpolate the peak ground acceleration (PGA) at the location of the uninstrumented buildings. The second phase trains an XGBoost model that is used to reconstruct the maximum (over the building height) peak story drift ratio (over the response history) (maxPSDR) and peak floor acceleration (maxPFA) in the uninstrumented buildings.

The performance of the dual model is then discussed, including a detailed analysis of the residuals. The next section provides an overview of the dual model followed by a detailed description of the dataset. The two phases of the model development, (kriging and CBRR) are then discussed. An evaluation of the model performance is presented and the paper concludes with a summary of the findings, the limitations of the study and suggestions for future related work.

3.2 Overview of Dual Model

An overview of the dual model formulation is shown in Figure 3.1. The raw data comprises of acceleration response histories measured at a subset of floors (usually three) in each building. This data is preprocessed to obtain the engineering demand parameters (EDP) of interest, namely, peak floor accelerations (PFA) and peak story drift ratio (PSDR). Note the distinction between the peak response demand at a given floor/story (i.e., PSDR and PFA) and the maximum of the peak response over the height of the building (maxPSDR and maxPFA). The latter represents the demand parameters that are reconstructed in the dual model. The data set comprising the EDPs of interest and model features (or input variables) is subdivided into training and testing sets. The specific details of the split are provided in Section 3.5.

In the first phase of the dual model development, empirical semivariograms are constructed based on the euclidean distance between data points as measured by the event magnitude, source-to-site distance and site separation. Several theoretical semivariograms are fit to the empirical data and the most appropriate one is chosen based on the root mean square error (RMSE) metric. Using the fitted semivariograms, universal kriging is applied to predict the PGA at the sites where structural response demands are to be reconstructed. The performance of the Kriging model is evaluated using the data in the test set.

In the second phase of the dual model development, the XGBoost algorithm uses the PGA estimated in Phase 1 along with the following additional features: the building location, distance from the corresponding event epicenter, number of stories, lateral force resisting system (LFRS) type, event magnitude, average shear wave velocity (Vs30), and the calculated ASCE 7-16 fundamental period. During the training process, a grid-search is performed along with K-fold cross validation, to tune and optimize the XGBoost hyperparameters. To evaluate its performance, the tuned model is then used to predict the maxPSDR and maxPFA on the test sets. The complete details of the dual model development are provided in Section 3.5.

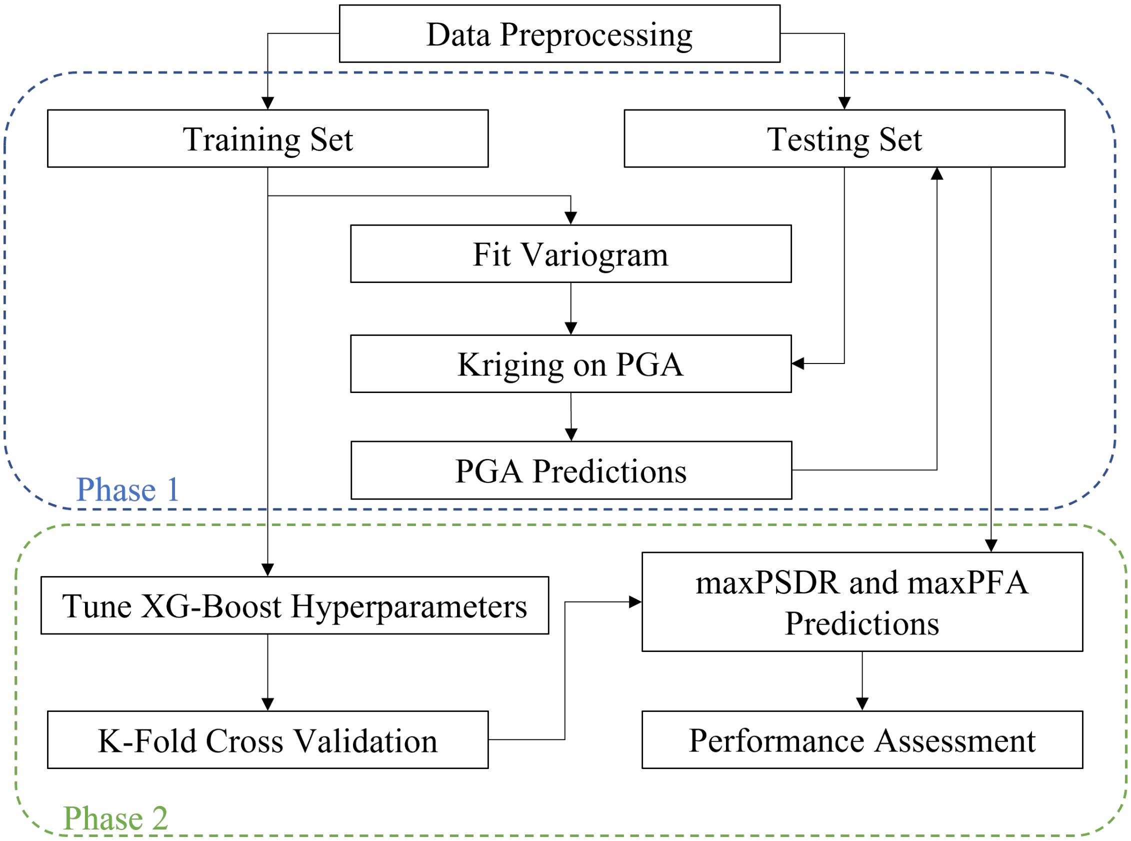


Figure 3.1: Overview of dual model development procedure

Figure 3.2 shows a compact summary of some of the more important features contained in the data set. The off-diagonal plots show interactions between the features and the diagonal plots show the empirical distributions. Note that there is a large skew towards buildings with fewer stories. This is generally consistent with the overall building population and the fact many taller buildings are not equipped with sensors. Some key limitations of this data set are its confinement to a single state (California) and the very limited representation of very tall buildings. Expanding the set to include data points outside of California and a more diverse building inventory will likely increase the robustness of the model.

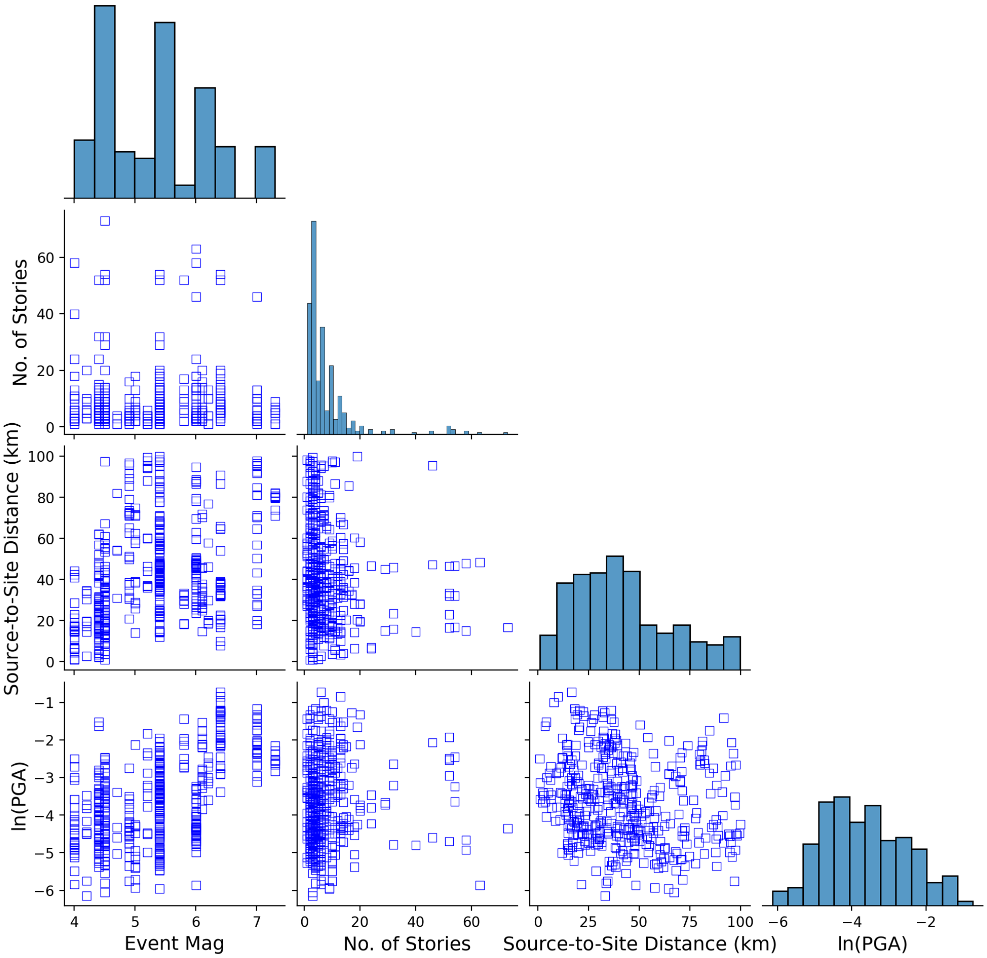


Figure 3.2: Pair plot summary of data features and their interactions

3.3 Kriging

3.3.1 Overview

The initial phase of this study uses spatial interpolation to predict a feature (PGA) that is known to be correlated with maxPSDR and maxPFA (more so the latter). Kriging was originally developed as a spatial interpolation algorithm that makes predictions on new data using a fitted model that represents the similarity between data points. It has since been advanced and applied to problems that include non-spatial features. This method was chosen as PGA is independent of the building in which the data was collected, but rather related to the seismic event itself and the location of the recording. The event magnitude and location of the building are features that would be known at the time of prediction, therefore, they can be used to predict the ground acceleration for a building that is not instrumented. The moderate and high correlation between PGA and recorded maxPSDR and maxPFA values, respectively, is an indication that it is an important feature in the learning model.

Kriging is commonly used on geospatial data sets but can also be used on non-spatial data as it utilizes Euclidian space to determine distance between data points. In other words, if the data display higher response similarity between points that are “close” together (i.e., share similar features) and lower similarity between points that are further apart, kriging may be a feasible strategy. The features used in the kriging algorithm for this study include both spatial and non-spatial data: the geographic location coordinates (spatial), event magnitude, and distance from event epicenter (non-spatial). Depending on which variation of the algorithm is being utilized, there are assumptions that are made about the data. Some of the most commonly used forms are simple, ordinary, and universal kriging. The general assumptions of kriging are that the data follow a stationary process (can be relaxed depending on which form is used) and a normal distribution (for best performance, but not required), and there are no trends. A “stationary” process refers to a constant trend (mean) and variance in a stochastic process. Simple kriging assumes a stationary process in which the mean and the variance are constant known quantities. This is often useful when analyzing residuals that are normal around 0. Ordinary kriging also assumes a stationary process, but the mean is instead an unknown constant. Universal kriging is the most relaxed. It utilizes a “weak stationarity” assumption, where the unknown mean follows some polynomial trend, and the variance is constant across the spatial field. If the raw data do not fit the criteria, data transformations and trend removal methods may be necessary to optimize the performance of kriging.

Kriging fits the data to either a semivariogram or covariance function (also known as a covariogram). Each function models the spatial correlation based on the distance between data points. According to these models, points that are closer together have higher correlation, which decreases as they become further apart. The commonly used spherical and powered-exponential covariance functions are defined in Equations 3.1 and 3.2, respectively.

$$C\Big(h, r, n\Big) = \Big(C(0) - n\Big)\Big(1 - \frac{3h}{2r} + \frac{h^3}{2r^3}\Big) \tag{3.1}$$

$$C\Big(h, r, n\Big) = \Big(C(0) - n\Big)e^{-3\left(\frac{h}{r}\right)^p} \tag{3.2}$$

$C(h)$ is the covariance between data points at separation distance, h. $C(0)$ is the total variance or the “sill”, r is the range, and n is the nugget. The sill is a term for the value that the variance approaches as the distance between points become larger and where there is no longer spatial correlation in the data. The range is the separation distance at which the sill is reached, and the nugget is a term used to account for the measurement error. A nugget value greater than 0 implies there is some non-zero difference in the values for observations at the same location. In Equation 3.2, the exponential covariogram is defined by $p = 1$ and the Gaussian covariogram is defined by $p = 2$. Kriging algorithms are often modeled using semivariograms rather than the covariogram. The relationship between the semivariogram and covariogram is given by:

$$\gamma(h) = C(0) - C(h) \tag{3.3}$$

Where γ is the semivariogram model. It should be noted that not all semivariograms have a corresponding covariogram. Examples of this are the linear and power models. Neither model has a sill, so the variance increases infinitely. For this study, the root mean squared error (RMSE) for each empirical semivariogram was calculated after being fit to the data. As shown in Figure 3.6, the data was fit to multiple covariance functions to select the most suitable model. The kriging procedure in this paper will use the covariogram for the derivations for consistency with the utilized model. The semivariogram-based formulation can be obtained by substituting Equation 3.3 into the one based on the covariogram.

3.3.2 Formulation

Universal kriging is used in this study as it makes fewer assumptions about the data than simple or ordinary kriging. It also allows for local trends to vary by location, which is important because, as shown in Figure 2.2, the data is primarily clustered around two distinct regions. Consider a stochastic process defined by random variable $Z(s)$, where s_i is the spatial location (or non-spatial feature) of data point i and $Z(s_i)$ is the corresponding variable value of interest (PGA in this study). Kriging uses a weighted average of the known PGA values to predict the unknown value, $\hat{Z}(s_0)$. The weights, w_i, are based on spatial similarity of the known data to the unknown data point, s_0.

$$\begin{gathered} \hat{Z}(s_0) = \sum_{i=1}^{n} w_i Z(s_i) \\ s.t. \sum_{i=1}^{n} w_i = 1 \end{gathered} \tag{3.4}$$

The objective function is defined as the minimization of the squared error between the predicted and observed values.

$$\begin{aligned} \min \sigma_e^2 &= E\Big[Z(s_0) - \hat{Z}(s_0)\Big]^2 \\ &= E\Big[Z(s_0) - \sum_{i=1}^{n} w_i Z(s_i)\Big]^2 \end{aligned} \tag{3.5}$$

Equation 3.5 can be rewritten using covariance,

$$\min \sigma_e^2 = C(0) - 2\sum_{i=1}^{n} w_i C(s_0, s_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j C(s_i, s_j) \tag{3.6}$$

To consider the constraint in Equation 3.4, the optimization problem becomes

$$\min \sigma_e^2 = C(0) - 2\sum_{i=1}^{n} w_i C(s_0, s_i) + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j C(s_i, s_j) - 2\lambda\left(\sum_{i=1}^{n} w_i - 1\right) \tag{3.7}$$

where λ is a lagrange multiplier for the constrained optimization problem. Differentiating with respect to w_i and λ and solving the minimization by setting equal to 0 gives

$$\frac{\partial \sigma_e^2}{\partial w_i} = \sum_{j=1}^{n} w_j C(s_i, s_j) - C(s_0, s_i)\lambda = 0, \quad i = 1, \ldots, n \tag{3.8}$$

$$\frac{\partial \sigma_e^2}{\partial \lambda} = \sum_{i=1}^{n} w_i - 1 = 0, \quad i = 1, \ldots, n \tag{3.9}$$

This can be written in matrix form and rearranged to solve for the weights and Lagrange multiplier

$$\begin{gathered}\mathbf{CW} = \mathbf{c} \\ \mathbf{W} = \mathbf{C}^{-1}\mathbf{c}\end{gathered} \tag{3.10}$$

$$\begin{pmatrix} w_1 \\ w_2 \\ \vdots \\ w_n \\ -\lambda \end{pmatrix} = \begin{pmatrix} C(s_1, s_1) & C(s_1, s_2) & \ldots & C(s_1, s_n) & 1 \\ C(s_2, s_1) & C(s_2, s_2) & \ldots & C(s_2, s_n) & 1 \\ \vdots & \ldots & \ddots & \ldots & 1 \\ C(s_n, s_1) & C(s_n, s_2) & \ldots & C(s_n, s_n) & 1 \\ 1 & 1 & 1 & 1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} C(s_0, s_1) \\ C(s_0, s_2) \\ \vdots \\ C(s_0, s_n) \\ 1 \end{pmatrix} \tag{3.11}$$

Once w and λ are computed, the equation for the error variance, σ_e^2, can be simplified using the differentiation in Equation 3.8. Multiplying the derivative by w_i, the equation becomes:

$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j C(s_i, s_j) - \sum_{i=1}^{n} w_i C(s_0, s_i) - \sum_{i=1}^{n} w_i \lambda = 0 \tag{3.12}$$

Equation 3.12 can be simplified through substitution into the error variance equation

$$\sigma_e^2 = C(0) - \sum_{i=1}^{n} w_i C(s_0, s_i) + \lambda \tag{3.13}$$

The DiceKriging R package [99] was used for model training and performance assessment.

3.4 Gradient Boosting

The second phase of the dual formulation develops the machine learning model that is used to make the response predictions. Gradient Boosting (GB) is a supervised learning technique that uses gradient descent to optimize the predictions of the response variable(s). In other words, GB fits a new model to the residuals of a previous model, taking a step in the direction that minimizes the error. This process continues until the change in prediction error between the current and previous models (the gradient) reaches a minimum. Extreme Gradient Boosting (XGBoost) is a variation of the GB technique that is modified to prevent overfitting while maintaining performance. This is achieved by including regularization terms. GB was selected not only for its performance, but because it is a highly customizable learning algorithm that can be adapted to the needs of the study.

To provide further insight into the inner workings of gradient boosting machines (GBMs), decisions trees must be introduced. GB is an ensemble learning method that sequentially combines a set of weak learners to obtain a final model. Each new learner improves upon the previous one and the final model performs better than any of the individual weak learners. Typically, this is done using decisions trees, where the tree leaves are assigned a prediction score and a weight. These weighted scores are summed for each input x to obtain a final prediction score. Decision trees can be utilized for both classification and regression (CART). As the response variables are continuous, regression trees were used to develop the model predictions in the current study. Given a dataset $D = \{(\mathbf{x}_i, y_i)\}(\|D\| = n, \mathbf{x}_i \in R^m, y_i \in R)$, where n represents the number of examples, and m is the number of features, the following equation represents a tree ensemble used to make a prediction for observation i, $\hat{y}_i$, as a function of x_i. The goal is to approximate the functional dependence between x and y.

$$\hat{y}_i = F(\mathbf{x}_i) = \sum_{i=1}^{K} f_k(\mathbf{x}_i), \quad f_k \in F \tag{3.14}$$

where K denotes the total number of decisions trees, and f_k defines decision tree k. The space F is defined by the trees and their corresponding leaves, which is given by $F = \{f(\mathbf{x} = w_{q(\mathbf{x})}\}(q : R^m \to T, w \in R^T)$. Q is a function that describes the structure (or the rules) of each tree, which has weights (or scores), w_i, for leaf i. T is the number of leaves in the tree.

Using this prediction method, an objective function is defined that tracks the performance of the model. This function includes a loss function and a regularization term. The goal is to minimize this objective function, and in turn, the loss, in order to optimize the predictive capability of the model. The objective function can be written as:

$$\begin{gathered} L\Big(F(\mathbf{x})\Big) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ \text{where} \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda \parallel w \parallel^2, \\ l(\hat{y}_i, y_i) = (\hat{y}_i - y_i)^2 \end{gathered} \tag{3.15}$$

The loss function is an error function that is used to compare the predicted value $\hat{y}_i$ to the target (y_i). In this study, the commonly used squared error is set as the loss objective. The regularization terms control the model complexity, which helps to reduce overfitting. The hyper-parameters used to construct the model include the type of regularization (e.g., L_2, L_1), the learning rate, and the maximum step size. L_2 regularization (represented by the equation above) is used in conjunction with a pre-defined learning rate and step size.

With the objective function defined, GB iteratively creates new trees, adding the tree (f_t) that minimizes the loss to the ensemble. This is one of multiple methods used in GB, called the exact greedy algorithm, detailed in [26]. This algorithm is effective for the relatively small dataset in this paper, but for larger data sets, an approximate algorithm may provide better performance. Equation 3.15 is rewritten to represent the iterative process, where t represents the current iteration. So, the objective function to be minimized is:

$$L^{(t)} = \sum_{i=1}^{n} l\bigg(y_i, \hat{y}_i^{t-1} + f_t(\mathbf{x}_i)\bigg) + \Omega(f_t) \tag{3.16}$$

Calculating every possible tree structure at each iteration and comparing the loss to find the best performing one is computationally inefficient. Instead, GB creates trees starting with a leaf node and expanding the tree based on the calculated losses at each split. It then uses a second-order approximation to find a local minimum of the loss function over all trees at the current iteration, $f_t(\mathbf{x})$. Here, g_i and h_i define the first and second order gradients of the loss.

$$\begin{gathered}\tilde{L}^{(t)} = \sum_{i=1}^{n} \bigg[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i)\bigg] + \Omega(f_t) \\ \text{where} \quad g_i = \frac{\partial \hat{y}_i^{t-1}}{\partial l(y_i, \hat{y}_i^{t-1})} \quad \text{and} \quad h_i = \frac{\partial \hat{y}_i^{t-1}}{\partial^2 l(y_i, \hat{y}_i^{t-1})}\end{gathered} \tag{3.17}$$

Expanding the regularization function, $\Omega(f_t)$, and approximating the optimal leaf weights using Equation 3.18, the simplified scoring function for a tree, q, is given in Equation 3.19, and the resulting change in loss calculation for each split is given in 3.20.

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda} \tag{3.18}$$

$$\tilde{L}^{(t)}(q) = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T \tag{3.19}$$

$$\begin{gathered}L_{split} = \frac{1}{2}\Bigg[\sum_{j=1}^{T} \frac{\left(\sum_{i \in I_L} g_i\right)^2}{\sum_{i \in I_L} h_i + \lambda} + \frac{\left(\sum_{i \in I_R} g_i\right)^2}{\sum_{i \in I_R} h_i + \lambda} + \frac{\left(\sum_{i \in I} g_i\right)^2}{\sum_{i \in I} h_i + \lambda}\Bigg] - \gamma \\ \text{where } I_j = \{i | q(\mathbf{x}_i) = j\} \text{ is the instance set of leaf } j \\ \text{and } I_L = I \ \text{ at the left split and } \ I_R = I \text{ at the right split}\end{gathered} \tag{3.20}$$

The overall additive ensemble that defines the model for the entire data set can be written as

$$F^{(t)}(\mathbf{x}) = F^{(t-1)}(\mathbf{x}) + \min_{f_t \in F} \tilde{L}^{(t)} \tag{3.21}$$

Figure 3.3 shows an example of one of the trees that was developed from training the XGBoost model. The leaf values are the predictions that a particular decision tree is making for a given data point. The final prediction value for a data point is the sum of all of the tree predictions at that point.

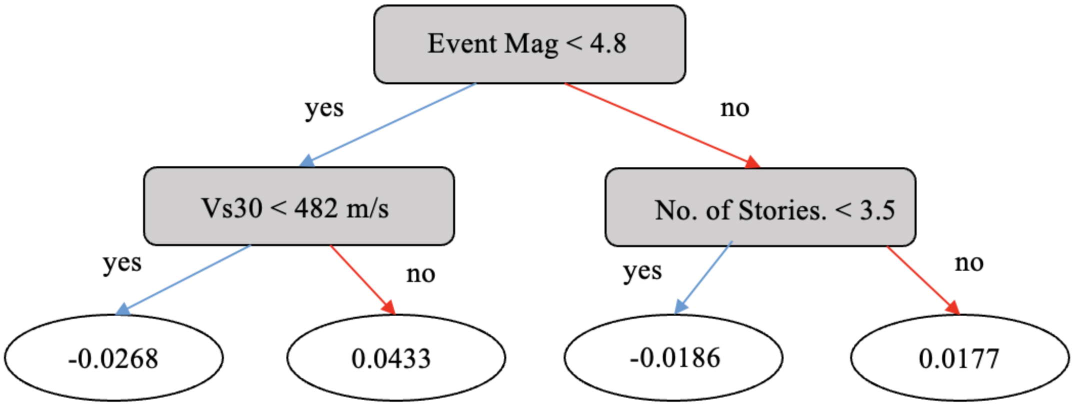


Figure 3.3: Example of a decision tree from maxPFA Gradient Boosting model

3.5 Dual Model Development

The training set consists of both categorical and numerical data, which presents a problem for many machine learning models as they can only interpret the latter of the two. Two methods of conversion from categorical to numerical values are considered. Target encoding replaces the categorical information with the average target value for the data points within the same category. For example, the LFRS-type categorical variable for all data points with reinforced concrete (RC) frames is replaced with the corresponding mean maxPSDR or maxPFA . The benefit of this method is that it maintains the number of training features, while still assigning different numerical values to the data based on the category. This should be done carefully, where the target encoding is performed separately for each of the training and testing data sets to prevent data leakage from the test data into the trained model. Alternatively, one-hot encoding involves replacing the categorical variables with an $n \times c$ matrix of ones and zeros, where n is the number of data points in the set and c is the number of categories. A "1" is placed at the position corresponding to the category associated with the current data point. The remaining values in the row will be zero. This method increases the overall number of training features by $c - 1$, so this is best used when there are a relatively small number of categories, otherwise the trade off of increased model complexity may outweigh the benefits. This method also avoids any artificial scaling of category importance since all categories are given the same value of 1.

One-hot encoding was ultimately chosen since the additional features are easily handled by the XGBoost algorithm. The encoding is performed on the LFRS feature to convert the categorical data to numerical values. For this data set, there are 6 LFRS-types: concrete shear wall, concrete moment frame, steel moment frame, masonry shear wall, wood frame, and steel braced frame. Note that some buildings use a combination of LFRS-types, and in those cases, two or more of the columns in the one-hot encoded vector have a 1 instead of a 0.

For developing and evaluating the model performance, a 70-30 train-test split is used. Only buildings within 100km of the corresponding event epicenter are considered, as the PGA trend dissipates at larger distances. This is especially important because the kriging model relies on the spatial correlation of the PGA values. The training and testing sets are developed using stratified sampling by event to ensure that the ratio of data points from each event is the same across both sets. This gives a more accurate representation of how the model would perform on new data as simple random sampling may result in a different training distribution than would be used in practice. The model is meant to be completely "event-agnostic". The model is not fit individually to the data for each event, rather to the entire set, so in instances where a new event occurs and there is very limited data, it can leverage data from other events with similar features and buildings to make predictions.

The approach used in [114] performed kriging on the SRPM residuals. As such, only data from the other buildings involved in the same event are used to estimate the residuals. The model presented in this paper takes a different approach. Rather than depending only on the spatial relationship between data points from the same event, buildings from other events are also included. This approach increases the generality of the model because it can also deal with events that have a limited number of recordings from instrumented buildings. The response values for data points involved in other events may share similar features and, in turn, similar EDPs. Whereas, for a single event with a small sample of recorded responses, some of them may appear to be outliers. However, when expanded into a larger data set, they actually fall within the expected range. Compared to the [114] model, the one developed in this paper is less dependent on having an ample number of instrumented buildings for the event of interest.

3.5.1 Predicting Peak Ground Acceleration

As shown in Figure 3.4, PGA is predictably found to be highly correlated with maxPFA and moderately correlated with maxPSDR. As such, it was determined that including PGA as a feature would be beneficial to the model. To do this, the model must be able to estimate the PGA at any given site location for an uninstrumented building. Since PGA is not a structural characteristic, but dependent on the seismic event itself as well as the location of the building with respect to the event epicenter, the recordings obtained from instrumented structures can be used to estimate the ground acceleration in various locations for those without sensors. The training set was used to construct the kriging model and estimate the PGA values for the test set.

As noted earlier, the following key assumptions are associated with kriging: stationarity, normality, and lack of trend. Because universal kriging is being used, the stationarity assumption can be relaxed. To address the normality assumption, Figure 3.5 shows a histogram and quantile plot for ln(PGA) . It is seen that the ln(PGA) follows an approximately normal distribution with the histogram having a bell shape and the quantiles falling mostly along the diagonal. It is also worth noting that the normality assumption is often mistaken

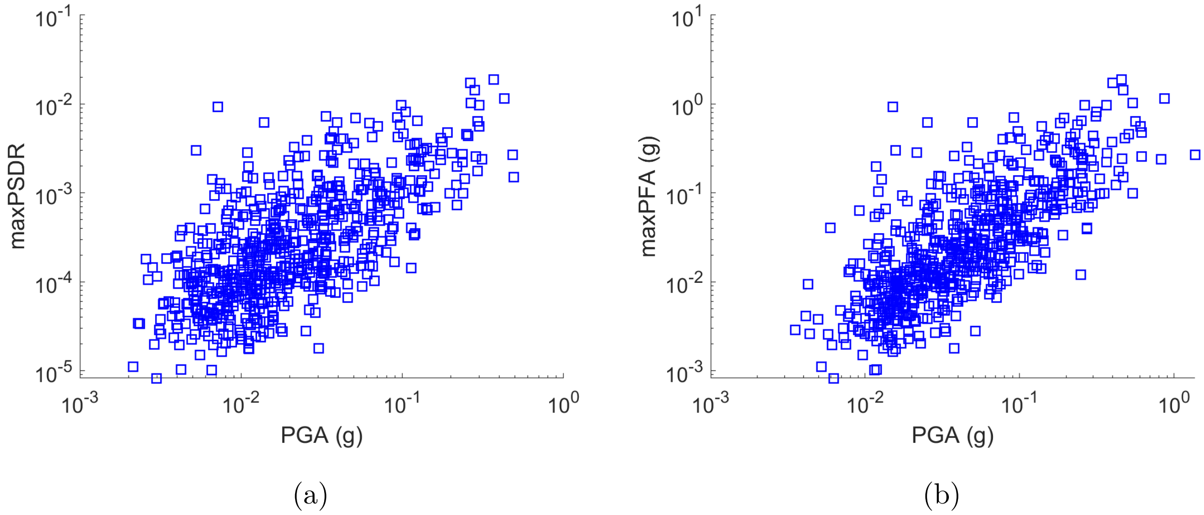


Figure 3.4: Scatter plots for (a) PGA vs maxPSDR with a correlation coefficient $\rho = 0.675$ and (b) PGA vs maxPFA with $\rho = 0.921$

as a strict requirement and that if the data do not fit the normal distribution, then kriging cannot be used. This is not entirely true, as is discussed by [13]. The use of kriging as a predictor does not necessarily require the data to be normal. Rather, for the model to be the best linear unbiased estimator (BLUE), the data should be normal. In other words, the data being normal generally improves the overall performance of the algorithm.

The main benefit of including PGA as a feature is the lower dispersion of the errors in the XGBoost model. Also, as shown in Figure B.1, PGA is one of the most important features in the Boosting model for maxPSDR and maxPFA predictions. Since the test data is a subset of the full dataset, some recorded PGA values were used to assess the kriging model's performance. Once the kriging model is determined to perform reasonably well, the PGA values in the test set are replaced with the predicted values. This is because for a new data point or uninstrumented building, the PGA value would be unknown and must be predicted. The original PGA values were initially maintained simply for assessing the kriging performance. Replacing the PGA values with the predicted ones from kriging provides a more accurate representation of how the model would perform on new data.

Kriging relies not only on correlation among the dependent and independent variables,

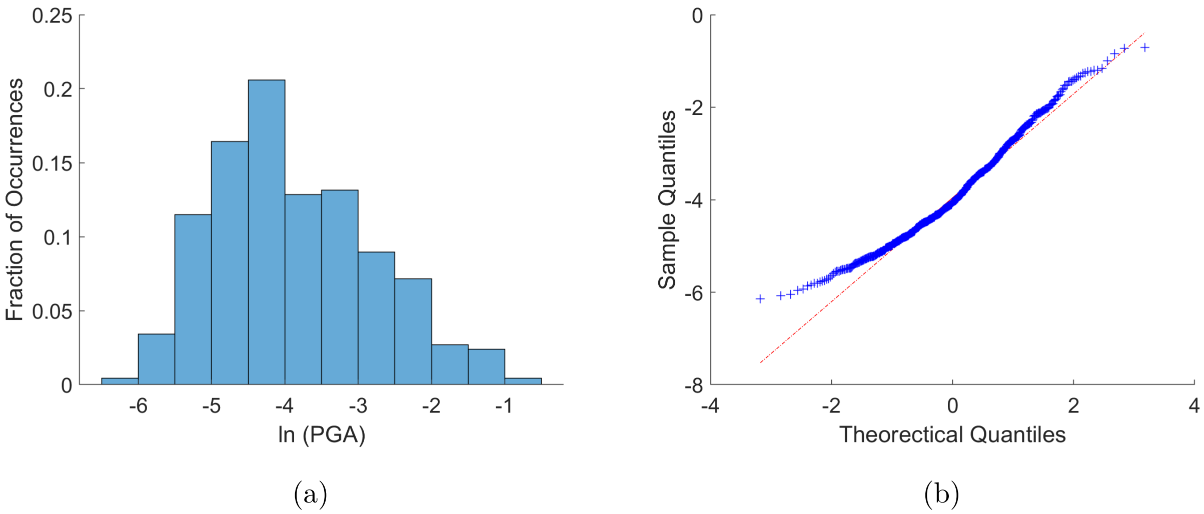


Figure 3.5: (a) Histogram and (b) normal quantile plot of the log-transformed PGA values

but on autocorrelation of the error. In other words, the errors are dependent on separation distance, not exact location. PGA is a function of the event magnitude, geographic conditions, and distance from the event where the measurement is being taken. Data points with similar features should, according to kriging, display similar response values. This can be seen in the empirical variogram plots presented in Figure 3.6 that are fit to the data for the purposes of determining which would be the most suitable one. The performance of the different variograms are compared through visual inspection and the root mean square error (RMSE). Figure 3.6 shows that the Stable, Cubic, Matern and Gaussian variograms have similar fit with RMSE = 0.16.

The kriging model is fit using the gaussian covariance function for its performance and familiarity. In the literature, kriging for PGA is often performed on the residuals rather than on PGA itself due to the assumption of residual normality. A study by [63] provides a thorough explanation on how ground motion models implement kriging into their prediction of PGA residuals. The current study uses building-independent features as the independent variables since the PGA does not rely on building characteristics. Specifically, the event magnitude, site distance, and the location coordinates are used as features. In Figure 3.2, a positive trend is observed between event magnitude and PGA, and a negative trend is

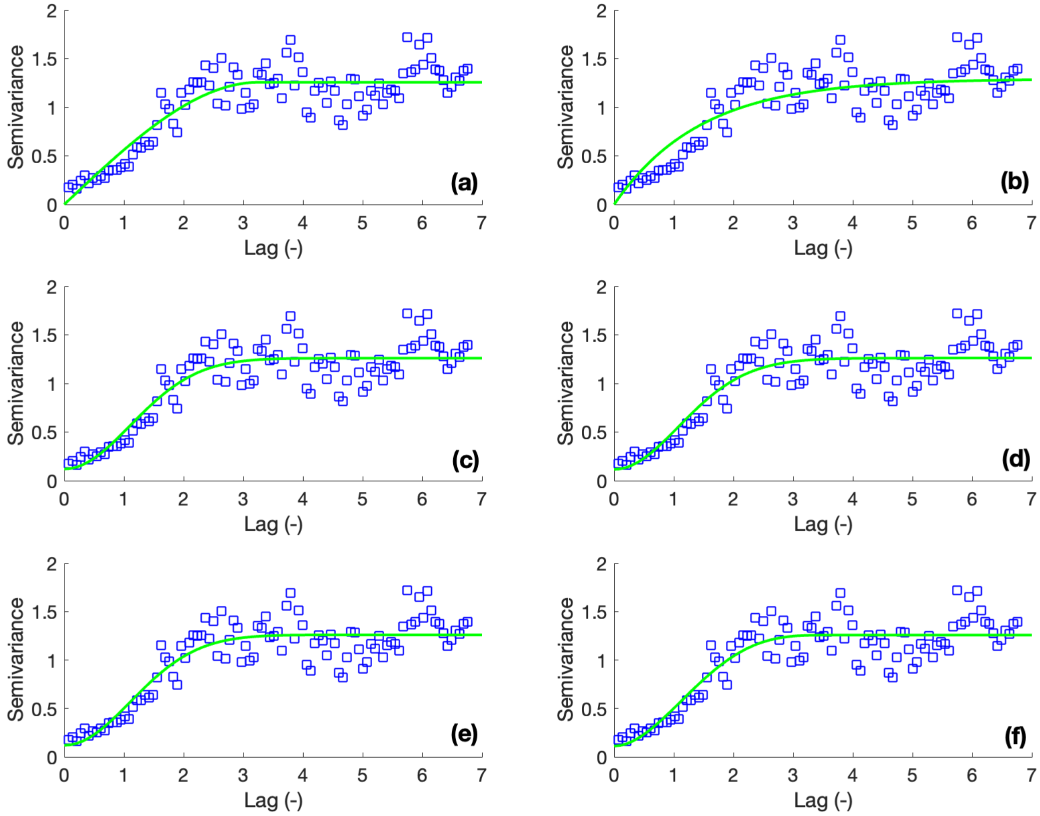


Figure 3.6: Fitted empirical semivariograms for PGA (a) Spherical (RMSE = 0.17), (b) exponential (RMSE = 0.18), (c) gaussian (RMSE = 0.16), (d) matern (RMSE = 0.16), (e) stable (RMSE = 0.16), (f) cubic (RMSE = 0.16)

observed between the site distance and PGA. Vs30 was thought to be a likely candidate, but was found not to improve the model. This may be due to the Vs30 being based on a mixture of measured and inferred data (according to the CESMD description), so it is subject to a higher level of uncertainty than the corresponding measured features. Nevertheless, the model performed well, with a mean-normalized root mean squared error (NRMSE) of 0.139. A leave-one-out analysis is also performed and plotted to show the match between the fitted and measured values and the residuals are plotted to verify normality (see Figure 3.7). To ensure that the relationship between the PGA and EDPs is maintained, the correlation coefficients between the original PGA and test EDPs and the predicted PGA and test EDPs are calculated and compared. The original test set coefficients are 0.618 and 0.896 for PSDR and PFA, respectively. The coefficient between the predicted PGA and test EDPs are 0.665 and 0.765, respectively, which are similar to the original values.

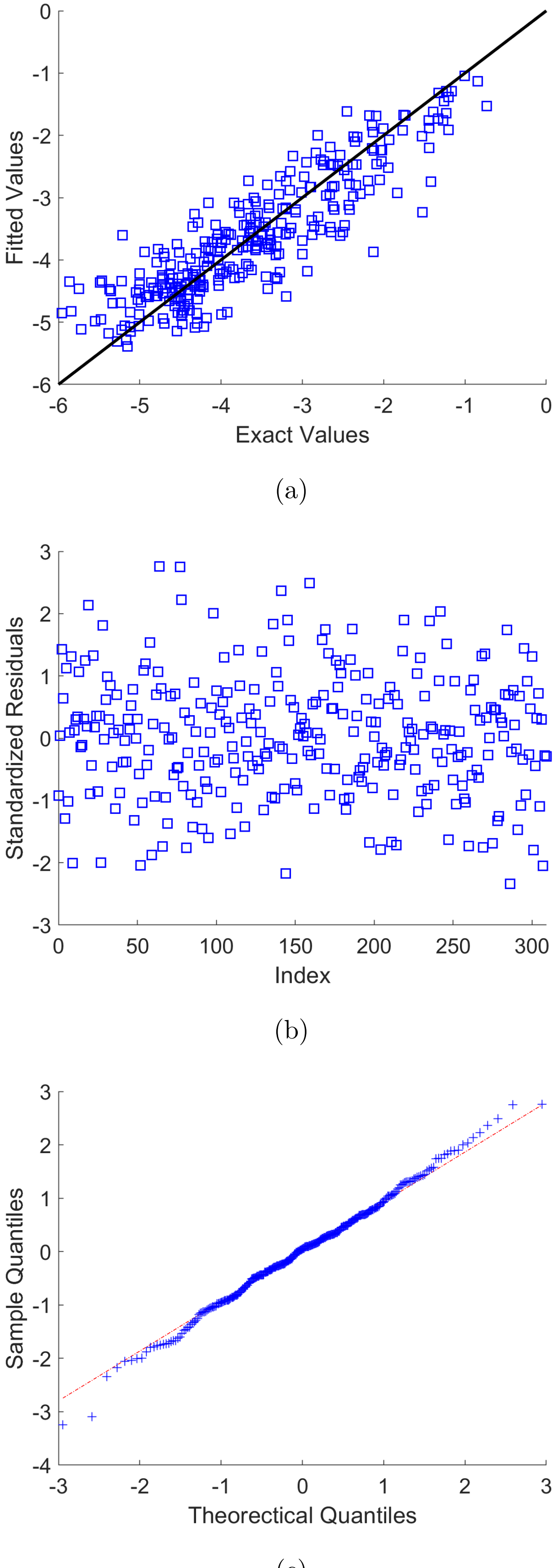
After using the kriging model to estimate the PGA values on the test set and analyze the performance, the real PGA values in the test set are replaced with the predictions. This ensures that when the XGBoost model is developed, it is the same as if an entirely new set of PGA values are being estimated.

3.5.2 Phase II - Gradient Boosting

The second phase of the dual formulation is training the XGBoost model that will be used to reconstruct the maxPSDR and maxPFA values for the test set. To evaluate the efficacy of the first phase of the dual formulation, two XGBoost models are developed i.e., with and without the inclusion of PGA as a feature. A grid search is performed to optimize the parameters for the mean squared error from the model. The optimized parameters are summarized in Table 3.1. The max depth parameter limits the number of splits within a tree to prevent overfitting. The learning rate shrinks the feature weights in each step by the defined factor. Gamma specifies the loss reduction needed for a split to occur. Lambda is an L_2 regularization parameter and the minimum child weight specifies the smallest number of samples required for a node.

To determine if the recorded features are useful in the prediction model, the F-score of each feature is computed. The F-Score is a numerical representation of how much influence each feature has on the model. XGBoost automatically ignores any features that do not contribute to the tree predictions, which is why the maxPFA importance plot has fewer feature scores than the maxPSDR plot. Figure B.1 shows plots of the feature importance for both maxPFA and maxPSDR. According to these plots, the displacement-based response is more sensitive to the LFRS than the acceleration-based response.

After fitting the models, the R^2 values are calculated to determine if the model fit the data well. The training set R^2 values are 0.888 and 0.863 for maxPSDR and maxPFA,

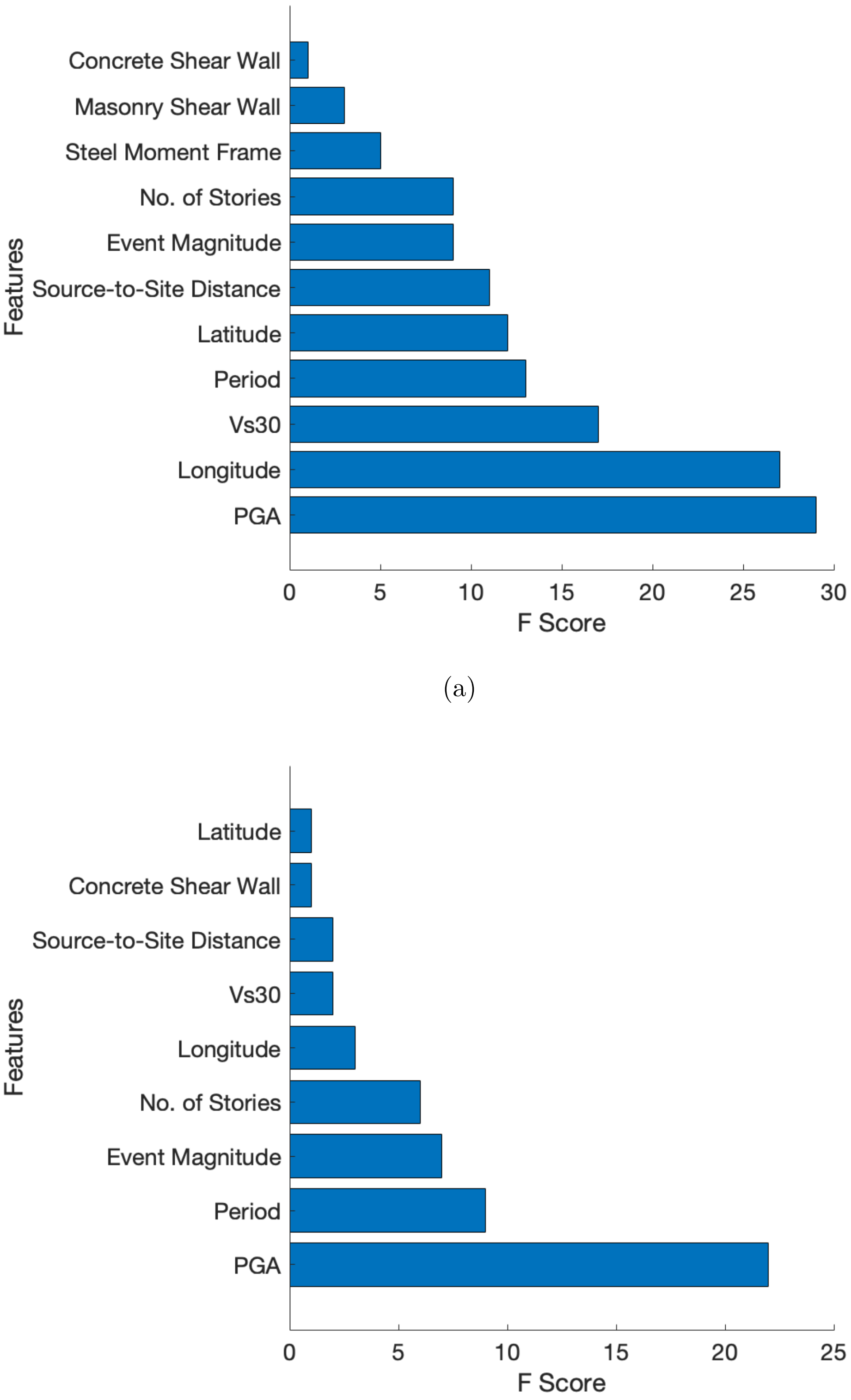


(a)

(b)

(c)

Figure 3.7: (a) Kriging Leave-One-Out Analysis, (b) standardized residuals and (c) residuals normal quantile plot



(a)

(b)

Figure 3.8: Feature importance scores for (a) maxPSDR and (b) maxPFA models

Table 3.1: Grid-search optimized parameters

Parameter	**Range**	**maxPSDR**		**maxPFA**	
		w/ PGA	w/o PGA	w/ PGA	w/o PGA
max_depth	[1,2,3,4,5]	2	5	2	5
learning_rate	[0.01,0.1,0.3]	0.3	0.1	0.3	0.3
gamma	[0,0.25,0.5,1]	0.5	0.25	0.25	0.25
lambda	[0,0.1,0.5,1]	1	1	0	0.1
min_child_weight	[0,1,2,3,5]	5	0	5	3

respectively. With a max possible value of 1, this shows a good fit for the models. As a comparison, the adjusted R^2 values considers how many features are being used, so if there are any unnecessary features, it will be penalized and be significantly smaller than the traditional R^2 which improves with more features regardless if the feature is useful. The adjusted R^2 values are 0.885 and 0.858 for maxPSDR and maxPFA respectively, which is very close to the traditional R^2. This indicates that the XGBoost algorithm successfully removed the unnecessary features.

Due to the relatively small number of data points, rather than having an additional validation set, a five k-fold cross validation (KFCV) repeated 50 times is performed on the training set. In KFCV, the data is split into k sets (folds). One of the folds is a “hold-out” or validation set, that is used for assessing the model performance. The remaining k-1 folds are used to train the model. This continues with each fold being used as the hold out set and the performance score is calculated as the mean score for that iteration. This process is repeated a specified number of times. More repeats typically leads to more consistent performance metric values. For maxPSDR, the RMSE over the 50 repeats is 0.84 with a standard deviation of 0.08 and maxPFA has a RMSE of 0.42 with a standard deviation of 0.05.

3.6 Results

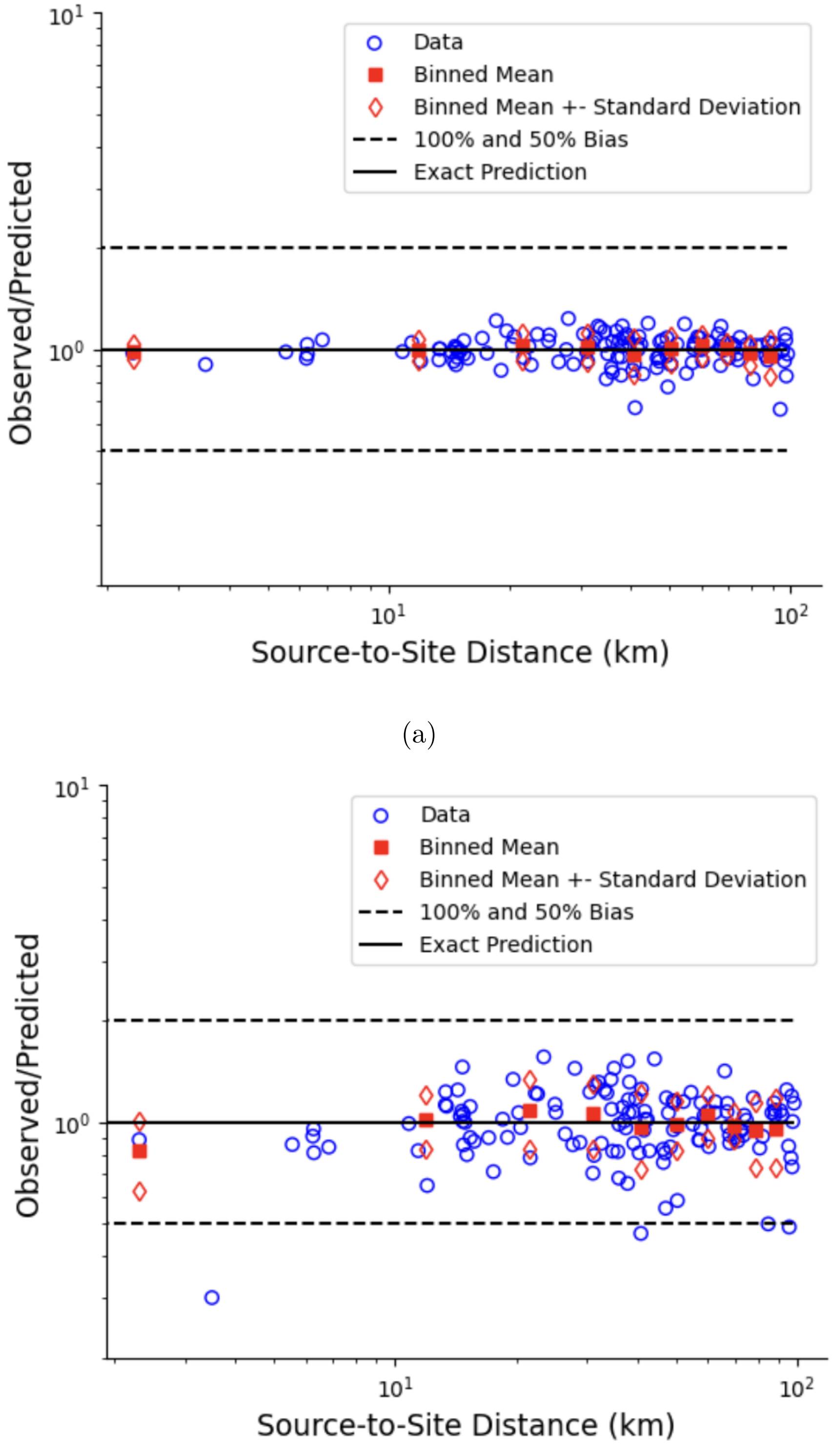
Figure 3.9 and Figure 3.10 show the observed/predicted ratios plotted against the Source-to-Site Distance for each data point. The ratios corresponding to 2 and 0.5 are indicated by the dotted lines. A majority of the data fall within these limits indicating good overall predictive performance. The median observed/predicted ratios for maxPSDR and maxPFA are 1.003 and 1.006, respectively, which shows that the model is unbiased. The absence of a trend between the observed/predicted ratios and the source-to-site distance is also a desirable property. The histogram in Figure 3.11 shows the residual distributions for both quantities of interest, which are centered around 0.

Another metric that can be used to evaluate the model performance is the absolute value of the mean normalized error/residual (MNE) which is computed in Equation 3.22.

$$MNE = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \tag{3.22}$$

where y_i and $\hat{y}_i$ are the observed and predicted responses (i.e., maxPSDR and maxPFA), respectively and N is the number of observations. The dual model achieved MNE values of 0.078 (maxPSDR) and 0.196 (maxPFA), indicating better performance for the former. To quantify the effect of including the PGA obtained from kriging as a feature in the reconstruction model, we compare the mean and standard deviation of the residuals of the "with-PGA" and "without-PGA" models. For PSDR, the mean residual with and without the inclusion of PGA is 0.015 and -0.04, while the standard deviation is 0.797 and 0.795, respectively. This result indicates that the inclusion of the PGA obtained from kriging as a feature in the XGBoost model slightly reduces bias and maintains dispersion, but did not significantly impact its ability to predict maxPSDR. For maxPFA with and without PGA being included as a feature, the residual had means of -0.008 and -0.08 and standard deviations of 0.56 and 0.62, respectively. These results suggest that the inclusion of PGA has a measurable effect on the model bias and accuracy when predicting maxPFA.

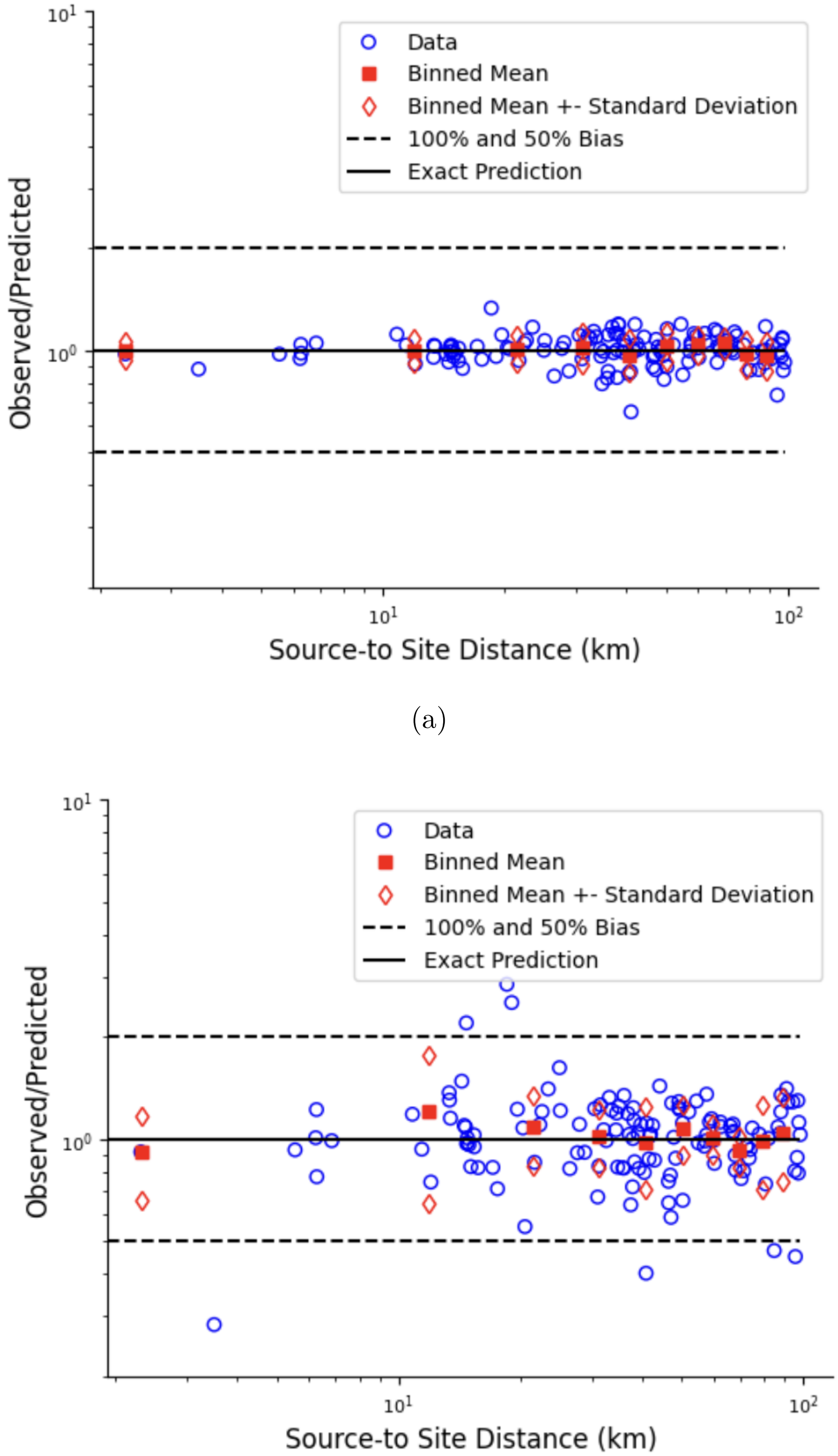
The inclusion of PGA provides a decrease in the residual dispersion for the maxPFA



(a)

(b)

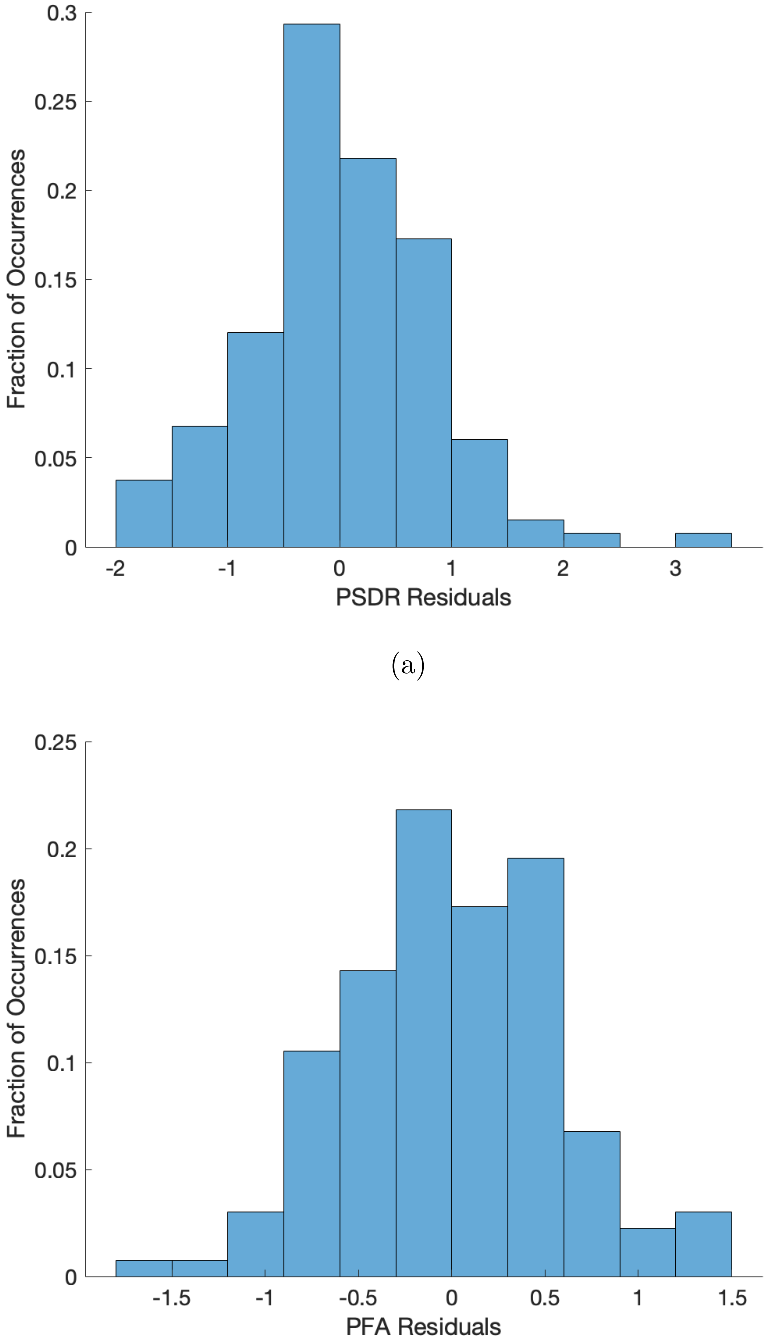
Figure 3.9: Observed/Predicted ratio vs source-to-site distance for model with PGA (a) maxPSDR and (b) maxPFA



(a)

(b)

Figure 3.10: Observed/Predicted ratio vs source-to-site distance for model without PGA (a) maxPSDR and (b) maxPFA



(a)

(b)

Figure 3.11: Histograms showing the residuals for (a) maxPSDR and (b) maxPFA

predictions. This is important when reconstructing uninstrumented building responses for events that have very few recordings, which is often a limitation in other models that depend on within-event residual distributions. Overall, the XGBoost model achieved good predictive performance, but was limited by how well the kriging model performed on estimating PGA. Unlike the one developed in [114], the model presented in this paper is capable of reconstructing EDP's in regions with a low number of instrumented buildings. However, more complex estimation methods can be explored to better estimate and minimize the error propagation from phase 1 to phase 2 of the model.

Also of note are the multiple sources of uncertainty in the data that, in future studies, can be analyzed to determine the effect that individual features have on the outcome. Sensitivity analyses utilizing a distribution of variables that have higher levels of uncertainty due to measurement error or were estimated by various means could illustrate whether the uncertainty has a significant effect on the model. For example, the Vs30 obtained from CESMD are a combination of measured and inferred values, so the data are subject to measurement error as well as misspecification. The Vs30 values could be obtained from a distribution while holding all other features constant and predicting the outcome to show how the predictions change over the adopted range. This could also be done on other features, such as the fundamental period, which was estimated using the ASCE 7-16 equation (ASCE 2017), or even the PGA, which was calculated by spatial interpolation. The earthquake engineering literature on uncertainty quantification (UQ) is primarily focused on structural response simulation. Future work can be done to apply UQ methods to research such as this, which utilizes measured data, to better understand how each feature influences the target predictions.

3.7 Conclusion

A dual Kriging-XGBoost model for predicting seismic structural demands was developed using measured response history data from 35 events and more than 200 buildings. Peak ground acceleration (PGA) was identified as a potentially useful feature in reconstructing the engineering demand parameters (EDPs) for individual buildings. The maximum peak story drift ratio (maxPSDR) and maximum peak floor acceleration (maxPFA) were chosen as the response quantities due to their known relationship to structural damage.

The first phase of the dual model used kriging to estimate PGA values at the sites of buildings without sensors. Based on its performance on the empirical data and general familiarity, the Gaussian semivariogram is adopted. The kriging model was able to predict the PGA at the site locations of buildings without sensors with a normalized root mean squared error (NRMSE) of 0.139. The second phase of the dual model used the PGA values obtained from kriging as part of the feature matrix and the Extreme Gradient Boosting (XGBoost) algorithm was implemented for predicting maxPSDR and maxPFA. The other considered features include the building location, source-to-site distance, period, number of stories, lateral force resisting system (LFRS) and the site Vs30.

The XGBoost-based structural response reconstruction produced near unity observed-to-predicted ratios (1.003 and 1.006 for maxPSDR and maxPFA, respectively) and near zero mean residuals, indicating unbiased models. Another desirable outcome was the absence of a trend between the observed-to-predicted ratios and source-to-site distance. The model predictions were assessed using Mean-Normalized Error (MNE) with values of 0.078 and 0.196 being obtained for maxPSDR and maxPFA, respectively. Including the PGA from kriging as a feature was found to improve the maxPFA prediction but did not have a measurable impact on the estimation of maxPSDR.

Some limitations of this study include the size and sparsity of the data set. Much of the data are from events with magnitudes between 4 and 6 and buildings with fewer than 20 stories located within 100km of the source event. Also, the data largely represents buildings with responses that are in the linear-elastic range, which reduces the generalizability of the model, especially in terms of its ability to detect damage. One of the inherent features of this model, its two-phase structure, could also be considered a limitation. The initial kriging that was used to estimate the PGA introduced an error into the data that was not previously present, although, as noted earlier, it did improve the overall model performance (relative to if PGA is excluded). The model could potentially perform better if the PGA residuals were reduced. PGA is known not to be well-correlated with displacement- and deformation-based EDPs, which explains the lack of improvement in maxPSDR predictions when it is included. Future studies should explore the other potential features, such as Peak Ground Velocity (PGV), Cumulative Absolute Velocity (CAV) or the first mode period spectral acceleration (SaT1).

Another limitation of the predictive model is that the methodology reduces the complex structural response to a single value (maxPSDR or maxPFA) in order to inform the performance of a large number of buildings, without providing information about localized damage. Lastly, the uncertainty in the response reconstruction was not quantified in the current model and should be considered in future studies. This research is an initial step towards the goal of developing a data-driven post-earthquake evaluation model that can be used for regional damage assessments as well as direct and indirect economic loss estimations.

CHAPTER 4

Evaluating the Effectiveness of Ground Intensity Measures using Building Strong Motion Data

4.1 Introduction

Structural health monitoring has been shown to be useful for informing rapid damage assessments of infrastructure following an earthquake. For individual facilities, strong motion response measurements can be incorporated into the decision-making process that determines whether an affected building is safe to occupy. For building portfolios, in addition to informing damage (or functionality) state determination, response measurements can also be used to guide the sequencing and timing of post-earthquake inspections. Numerous studies have been conducted to develop methodologies that use strong motion response measurements to assess post-earthquake structural integrity. While the earliest studies used metrics related to the change in the physical state of the structure [97, 95, 46, 28], more recent efforts have sought to leverage the performance-based earthquake engineering (PBEE) framework to assess immediate impacts using decision variables such as monetary losses and/or downtime [24, 80, **?**, 96].

An important aspect of the PBEE methodology is determining the most suitable ground motion intensity measure (IM) for structural response estimation. In simulation-based PBEE assessments (i.e., without the use of strong motion data), structural responses or engineering demand parameters (EDPs) are estimated using nonlinear response history analyses (NRHAs). In this context, numerous studies have been carried out to appraise the suitability of different IMs for structural response estimation. The basic formulation of such studies includes constructing a nonlinear structural model, performing NRHAs using a set of either site-specific or site-agnostic ground motions, and using the empirical distributions of the generated EDPs to evaluate alternative IMs using different criteria. Efficiency and sufficiency are the most commonly used criteria, however, IMs have also been evaluated based on their robustness to scaling and predictability.

The efficiency criterion evaluates an IM based on its ability to minimize the dispersion in the generated EDPs. The earliest studies focused on efficiency-based IM evaluations date back to the early 1990's where the primary IMs of interest were the peak ground acceleration (PGA) and the spectral acceleration at the first mode period (Sa_{T1}). The latter was found to be more suitable for estimating displacement-based EDPs such as peak story drift ratio [104]. Later studies considered other spectral acceleration-based parameters that were conditioned on two or more periods. Studies by [79] and [32] showed that the performance of Sa_{T1} could be enhanced by including additional modal periods (e.g., 1st and 2nd mode). This idea has since been extended by using the geometric mean spectral acceleration over a range of periods (Sa_{avg}), which has been shown to be very efficient for structures with highly nonlinear response or whose behavior is influenced by multiple modes (e.g., taller buildings) [42, 103]. Peak ground velocity (PGV) has also been shown to be effective (from an efficiency standpoint) for estimating peak story drift responses, especially in tall structures [21] while the filtered incremental velocity has been demonstrated to be suitable for collapse performance estimation [38]. In general, PGA has been shown to be the IM that is most suitable for estimating acceleration-based EDPs (e.g., peak floor acceleration) [103, 21].

The sufficiency criterion evaluates an IM based on the conditional independence assumption that is embedded in the PBEE triple integral. Specifically, this assumption implies that, conditioned on a given IM, the EDPs are independent of the ”upstream” parameters used in ground motion models (e.g., earthquake magnitude, source-to-site distance) [72]. There are alternative approaches to assessing the sufficiency of an IM. One uses a two-stage regression, the first step of which develops a linear model that predicts the EDP as a function of the

IM. The second step regresses the residuals from the first model against the upstream parameter of interest. The statistical significance of the coefficient from the second regression determines the sufficiency of the IM with respect to the considered upstream parameter. A p-value that exceeds some predefined significance threshold (usually 5%) implies that the upstream parameter is statistically independent of the EDP after conditioning on the IM and is therefore sufficient. An alternative approach to the two-stage regression is to using scaling to condition the ground motions on a predefined IM level. The EDPs generated by the scaled ground motions are then regressed against the upstream parameter and the resulting p-value is interpreted as described earlier. Like efficiency, the sufficiency criterion has been widely used in prior simulation-based IM evaluation studies [72, 104, 117, 42, 103, 38]. However, a recent study by [21] showed that the results of the sufficiency-based assessment varies depending on the number of ground motions considered in the analysis. Recently, alternatives to the sufficiency-based assessments have been proposed including the use of information theory [41, 39, 62] or methods and principles from the field of causal inference [21]. The latter has recently been shown to produce results that are invariant to the number of considered ground motions.

Because PBEE assessments are typically performed using simulated EDPs, the aforementioned studies utilized responses generated from NRHAs to evaluate the adequacy of IMs. However, as noted earlier, there is a growing emphasis on using strong motion building response data to conduct rapid PBEE-type assessments in the post-earthquake environment. A major challenge in this context is that the response measurements in a given building or portfolio are never ”complete”. Meaning, within in a given building, only a subset of the floor levels are instrumented. Also, in a given portfolio, only a subset of buildings are instrumented. For this reason, response reconstruction models are used to generate EDPs in the uninstrumented locations (i.e., a subset of the stories in a given building or a subset of the buildings in a given inventory) [97, 2, 113, 111, 114, 87, 98]. These response reconstruction models often utilize one or more IMs as predictor variables. Therefore, understanding the relative suitability of different IMs based on different criteria can inform the development of the EDP reconstruction models, especially when they are incorporated into a PBEE-type evaluation.

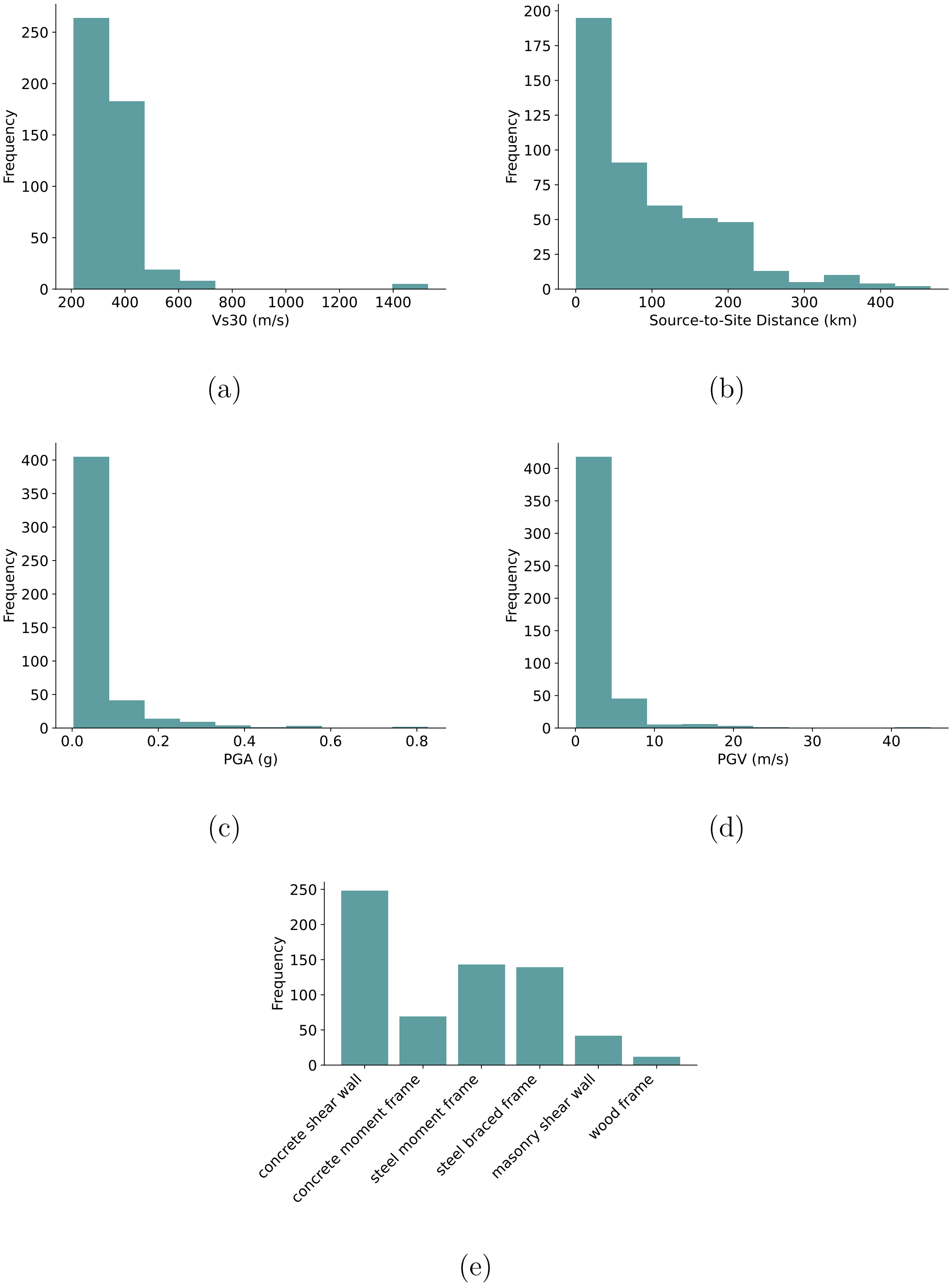
In this study, we assess the effectiveness of several different IMs using data from building strong motion response measurements. In addition to the sufficiency- and efficiency-based criteria, we evaluate the IMs using the principles and methods from the field of causal inference. The six IMs considered in the evaluation include PGA, PGV, cumulative absolute velocity (CAV), arias intensity (I_a), Sa_{T1} and Sa_{avg}. The response measurements are taken from the database developed by [4], which contains historical data from more than 200 buildings located in California that have been subjected to $\mathbf{M} > 4$ earthquakes during a period of more than 35 years. The recorded floor acceleration histories are used to generate peak story drift ratios (PSDRs) and peak floor accelerations (PFAs), which are the considered EDPs. The acceleration history measured at the ground level is taken as the input motion for each building and used to compute IM values. Several different causal inference methods are considered in the evaluation to determine whether the findings are invariant to the adopted model. We also evaluate the consistency of the findings for the three types of assessment (sufficiency, efficiency and causal inference) with respect to the size of the data set or the number of considered responses.

The next section provides an overview of the considered data set including the geometric and structural characteristics of the buildings that make up the inventory and the IM and EDP distributions. A brief but broad overview of the relevant aspects of causal inference is then presented followed by a detailed description of the considered methods. The sufficiency- and efficiency-based assessment criteria are then summarized. The subsequent section presents the results of the IM evaluation using the three approaches (sufficiency, efficiency and causal inference). The paper concludes with a summary of the key findings, limitations of the study, and suggestions for future related work.

4.2 Summary of Building, Event and Strong Motion Response Data

The building strong motion responses are obtained from the relational database developed in chapter 2. After removing data points with relevant missing information, 150 buildings with recordings from 35 earthquakes are considered, giving a total of 479 response measurements. The pertinent building properties include the number of stories (above ground), lateral force resisting system (LFRS) type, period estimated by ASCE 7-16 [9] ($T_{1,ASCE}$), and the response measurements, including their locations within the structure. The site properties include the location, source-to-site distance (epicentral), average shear wave velocity to a depth of 30m (V_{S30}), and site class. The event magnitude associated with each response measurement is also considered. Figure 6.7 shows histograms of the V_{S30}, source-to-site distance, PGA, PGV, and the LFRS type. For the LFRS type histogram, many of the buildings have multiple systems, so each system is counted individually for this plot. In Figure 4.2, a histogram of the number of stories is overlayed by plots showing the bin averages of (a) the ASCE period and (b) Sa_{T_1} and Sa_{Avg}. In (c), the coefficient of variation (COV) for each bin is plotted. These plots help to understand the data distribution and serve as a preliminary check. For example, in plot (a), we see that as the number of stories increases, the average period within a bin increases. Conversely, the spectral IMs in (b) decrease with the increase in number of stories (i.e. increase in period). From a physical standpoint, this is expected. The COV plots in (c) help to understand how the number of data points within a bin may effect the results we obtain from the analysis performed later. There are multiple factors considered in the analysis, and given that there are more data points in the low to mid-rise buildings, there is more variation in the building, site, and event characteristics, leading to the higher observed COV. This information will aid in interpreting the results and drawing conclusions.

The ASCE 7-16 estimated period is used to determine Sa_{T_1} and Sa_{avg}. Figure 4.3 shows the geometric mean spectra (considering ground motions in the two orthogonal directions) for all buildings in the set. Sa_{avg} is calculated as the geometric mean for the range $0.2T_1$



(a)

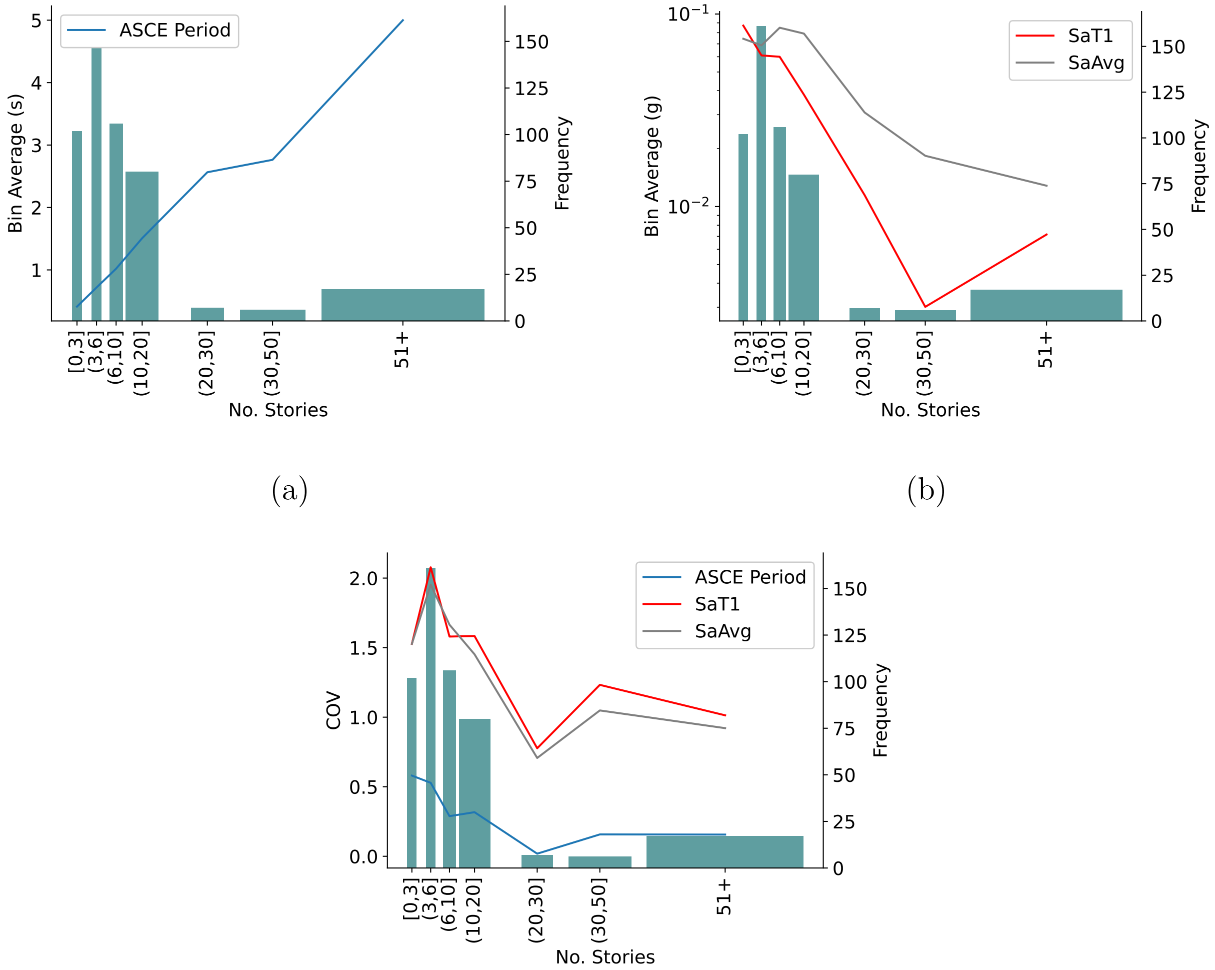
(b)

(c)

(d)

(e)

Figure 4.1: Histograms showing the empirical distribution of (a) V_{S30}, (b) source-to site distance, (c) PGA, (d) PGV, and (e) the lateral force resisting system types



(a)

(b)

(c)

Figure 4.2: Bin averages and histograms for (a) the ASCE Period (b) Sa_{T_1} and Sa_{avg} and (c) the corresponding coefficients of variation (COV)

to $3T_1$, as recommended by [43], where this range was found to perform well for collapse prediction in the study. It is important to note that the lack of moderate to highly nonlinear responses is one of the limitations of the current study. Specifically, the findings in terms of the most suitable IM(s) would only be applicable to problems involving structures that respond in the near-elastic range. Nevertheless, we believe the IM evaluation using measured responses provides new information that has overall relevance to PBEE-type assessments.

The EDPs (maxPSDR and maxPFA) are extracted from the time-series data. The story drift ratio is calculated as the difference in the displacement values in two adjacent stories divided by the story height. This value is calculated at each floor, using the maximum over the entire response history as the PSDR for that floor. The maximum over all floors is then taken as the maxPSDR. The instruments record the accelerations at their respective floors. The maximum at each floor over the entire response history is computed (PFA), and the maximum of these values over all floors is the maxPFA for a given building.

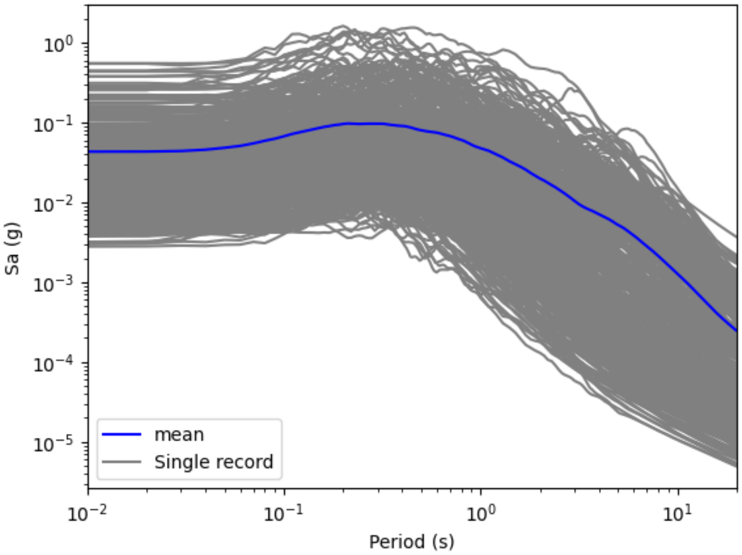


Figure 4.3: Geometric mean spectra (considering ground motions in the two orthogonal directions

4.3 Fundamentals of Causal Inference

4.3.1 Potential Outcomes

This section provides an overview of the relevant fundamental aspects of causal inference. In a broad sense, causal inference aims to utilize observational data (i.e., data not generated by a controlled experiment) to identify and/or quantify the causal effect of one variable (described as the ”treatment”) on another (described as the ”outcome” variable) in the presence of ”covariate” or ”confounding” variables. This can be contrasted with traditional statistics which is centered around establishing associative relationships among a set of variables. Potential outcomes, a cornerstone of causal inference, provide a conceptual mathematical framework for representing causal relationships. Originating from the works of Neyman and Rubin, potential outcomes are hypothetical outcomes that would have been observed under different treatment conditions [108, 100]. This framework lays the foundation for estimating causal effects and drawing causal conclusions.

In any causal problem, each individual or unit possesses two potential outcomes: the outcome that would be observed if the unit receives the treatment (denoted as Y_1), and the outcome under the condition that the treatment is not assigned to the unit (denoted as Y_0). However, the fundamental challenge of causal inference stems from the fact that only one of these potential outcomes can be realized in an observational data set, depending on whether a given unit actually receives the treatment or not. The potential outcomes framework allows us to define the causal effect of a treatment as the difference between Y_1 and Y_0 or $Y_1 - Y_0$. This causal effect captures the change in the outcome that can be attributed solely to the treatment itself while holding all other factors constant. To estimate causal effects within this framework, researchers rely on data from various study designs such as observational studies or randomized controlled trials. Regardless of the context, the main challenge lies in addressing the unobserved potential outcomes. Overcoming this challenge requires making identification assumptions that relate to the relationship between the potential outcomes and the treatment assignment mechanism. These assumptions play a critical role in estimating causal effects.

One widely recognized identification assumption is the stable unit treatment value assumption (SUTVA) [59]. SUTVA assumes that the potential outcome for a given unit is not influenced by the treatment assignment or potential outcomes of other units. This assumption ensures that the estimated causal effect can be solely attributed to the treatment itself. Another important criterion within the potential outcomes framework is related to the assignment mechanism or the process by which units are assigned to different treatment conditions. In randomized controlled trials, the assignment mechanism is typically random, ensuring that treatment assignment is independent of unit characteristics or covariate variables. In observational studies, however, the assignment mechanism may be affected by confounding factors, which can introduce bias into the estimated causal effects. This leads to the ignorability assumption, which says that the treatment assignment should be independent of potential confounding factors (variables affecting both the treatment and the outcome). Lastly, the consistency assumption says that the potential outcome of unit i corresponding to the treatment condition for the same unit is the outcome that is observed. So, in other words, if the treatment condition for unit i, $T_i = t$, then outcome, $Y_i^t = Y_i^T = Y_i$ [90].

The potential outcome framework provides a principled approach to understanding causal relationships. By considering the hypothetical potential outcomes and making appropriate identification assumptions, effect estimates can be quantified and more defensible conclusions can be drawn about the causal relationships between variables.

4.3.2 Directed Acyclic Graphs (DAGs) and Structural Causal Models (SCMs)

Causal graphical models have been proposed as a compliment to the potential outcomes framework [90]. Causal graphical models provide a visual representation of causal relationships and facilitate the identification of causal effects based on structural causal models. Directed Acyclic Graphs (DAGs) and Structural Causal Models (SCMs) are used to represent and understand causal relationships between variables. They provide a formal framework for modeling causal structures and making causal claims based on observed data.

In a DAG, variables are represented as nodes, and causal relationships are depicted by directed edges connecting the nodes. Importantly, DAGs impose a structure where causal effects flow from parent nodes to child nodes, creating a clear directionality of influence. The acyclic property ensures that there are no feedback loops or circular dependencies among variables. DAGs are useful for encoding prior knowledge about the causal relationships between variables. They help define the assumptions underlying the causal models and identify the confounding variables that need to be addressed when estimating causal effects. Figure 4.4 shows an example of a DAG for a hypothetical causal problem with outcome Y, treatment T, a set of covariates [X], and the corresponding unmeasured influences, U.

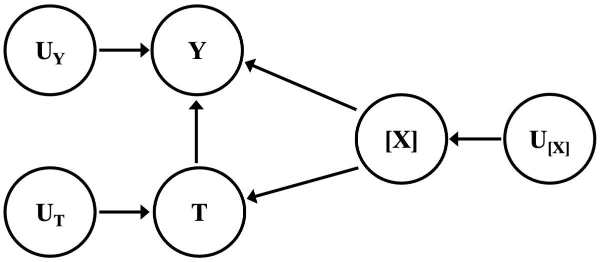


Figure 4.4: DAG for a hypothetical causal problem with treatment T, outcome Y and covariates [X]

SCMs are mathematical representations of the causal relationships described by a DAG. SCMs specify the interactions among variables in the system and how their values are determined. Each variable is associated with a structural equation that describes its dependence on its parent variables. The structural equations encode the causal mechanisms or processes that generate the values of the variables. In an SCM, there are endogenous variables and exogenous variables. Endogenous variables are those whose values are determined by other variables in the system, while exogenous variables are treated as inputs or sources of external influence, which are not affected by other variables in the system. The structural equations in an SCM provide a mathematical representation of the causal relationships implied by the DAG. By specifying the functional forms and parameters of these equations, causal effects can be estimated using statistical methods under different scenarios. The SCM associated with the DAG in Figure 4.4 is described in Equation 4.1.

$$X_i = f_{X_i}(U_{X_i}) \tag{4.1a}$$

$$T = f_T(X_1, ..., X_n, U_T) \tag{4.1b}$$

$$Y = f_Y(X_1, ..., X_n, T, U_Y) \tag{4.1c}$$

In summary, the combination of DAGs and SCMs enables causal conclusions to be drawn from observational data. DAGs provide a clear graphical representation of the causal assumptions, while SCMs formalize these assumptions into mathematical models. These models can the be used to estimate causal effects, perform counterfactual analyses, and assess the sensitivity of results to different assumptions.

4.4 Methods used to Evaluate the Efficacy of Ground Motion Intensity Measures

In addition to the well-established efficiency and sufficiency criteria, three different causal inference methods are used to evaluate the relative performance of the considered IMs. The subsequent subsections provide an overview of the causal inference approach to the IM evaluation problem as well a detailed description of the three methods. A brief summary of the efficiency and sufficiency-based assessment is also presented.

4.4.1 Causal Inference-Based Intensity Measure Evaluation

4.4.1.1 Overview

According to the sufficiency criterion, an effective IM is one that produces EDPs that are independent of the upstream parameters used in ground motions models (e.g., event magnitude, source-to-site distance) after conditioning on said IM. As discussed in the Introduction section, the sufficiency-based assessment of this conditional independence criterion can be evaluated using one of two approaches: (i) regressing the conditional EDP (i.e., EDPs associated with a single IM level) against the upstream parameter of interest or (ii) a two-stage process that first regresses the EDP against the IM and then regresses the residuals from the first model against the upstream parameter of interest. In either approach, if the upstream parameter of interest is found to be statistically insignificant (based on the associated p-value), the IM is deemed sufficient.

As described in [21], the conditional independence criterion can also be evaluated through the lens of causal inference. The DAG shown in Figure 4.5 provides a conceptual representation of the IM evaluation problem from a causal inference perspective. In this formulation, the EDP is the outcome variable and the IM is the treatment. The considered upstream parameters include the event magnitude (M), epicentral source-to-site distance (R), and site class (as defined by ASCE 7-16). Note that site class (SC in Figure 4.5) is used (as a categorical variable) in lieu of Vs_{30} because of the large uncertainty associated with the latter. Also, many of the Vs_{30} values provided by the Center for Strong Motion Data [25] are not directly measured but inferred from topological features. In the Figure 4.5 DAG, M, R and the site class (denoted by the vector $X_1 = \{M, R, SC\}$) are confounders because they affect both the IM and the EDP. It then follows that the causal effect of a given upstream parameter (e.g., M) on the EDP can be interpreted as having two parts. The direct effect represented in Figure 4.5 by the arrow from the upstream parameter to the EDP and the ”mediated” effect which flows through the IM. From a conditional independence standpoint, a desirable IM is one that maximizes the ”mediated” effect and minimizes the direct effect. However, since the direct effect is independent of the IM, the IM that maximizes the effect on the EDP while controlling for the upstream parameters is the most desirable (from a causal inference perspective).

In a typical NRHA-based IM evaluation study (like the prior studies described in the Introduction section), each structure is considered individually and the uncertainty is derived from using multiple ground motions. In the current study, the set of measured EDPs are from a portfolio of buildings. In other words, the IMs are being evaluated across multiple structure types. As such, the number of stories (N_s), LFRS-type and ASCE 7-16 estimated period (T_1) (denoted by the vector $X_2 = \{N_s, LFRS, T_1\}$) are also confounders because they affect both the IM and the EDP. Therefore, the causal inference-based IM evaluation considers both the upstream parameters and building/structural properties shown in Figure 4.5 as confounders (denoted by the vector $X = \{X_1, X_2\}$). Note that T_1 only affects the Sa_{T1} and Sa_{avg} IMs, which is why the arrow between the two variables is shown using a dashed line.

The IM evaluation problem shown in Figure 4.5 is solved using the following three causal inference methods: Inverse Propensity Weighting (IPW), Covariate Balancing Propensity Score (CBPS) and Entropy Balancing for Continuous Treatments (EBCT). Recall that for any causal problem, a primary challenge with estimating the effect of the treatment on the outcome variable is the presence of one or more confounders. One approach to addressing this challenge (illustrated in Figure 4.6) is to ”balance” the effect of the confounding variables (or covariates) before estimating the treatment effects using statistical regression. The differences among the IPW, CBPS and EBCT methods lie primarily in the covariate balancing strategy that is used before estimating the treatment effect. Considering all three methods will allow us to assess the invariance of the causal inference-based IM evaluation results to the covariate balancing strategy. Ideally, the relative performance of the considered IMs in estimating the EDPs should should not vary across the different covariate balancing approaches. As shown in Figure 4.6, two types of regression are considered after the covariate balancing is performed. Using simple linear regression provides a single average treatment effect (ATE) value. However, to consider any nonlinearity in the post-balancing relationship between the treatment and outcome, local linear regression is also performed to obtain an exposure outcome relationship.

4.4.1.2 Inverse Propensity Weighting (IPW)

To account for non-random assignment of the treatment in observational studies, inverse propensity weighting (IPW) [101] is a widely used technique in causal inference. It assigns weights to each observation based on their propensity scores, which represent the conditional

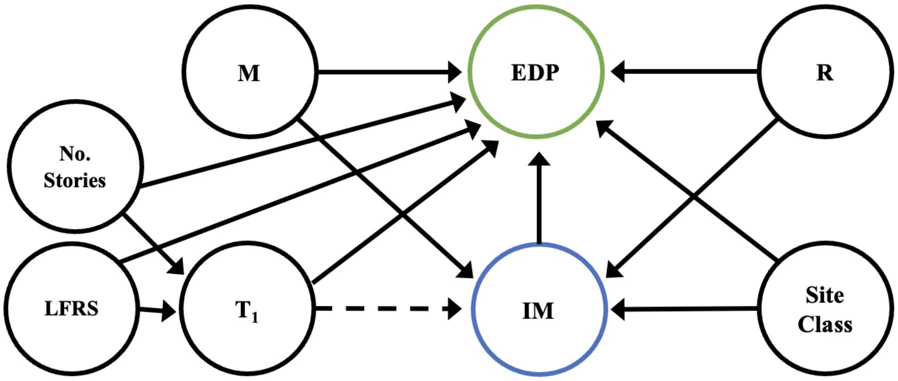


Figure 4.5: DAG showing a conceptual representation of the IM evaluation problem. The unmeasured influences are not shown.

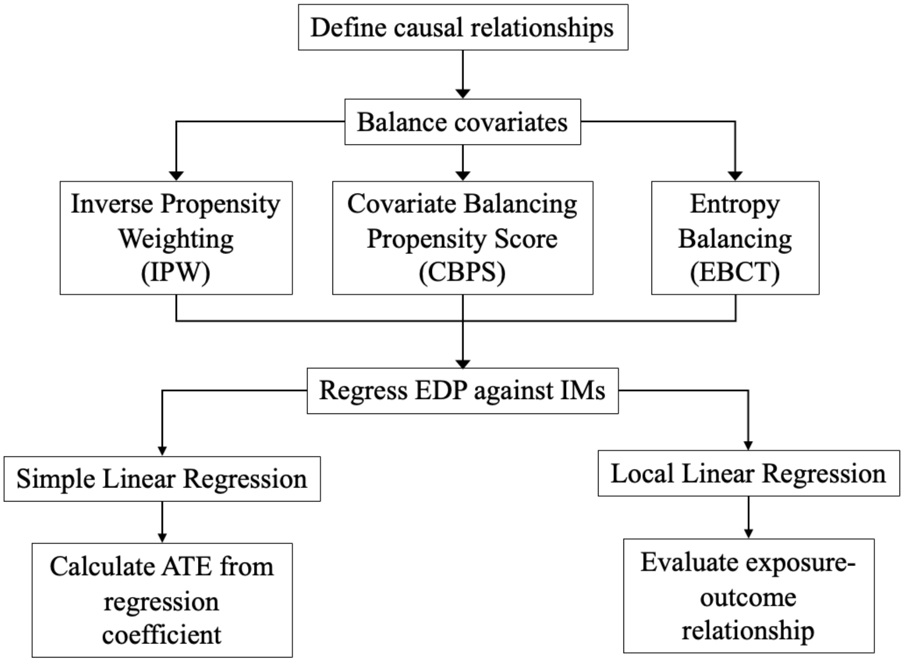

Define causal relationships
Balance covariates
Inverse Propensity Weighting (IPW)
Covariate Balancing Propensity Score (CBPS)
Entropy Balancing (EBCT)
Regress EDP against IMs
Simple Linear Regression
Local Linear Regression
Calculate ATE from regression coefficient
Evaluate exposure-outcome relationship

Figure 4.6: Schematic illustration of the covariate balancing approach used to solve the IM evaluation problem using causal inference

probability of receiving the treatment given the observed covariates.

Consider a binary treatment setting with a dataset comprised of N observations denoted by (Y_i, T_i, X_i) for $i = 1, 2, ..., N$. Here, Y_i is the outcome of interest, T_i is the binary treatment indicator (0 for control, 1 for treatment), and X_i represents the vector of observed covariates for the observation i. The propensity score, denoted as $\pi(X_i)$, is the conditional probability of receiving the treatment given the covariates X_i, i.e., $\pi(X_i) = P(T_i = 1|X_i)$. The goal is to estimate the average treatment effect (ATE), which is the average difference in outcomes between the treated and control groups. To balance the covariates across the treatment groups, we assign weights to each observation based on the inverse of the propensity scores. The weight for the i^{th} observation is given by $w(X_i) = 1/\pi(X_i)$ for the treated group $(T_i = 1)$ and $w(X_i) = 1/(1 - \pi(X_i))$ for the control group $(T_i = 0)$. The $w(X_i)$ operates by upweighting the underrepresented group and downweighting the overrepresented group, effectively balancing the covariate distributions between the treatment groups. The IPW estimate of the ATE is then calculated as:

$$ATE_{IPW} = (1/N)\sum[T_iY_iw(X_i)] - (1/N)\sum[(1 - T_i)Y_iw(X_i)] \tag{4.2}$$

where the first term on the right-hand side represents the weighted sum of outcomes in the treated group, and the second term represents the weighted sum of outcomes in the control group. By incorporating the weights based on the inverse propensity scores, the IPW estimator effectively creates a pseudo-population where the treatment assignment is independent of the observed covariates. This allows for unbiased estimation of the treatment effect by removing the confounding bias induced by the non-random treatment assignment.

For continuous treatment variables (as is the case in the IM evaluation problem), the IPW methodology can be extended by replacing the binary treatment indicator with the continuous treatment values. The propensity scores are then estimated using the appropriate models, such as logistic regression for binary treatments or generalized propensity score models for continuous treatments. The weights are calculated as the inverse of the estimated propensity scores, similar to the binary treatment case. The IPW estimator for the average treatment effect with continuous treatments follows a similar formulation as the binary treatment case:

$$ATE_{IPW} = (1/N) \sum (T_i - \mu(T)) w_i Y_i \tag{4.3}$$

where T_i now represents the continuous treatment values, and $w(X_i)$ is the corresponding weight based on the inverse propensity score. μ_T is the conditional mean of the treatment given the covariates.

Estimating the propensity scores and choosing of the appropriate model for estimating treatment effects are the two main steps in the IPW methodology. Various techniques, such as logistic regression, machine learning algorithms, or propensity score matching, can be employed to estimate the propensity scores.

IPW is a useful technique in causal inference that addresses the issue of non-random treatment assignment in observational studies. By assigning weights based on the inverse of propensity scores, IPW balances the covariate distributions between treatment groups and allows for unbiased estimation of treatment effects. The methodology can be applied to both binary and continuous treatment settings, with appropriate modifications in the calculation of weights and estimation procedures.

4.4.1.3 Covariate Balancing Propensity Score (CBPS)

Covariate Balancing Propensity Score (CBPS), originally developed by [61], is a method used in causal inference to estimate treatment effects while achieving balance in the covariates between the treatment and control groups. The method was extended by [50] to consider continuous treatments. It combines the principles of propensity score estimation and weighting techniques to address confounding bias and enhance the validity of causal analyses. Similar to IPW, the CBPS method begins by estimating the propensity scores, $\pi_\beta(X_i) = P(T_i = 1|X_i)$ using logistic regression:

$$\pi_\beta(X_i) = \frac{exp(X_i^T\beta)}{1 + exp(X_i^T\beta)} \tag{4.4}$$

where β is an unknown parameter vector of size equal to the number of covariates, which can be estimated by maximizing the log-likelihood in Equation 4.5. Once the propensity scores are estimated, the CBPS method assigns weights to each unit based on their propensity score and covariate values. The definition of weights based on the covariates is where CBPS differs from IPW. In the CBPS framework, the weight assigned to unit i, W_i, is given by Equation 4.6.

$$\hat{\beta}_{MLE} = \arg\max_\beta \sum_{i=1}^{N} T_i log(\pi_\beta(X_i)) + (1 - T_i) log(\pi_\beta(X_i)) \tag{4.5}$$

$$W_\beta(T_i, X_i) = \frac{T_i - \pi_\beta(X_i)}{\pi_\beta(X_i)(1 - \pi_\beta(X_i))} \tag{4.6}$$

This ensures that units with covariate values that are unbalanced between the treatment and control groups receive higher weights, facilitating covariate balance. When extended to continuous treatments, transformation of the treatment and covariate vectors is performed so that the means and variances are 0 and 1, respectively. The transformed vectors are denoted as $\tilde{T}_i$ and $\tilde{X}_i$. The transformation of the treatment variable is shown in Equation 4.7. The same equation is used to transform the covariate vector.

$$\tilde{T}_i = \frac{T_i - \bar{T}}{\sqrt{s_T}} \tag{4.7a}$$

$$\bar{T} = \sum_{i=1}^{N} \frac{T_i}{N} \tag{4.7b}$$

$$s_T = \sum_{i=1}^{N} \frac{(T_i - \bar{T}^2)}{N - 1} \tag{4.7c}$$

The weights for the continuous treatment case are defined by:

$$w_i = \frac{f(\tilde{T}_i)}{f(\tilde{T}_i|\tilde{X}_i)} \tag{4.8}$$

To satisfy the covariate balancing conditions that the weights (i) remove the correlation between $\tilde{X}_i$ and $\tilde{T}_i$ and (ii) maintain their marginal means, the following constraints are applied.

$$\sum_{i=1}^{N} w_i g(\tilde{X}_i, \tilde{T}_i) = 0 \tag{4.9a}$$

$$\sum_{i=1}^{N} w_i - N = 0 \tag{4.9b}$$

where $g(\tilde{X}_i, \tilde{T}_i) = (\tilde{X}_i, \tilde{T}_i, \tilde{X}_i\tilde{T}_i)^\top$. Equation 4.8 can be rewritten in terms of w_i as $f(\tilde{T}_i, \tilde{X}_i) = \frac{1}{w_i} f(\tilde{T}_i) f(\tilde{X}_i)$. To maximize the likelihood of the full sample, we maximize the following function with the corresponding contraints:

$$\prod_{i=1}^{N} f(\tilde{T}_i, \tilde{X}_i) = \prod_{i=1}^{N} \frac{1}{w_i} f(\tilde{T}_i) f(\tilde{X}_i) \tag{4.10}$$

$$\begin{aligned} \text{s.t.} \sum_{i=1}^{N} w_i g(\tilde{X}_i, \tilde{T}_i) &= 0 \\ \sum_{i=1}^{N} w_i &= N \\ \sum_{i=1}^{N} w_i \tilde{X}_i \tilde{T}_i &= 0 \\ w_i &> 0 \text{ for all i} \end{aligned}$$

The likelihood is equivalent to $\sum_{i=1}^{N} log(f(\tilde{T}_i)) + log(f(\tilde{X}_i)) - log(w_i)$. Since we are only interested in w_i, this equation simplifies to

$$\arg\min \sum_{i=1}^{N} log(w_i) \tag{4.11}$$

Using the Lagrange method to evaluate the optimization problem and solve for w_i, we obtain

$$w_i = \frac{1}{1 - \gamma^\top g(\tilde{X}_i, \tilde{T}_i)} \tag{4.12}$$

where γ is the Lagrange multiplier associated with the first constraint. With the assigned weights, the CBPS method computes the weighted average treatment effect. The w_i's are incorporated in the estimation of the ATE by giving more weight to units whose covariate values are unbalanced between the treatment and control groups.

By achieving covariate balance, the CBPS method aims to reduce confounding bias and provide more reliable estimates of the treatment effect. It allows for the adjustment of multiple covariates simultaneously and can handle both binary and continuous treatments.

CBPS combines propensity score estimation and weighting techniques to achieve covariate balance. By assigning weights based on the estimated propensity scores and covariate values, CBPS aims to reduce confounding bias and provide more accurate treatment effect estimates.

4.4.1.4 Entropy Balancing for Continuous Treatments (EBCT)

Entropy Balancing for Continuous Treatments, or EBCT, was developed by [118] as a continuous treatment extension to the original entropy balancing method for binary treatments [54]. The primary function of EBCT is to minimize the correlation between the moments of the pre-treatment covariates and the treatment by weighting each unit. The weighted set can then be used to estimate the ATE with minimal influence from the confounders. EBCT works by first partitioning the observations into bins based on the treatment variable. Then, within each bin, the covariate distributions are balanced using a weighting scheme that minimizes the dispersion of the weights (the entropy). This ensures that the weights are evenly distributed across the treatment and control groups. Once these weights are obtained, they can be used in techniques such as weighted linear regression to estimate the ATE.

The binary entropy balancing case will be first introduced then expanded to deal with continuous treatments. The binary treatment variable is denoted as T such that each data point assigned to either the treatment (1) or control (0) group. For a set of N observations ($N = N_0 + N_1$), each unit, i, has a treatment status T_i and covariates X_i. Entropy balancing seeks to find the weight, w_i, for each unit that balances the covariate distribution in the treatment and control groups. The formulation uses the Kullback entropy metric ($h(w)$ in Equation 4.13).

$$h(w_i) = w_i \ln \left(\frac{w_i}{\pi_i} \right) \tag{4.13}$$

where π_i, the selected base weight, is set equal to $1/N_1$ for all i when not specified. The goal is to minimize this entropy subject to certain constraints via the loss function $H(w)$ in Equation 4.14.

$$\min_w H(w) = \sum_{i|T_i=0} h(w_i) \tag{4.14}$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{i|T_i=0} w_i c_{ri}(X_i) = m_r, \ r \in 1, ..., R \\ & \sum_{i|T_i=0} w_i = 1 \\ & w_i \geq 0 \text{ for all } i \text{ such that } T_i = 0 \end{aligned}$$

where $c_{ri}(X_i)$ is the covariate moment functions of the control group, which can be defined as $c_{ri}(X_i) = X_{ij}^r$ or a de-meaned version of the moment functions given by $c_{ri}(X_{ij}) = (X_{ij} - \mu_j)^r$, where μ_j is the mean and j is a corresponding covariate for observation i. m_r represents the r^{th} moment of the treatment group for covariate X_j. For example, if r = 1, the constraint is such that only the means of the covariates are balanced. If r = 2, then the means and the variances are balanced, and so on. In summary, the constraints impose that the sum of the weighted covariate moments in the control group must equal the moments of the covariates in the treatment group.

The extension of EBCT to consider continuous treatment variables is based on [118]. The treatment T is now a continuous variable, with $T_i > 0$ for treated units. The covariate moment function for the binary case, $c_{ri}(X_i)$, is rewritten as $\tilde{X}_i = f(X_i) - \frac{1}{N_1}\sum_{i|T_i>0} f(X_i)$ where $f(\cdot)$ is a function of the covariate and any interaction terms to be balanced, and the de-meaning is done with respect to the treatment group. The variables p and $\tilde{T}$ are used to denote the order of the treatment term to be balanced and a de-meaned representation of the treatment (similar to $\tilde{X}$ for the covariates), respectively. The column vector $g(p, \tilde{X}_i, \tilde{T}_i) = [\tilde{X}_i{}', \tilde{T}_i, ..., \tilde{T}_i{}^p, \tilde{X}_i{}' \cdot \tilde{T}_i, ..., \tilde{X}_i{}' \cdot \tilde{T}_i{}^p]$ considers all necessary covariate and treatment moments and interactions. Equation 4.14 can then be re-written for continuous treatments as

$$\min_{w} H(w) = \sum_{i|T_i=0} h(w_i) \tag{4.15}$$

$$\begin{aligned} \text{s.t.} \quad & \sum_{i|T_i=0} w_i g(p, \tilde{X}_i, \tilde{T}_i) = 0 \\ & \sum_{i|T_i=0} w_i = 1 \\ & w_i > 0 \text{ for all } i \text{ such that } T_i = 0 \end{aligned}$$

To obtain the weights, the constrained optimization problem can be reformulated by applying Lagrange multipliers (λ, γ) to the constraints. Recall from Equation 4.13 that $h(w) = w_i \ln\left(\frac{w_i}{\pi_i}\right)$, so the new unconstrained problem becomes

$$\min_{w,\lambda,\gamma} L(w, \lambda, \gamma) = w_i \ln\left(\frac{w_i}{\pi_i}\right) - \lambda\Big\{\sum_{i=1}^{N_1} w_i - 1\Big\} - \gamma'\Big\{\sum_{i=1}^{N_1} w_i g(p, , , \tilde{X}_i, , , \tilde{T}_i)\Big\} \tag{4.16}$$

Differentiating with respect to λ and w_i produces the equation for calculating the balancing weights

$$w_i = \frac{q_i \exp\Big\{\gamma' g(p, , , \tilde{X}_i, , , \tilde{T}_i))\Big\}}{\sum_{i=1}^{N_1} q_i \exp\Big\{\gamma' g(p, , , \tilde{X}_i, , , \tilde{T}_i)\Big\}} \tag{4.17}$$

4.4.2 Efficiency and Sufficiency

In general (statistical) terms, efficiency and sufficiency are two approaches to quantifying the quality of an estimator. The efficiency criterion gauges how effectively the available data is utilized to estimate the outcome of interest. Sufficiency assesses whether the estimator alone is sufficient with respect to one or more covariates in explaining the response. Sufficiency is evaluated by conditioning on the estimator and evaluating the dependence of the response on the covariate(s) of interest. In the context of the current study, the IM is the estimator and the upstream parameters are the covariates.

The primary objective in efficiency-based evaluation is to identify the IM that offers the most precise estimates of the EDPs given certain assumptions or conditions. Specifically, an estimator is deemed efficient when it achieves the smallest variance among the class of estimators being compared. In other words, it strives to minimize the dispersion between estimated values and the true value of the parameter. Consider the following regression:

$$ln\,\hat{Y} = ln\,a + b \cdot ln\,X + \epsilon \tag{4.18}$$

where $\hat{Y}$ is the response variable of interest, X is a predictor variable, a and b are regression parameters and ϵ is the error. An efficient estimator is one that minimizes the variance in the error term ϵ.

The sufficiency criterion is a measure of how well the estimator alone predicts the outcome. The two stage regression approach is adopted in the current study. The first regression is between the intensity measure and the response, the same as for efficiency in Equation 4.18. The second regression takes the residuals from the first stage and regresses them against each covariate, shown in Equation 4.19. The p-values obtained from this second regression indicate the sufficiency of the IM. If the p-value is greater than 0.05 (a commonly used threshold for statistical significance), then the covariate is rendered independent of the first stage regression residuals conditioned on the treatment, deeming the IM sufficient.

$$\epsilon = c \cdot Z + \eta \tag{4.19}$$

where Z is the covariate or upstream parameter, c is a regression parameter and η is a second error term. The baseline unbalanced data set is evaluated for both efficiency and sufficiency for comparison with the causal inference approaches that use covariate balancing.

4.5 Ground Motion Intensity Measure Evaluation and Results

This section evaluates the relative effectiveness of five ground motion intensity measures in estimating two types of engineering demand parameters. The considered IMs include PGA, PGV, CAV, I_a, Sa_{T1} and Sa_{avg}. Each sensor in a building that is subjected to a given event provides response history data at a specific location (i.e., floor level). The response histories are recorded at small intervals over the duration of the event. In most buildings, sensors are not placed on every floor, so linear interpolation was performed to calculate the response histories on those floors that are not instrumented [114]. This information is then used to calculate the peak response demands. The peak story drift ratio (PSDR) and peak floor acceleration (PFA) are used to denote the peak response corresponding to each story and floor respectively. The maximum PSDR and PFA over all stories and floors, respectively, denoted as maxPSDR and maxPFA, are the considered EDPs. To calculate the peak response demands, the diaphragm is assumed to be rigid and torsional effects are neglected. The causal inference-based evaluation via the three alternative covariate balancing methods are implemented in addition to the efficiency and sufficiency criteria.

4.5.1 Covariate Balancing

This subsection presents the results of the covariate balancing sub-step of the causal inference-based evaluation illustrated in Figure 4.6. Recall that the goal of covariate balancing is to remove the correlation between the confounding covariates and the treatment variable (EDP) to minimize bias in the effect estimation. Balancing also reduces the effective sample size

(ESS) which are summarized in Table 4.2. ESS measures the amount of precision in the data. Specifically, higher correlations in data results in lower ESS due to data redundancy. The considered covariates include M, R, SC, and T_1. Since the number of stories (N_s) and lateral force resisting system (LFRS) only affect the IM through the first mode period, they are not included as covariates. More formally, it can be shown that considering T_1 in the covariate balancing also addresses confounding effects of LFRS and N_s. For this data set, it is expected that the ESS will be significantly reduced given the high correlations that M has with the exposure variables.

The results from the covariate balancing are shown in Table 4.1 which summarizes the Pearson correlations between the confounding covariates and treatment variable (IM) before and after balancing is performed. The post-balancing correlations, which are computed after applying the weights (w_i) to each observation, serve as an indication of the relative balancing ability of each method. The values reported in the table are the Pearson correlations between the covariates and treatment variables. Prior to balancing, the earthquake magnitude has the highest absolute correlation - which is positive, as expected - with the IMs. Also, as expected, T_1 has relatively high pre-balancing correlations with Sa_{T1} and Sa_{avg}. R has low to moderate correlations with PGV, Sa_{T1} and Sa_{avg} and low positive correlations with PGV, CAV and I_a before balancing. Recall that SC is a categorical variable, which, for the purpose of the analysis, is assigned values of 0 and 1 for site class C and D, respectively. As such, the positive (but low) pre-balancing correlations between SC and the various IMs indicate that site class D increases the IM level relative to site class C. Based on the observed post-balancing correlations, it is clear that EBCT provides the best balance, effectively eliminating the correlations for all variables while maintaining a larger ESS (see Table 4.2) than the other two methods. Nevertheless, the results using all three balancing methods are reported to evaluate the consistency of the relative performance of the different IMs.

Table 4.1: Pearson correlations between the IMs and covariates before and after balancing

Covariate	**IM**	**Unweighted**	**IPW**	**CBPS**	**EBCT**
M	PGA	0.279	-0.143	0.013	0.000
	PGV	0.661	0.196	0.001	0.000
	CAV	0.742	0.367	0.198	0.000
	I_a	0.742	0.330	0.166	0.000
	Sa_{T1}	0.514	0.001	0.041	0.000
	Sa_{avg}	0.533	-0.055	0.110	0.000
R	PGA	-0.381	-0.354	-0.111	0.000
	PGV	0.046	0.146	0.003	0.000
	CAV	0.091	0.084	-0.002	0.000
	I_a	0.092	0.051	-0.020	0.000
	Sa_{T1}	-0.007	-0.263	-0.161	0.000
	Sa_{avg}	-0.041	-0.284	-0.115	0.000
Site Class	PGA	0.011	-0.039	0.015	0.000
	PGV	0.056	0.078	-0.005	0.000
	CAV	0.079	0.047	0.071	0.000
	I_a	0.070	0.021	0.046	0.000
	Sa_{T1}	0.196	0.069	0.056	0.000
	Sa_{avg}	0.135	0.039	0.049	0.000
T_1	Sa_{T1}	-0.470	-0.521	-0.201	0.000
	Sa_{avg}	-0.478	-0.313	-0.153	0.000

4.5.2 Causal Effect Estimation Results

After the balancing is performed and the weights (w_i) for all samples are determined, a weighted linear regression is performed for each method. Both simple and local linear regressions are performed. The local linear regression provides more flexibility in the functional relationship between the exposure and outcome, so the form does not need to be assumed.

Table 4.2: Effective sample size corresponding to the different covariate balancing methods

IM	**IPW**	**CBPS**	**EBCT**
PGA	132.9	69.9	220.1
PGV	99.8	52.8	171.7
CAV	126.3	53.9	129.3
I_a	90.5	36.2	130.5
Sa_{T1}	46.8	122.3	154.0
Sa_{avg}	140.5	85.6	1141.3

It serves as a way to evaluate how the effect may vary at different intensity levels of the IMs. The simple linear regression allows for a more direct and intuitive comparison of the population ATE, which is taken as the regression coefficient. When the relationship is nonlinear, the effect cannot be simply summarized into a single value for the population. However, the exposure-outcome curve is useful for highlighting important aspects of the relationship between the treatment (IM) and outcome (EDP).

The results of the simple linear regression are summarized in Table 4.3 and Table 4.4. The reported values are the regression coefficient (β) (causal effect) which is interpreted as the β% increase in the outcome (EDP) caused by a 1% increase in the treatment (IM). Recall that, from a conditional independence standpoint, the most desirable IM is the one with the largest β value (shown in bold font). From Table 4.3, it is observed that, across all three covariate balancing methods, CAV has the highest causal effect for PSDR and is therefore the best performing IM (from a causal inference perspective). It is notable that CAV is one of the least studied IMs specifically as it relates to its relative performance in estimating EDPs. However, this finding is consistent with that of the study by [82] which found that CAV is well-correlated with damage, even moreso than other IMs such as PGA. The worst performing IM for PSDR, which is the one with the lowest causal effect, is PGA. Again, this is consistent with several prior studies which showed that PGA does not perform well for displacement-based EDPs [42, 103].

The results from Table 4.4 show that CAV is also the best performing IM for PFA across all balancing methods. It is a bit unusual that PGA is outperformed by most of the other IMs (with the exception of Sa_{T1}) when estimating PFA. Most prior simulation-based studies have shown that, at least from an efficiency standpoint (discussed later in this section), PGA performs relatively well for acceleration-based EDPs. On the other hand, the fact PGA generally performs much better for PFA than for PSDR (when comparing the causal effects in Table 4.3 and Table 4.4)) is consistent with the principles of structural dynamics. Further, when attempting to benchmark the current results against findings from the prior literature, it is important to highlight that (i) all prior studies utilize simulation-based data whereas this study uses response measurements and (ii) the current study is based on aggregated data from multiple structure types and configurations whereas all prior studies utilize data from one structure type at a time. Although the covariate balancing is implemented to consider these differences, the analysis is limited by the amount of available data. The strata for some variables have few data points compared to others (i.e. high-rise vs low-rise buildings). The results are also dependent on the performance of the balancing methods and the resulting effective sample sizes. Results where the ESS is less than 10% of the original sample size are excluded due to the low precision.

When evaluating the results across the different methods, it is important to acknowledge their respective performance. For example, it can be observed that the causal effect estimation for PGA on PSDR is significantly larger when balanced using EBCT compared to IPW and CBPS. This may initially seem concerning without considering that not only was EBCT better at eliminating the correlations in the covariates, but for this IM-EDP pair, EBCT maintained a larger ESS. The ESS for CBPS, for example, is only just barely above the 10% cutoff criterion.

The results of the weighted local linear regressions are shown in Figure 4.7 and Figure 4.8. The estimated relationships using weights from each balancing method are overlaid on the original data and plotted with the solid colored lines. We see that although the linear model assumption was not unreasonable for most IM-EDP pairs, there is a more complex relationship as the intensity level increases. Specifically, for CAV, Ia and Sa_{avg}, the relationships take a sigmoidal shape, with the effect on the EDP's reduced at the lower and higher ends of the IM distributions. In one case (Sa_{avg} effect on the EDPs), an unrealistic negative trend is observed when using equally-weighted or CBPS-weighted regression. As noted by Table 4.1 and Table 4.2, not only was CBPS unsuccessful in eliminating the correlations, but it reduced the ESS substantially. It is unsurprising that in a range of the data that has very few points to begin with, that the relationship may not be appropriately captured. Similarly for the equally weighted case, equal weight is being given to points that represent a very small subset of the entire data, as is noted by the scatter plots in Figure 4.7 and Figure 4.8. EBCT, which performed best, relative to the other methods, maintaining the highest ESS and lowest correlation, does not display this negative trend.

Table 4.3: Estimated causal effect (β) for PSDR from post-balancing linear regression

IM	**IPW**	**CBPS**	**EBCT**
PGA	0.324	0.379	0.686
PGV	0.926	0.85	0.849
CAV	**1.316**	**1.635**	**1.698**
I_a	1.053	-	1.401
Sa_{T1}	-	0.58	1.122
Sa_{avg}	0.335	0.867	1.091

4.5.3 Efficiency and Sufficiency

This section presents the results of the efficiency and sufficiency-based evaluations. Table 6 summarizes the standard errors from the probabilistic seismic demand models (based on Equation 4.18) which serve as a measure of efficiency. For PSDR, PGV produces the lowest dispersion and is therefore the best performing IM for that EDP. Although, it is noteworthy that CAV and I_a have dispersion values that are comparable with (5% higher than) PGV, which is generally consistent with the findings from the causal analysis. It is interesting that

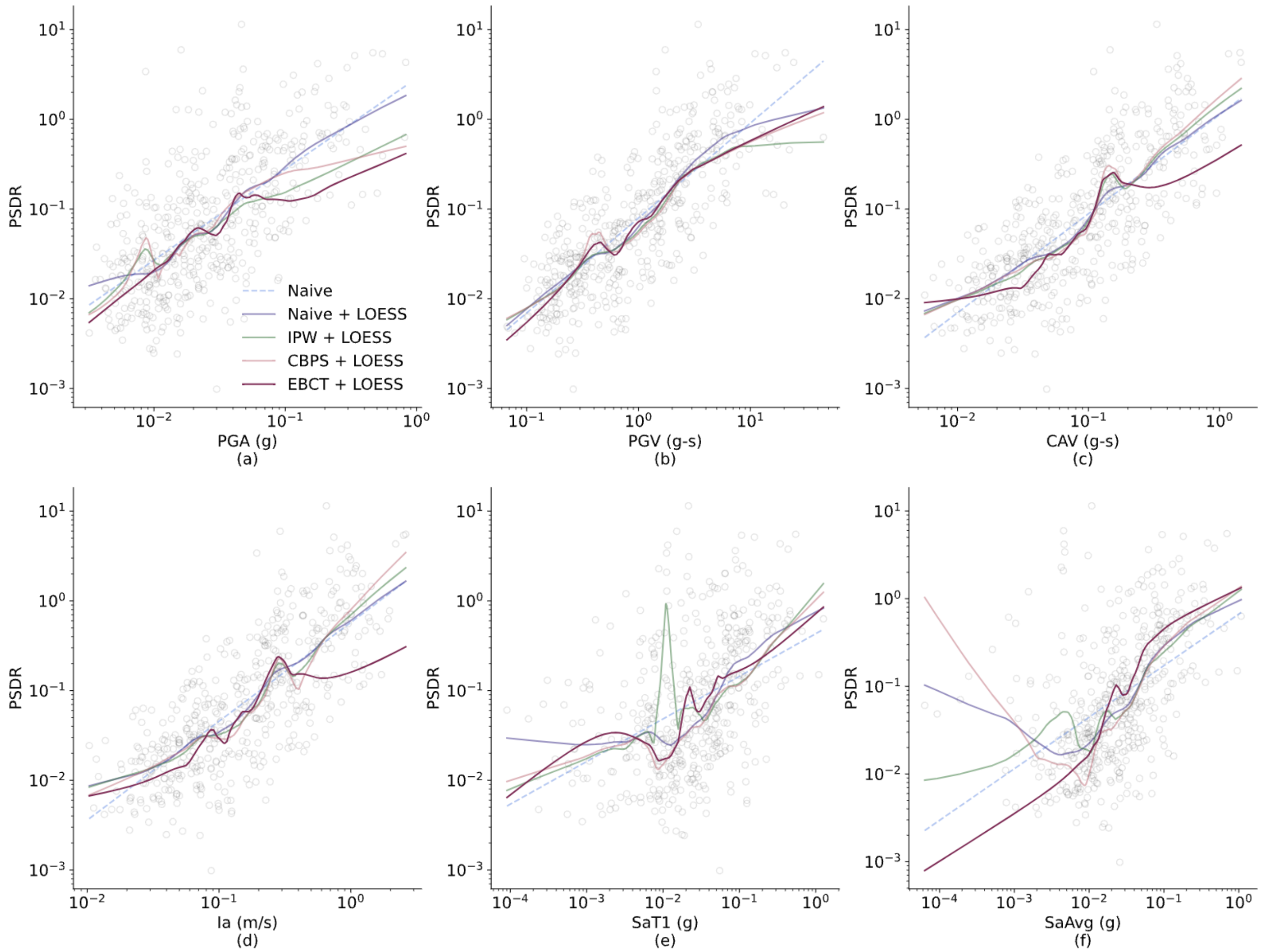
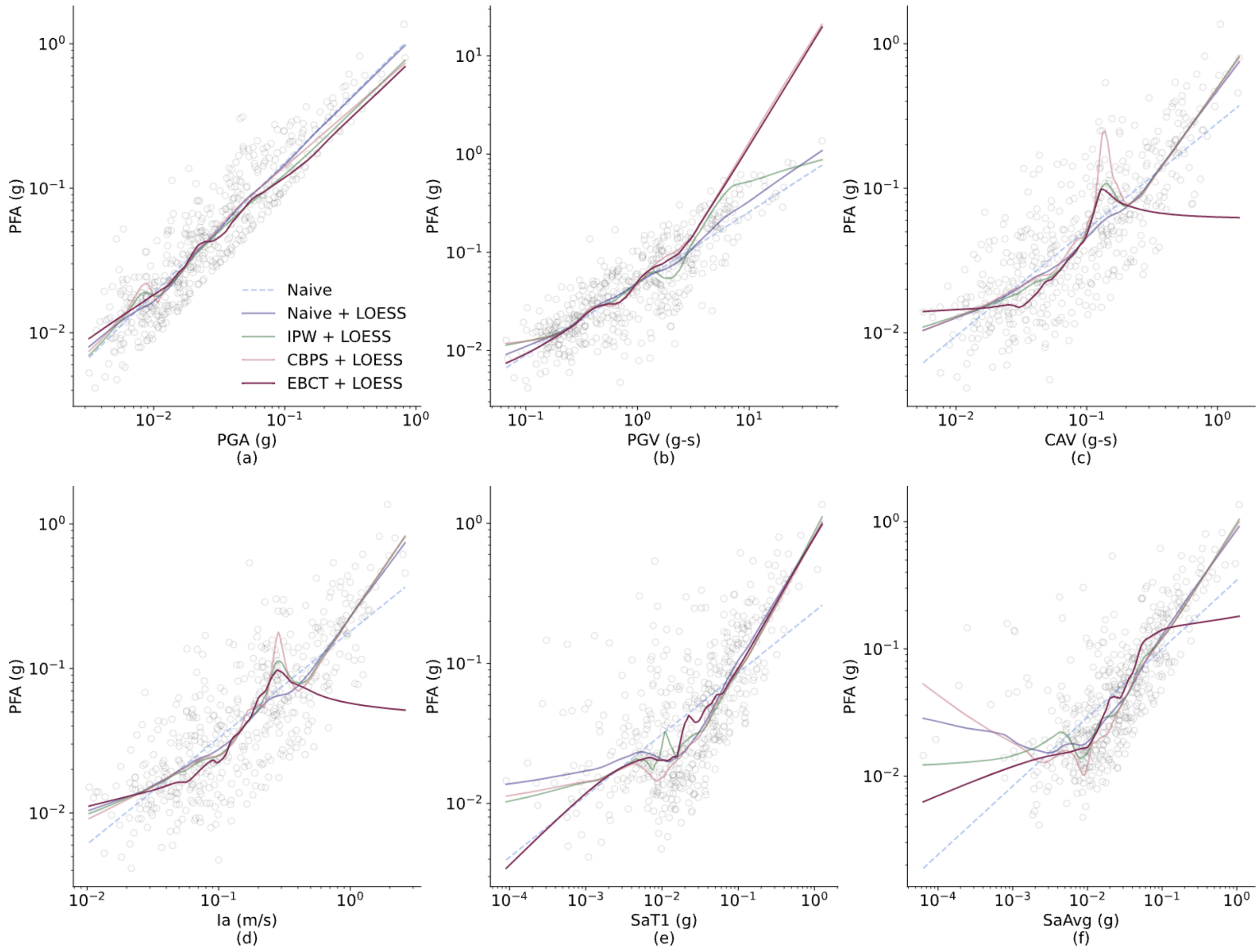


Figure 4.7: Exposure-outcome Curves for PSDR vs a) PGA, b) PGV, c) CAV, d) Ia, e) Sa_{T_1} and f) Sa_{Avg} using Naive simple linear regression, Naive LOESS, IPW-weighted LOESS, CBPS-weighted LOESS, and EBCT-weighted LOESS



(a)

(b)

(c)

(d)

(e)

(f)

Figure 4.8: Exposure-outcome Curves for PFA vs a) PGA, b) PGV, c) CAV, d) Ia, e) Sa_{T_1} and f) Sa_{Avg} using Naive simple linear regression, Naive LOESS, IPW-weighted LOESS, CBPS-weighted LOESS, and EBCT-weighted LOESS

Table 4.4: Estimated causal effect (β) for PFA from post-balancing linear regression

IM	**IPW**	**CBPS**	**EBCT**
PGA	0.687	0.672	0.689
PGV	0.758	1.041	1.037
CAV	**1.194**	**1.597**	**1.549**
I_a	0.992	-	1.371
Sa_{T1}	-	0.625	0.609
Sa_{avg}	0.718	0.935	1.053

PGA outperforms Sa_{T1} and Sa_{avg} in estimating PSDR. This runs counter to the findings from prior simulation-based studies. One possible explanation is that the inventory considered for the current study comprises primarily low-rise buildings (as summarized in Figure 4.2). In fact, more than half of the buildings in the inventory have five stories or less. This, coupled with the fact that most of the responses are in the linear elastic realm may explain the superior performance of PGA relative to Sa_{T1} and Sa_{avg}. For PFA, PGA produces the lowest dispersion and is therefore the most efficient. This is consistent with the findings from prior simulation-based studies which have shown that PGA performs well for acceleration-based EDPs.

The p-values from the sufficiency-based evaluations are summarized in Table 7. For each upstream parameter (i.e., M and R), the p-values for the sufficient IMs (i.e., p-values above the 5% significance threshold) are highlighted in bold. The fact that CAV and I_a are sufficient with respect to both M and R when estimating PSDR while PGA isn't, is consistent with the findings from the causal analysis. For PFA, only Sa_{T1} and Sa_{avg} are sufficient with respect to M and none of the IMs are sufficient with respect to R. The latter finding is a bit surprising and inconsistent with the results from prior studies [72, 89, 16]. However, as described in [21], this may be explained by the sensitivity of the sufficiency-based findings to the size of the data set.

Table 4.5: Efficiency of IMs as measured by the standard error from the PSDM (Equation 4.18)

IM	**PSDR**	**PFA**
PGA	1.30	**0.43**
PGV	**1.07**	0.58
CAV	1.12	0.66
I_a	1.13	0.67
Sa_{T1}	1.55	0.85
Sa_{avg}	1.51	0.78

Table 4.6: The p-values from the sufficiency-based IMs evaluations

Covariate	**IM**	**PSDR**	**PFA**
M	PGA	0	0
	PGV	**0.33**	0
	CAV	**0.58**	0
	I_a	**0.58**	0
	Sa_{T1}	0	**0.24**
	Sa_{avg}	0	**0.06**
R	PGA	0	0
	PGV	**0.41**	0
	CAV	**0.46**	0
	I_a	**0.45**	0
	Sa_{T1}	**0.26**	0
	Sa_{avg}	**0.94**	0

4.6 Conclusion

The goal of this study was to assess the effectiveness of ground intensity measures (IM's) using strong motion building response data. Sufficiency, efficiency, and causal analyses are used to evaluate the performance of different IMs for potential use in a structural response reconstruction context. In the causal analyses, covariate balancing methods are implemented to control for known confounding variables that contribute to building response, and the weights from these methods were used to perform regressions that define the causal relationship between the IMs and engineering demand parameters (EDPs).

The IMs considered include peak ground acceleration (PGA), peak ground velocity (PGV), cumulative absolute velocity (CAV), Arias Intensity (Ia), first mode period spectal acceleration (Sa_{T_1}), and the geometric mean spectral acceleration over a range of periods (Sa_{avg}). For the causal analysis, each IM was balanced against common cause covariates of the IM and EDP (or confounders), then regressed using weights calculated from the covariate balancing. The simple linear regression provided an intuitive understanding of the population level average treatment effect (ATE), which represents how a change in the intensity level of the IM, on average, changes the response. A local linear regression was also performed to allow for a more complex functional relationship that captures the variation in the effect across the intensity levels. This revealed potential heterogeneous effects in the data, where the effect varies among covariate subgroups. The effects of the IMs on the EDPs, including when accounting for confounding, were shown to be mostly linearly positive, but the results also highlighted some complexity in the relationships.

The results from the causal analysis indicated that CAV generally outperforms the other IMs. A prior study by [82] also found that CAV generally outperforms other IMs, especially in terms of damage detection. Also consistent with prior studies, PGA was found to be the worst performing IM when estimating PSDR. There were some surprising observations, such as PGA being outperformed by most of the other IMs in the causal analysis. However, this was explained by a number of factors including the structural heterogeneity within the data. In terms of efficiency, the findings were generally consistent with the causal evaluation.

The sufficiency-based evaluation also produced some unexpected results. For example, the spectral IMs (Sa_{T1} and Sa_{avg}) were generally outperformed by PGA and none of the IMs were sufficient with respect to source-to-site distance. The inconsistencies (with respect to prior studies) in the sufficiency analysis may be explained by the relatively small size of the data set as it has been shown in past studies to be sensitive to sample size [21].

There are several limitations of this study which serve as motivation for future related research. The EDP data set comprises mostly elastic responses, which means that the findings cannot be generalized to nonlinear demand levels. The imbalance in the distribution of structural characteristics also produced some challenges. For instance, the majority of buildings can be classified as low to mid-rise with very few high rises. While it is conceptually possible to account for heterogeneous effects (e.g., variations in IM performance across different building heights) in the causal analysis, this was not feasible in the current study because of the relatively small size of the data set.

Future studies can address some of the aforementioned limitations by expanding the suite of strong motion data to consider a broader range of demand levels (i.e., elastic and inelastic). Alternatively, a hybrid data set can be constructed by combining recorded and simulated responses. In addition to addressing issues related to sample size, a hybrid data set can be used to ensure greater balance in the structural characteristics of the considered building inventory. This would enable an assessment of heterogeneous effects (in the causal evaluation) using subgroup analysis.

CHAPTER 5

An end-to-end framework for portfolio-scale seismic design, analysis and performance assessment of building inventories with multiple lateral force resisting systems

5.1 Introduction

Evaluating building performance using the PBEE paradigm involves multiple steps, including hazard characterization, structural response analysis, and damage and loss assessment. When performed manually (i.e., without automation), this process can be very time consuming, requiring on the order of days to evaluate a single building. This time sink becomes especially challenging for regional assessments where the PBEE steps are applied to hundreds or even thousands of buildings. One solution to this challenge is to generate an automation procedure that implements the PBEE steps in sequence where human input is only required at the first (i.e., defining the necessary inputs) and last (i.e., receiving the outputs) steps. This could potentially cut down the amount of time engineers spend on the design, analysis, and performance assessment processes.

The evolution of structural engineering software has led to the creation of programs capable of automating various aspects of the design process. Early efforts in automation included software for structural analysis and design that required substantial user input and expertise. Recent advancements have produced more sophisticated tools that integrate multiple design stages, from conceptual design to detailed analysis, significantly reducing the need for manual intervention. Programs such as ETABS, SAP2000, and RAM Structural System represent milestones in automating structural design and analysis [30, 31, 115]. These tools not only perform complex calculations but also offer design optimization features, user-friendly interfaces, and the ability to handle a variety of structural systems. They support various materials and structural forms, including steel, concrete, and timber, aligning with modern design codes and standards. More recently, python-based programs have been developed that perform code-conforming building design, nonlinear response history analysis, and in some cases, damage and loss assessment. In fact, multiple python-based programs have been developed to automate some or all PBEE steps for individual lateral force resisting systems [53, 36, 5], requiring minimal inputs from the engineer. The Automated Seismic Design and Analysis (AutoSDA) [53], WoodSDA [36], and RCWallSDA [5] platforms were developed to advance the state of the art in automated design, analysis and performance assessment. The advancements in automated design have the potential to allow engineers to conduct performance-based design optimization and regional scale seismic performance assessments using the PBEE framework [49].

The current study aims to streamline the design and analysis process for buildings for steel moment frame, wood frame, and concrete shear wall lateral force resisting systems (LFRS). The program is developed in Python and includes an additional module for performing loss assessment based on Federal Emergency Management Agency (FEMA) P-58 PBEE methodology [49]. The program can be used to perform end-to-end design and analyses of a single building or a portfolio of buildings. This type of program could be used for regions where there is limited information on the building composition and instrumentation data. Reasonable building designs can be generated, ground motions applied to the resulting model, and nonlinear analyses and loss assessments performed to determine the potential impact.

5.2 Program Overview

AutoSDA is comprised of three previously developed and tested seismic design and analysis programs for buildings with a steel moment frame (SteelSDA), wood frame (WoodSDA), or concrete shear wall (RCWallSDA) LFRS. The intention of this program is to streamline the design, analysis, and assessment of building portfolios by combining these into a unified program. With general information about the building geometry, location, loading conditions, and occupancy, an entire region of buildings can be assessed using this program. The program structure is illustrated in Figure 5.1. Each of the previously developed LRFS-specific platform is included as a module in AutoSDA and an additional module is developed for performing loss assessment using the pelicun package [129], which follows FEMA P-58. The user inputs include basic building information and ground motions to be applied to the models. Based on the LFRS type, the program executes the appropriate design module, where an iterative procedure is implemented (outlined in the corresponding publications). Structural models are created in the Open System for Earthquake Engineering Simulation (OpenSees) [77], applying the selected ground motions and the design and response information are generated as outputs into excel, csv, and text files. These outputs can then be used to execute the Loss Assessment module, which takes the EDPs and information on the building and its components to provide estimates of the potential loss and uncertainty. The program is structured such that a portfolio of buildings that comprises some combination of these three LFRS types can be run simultaneously when provided with the necessary inputs for each building. The loss assessment module uses the story- and floor-level demands of the building along with component input information and applies the FEMA P-58 framework to assess potential damage and losses. Each sub-routine takes in excel files for the inputs to make its use accessible.

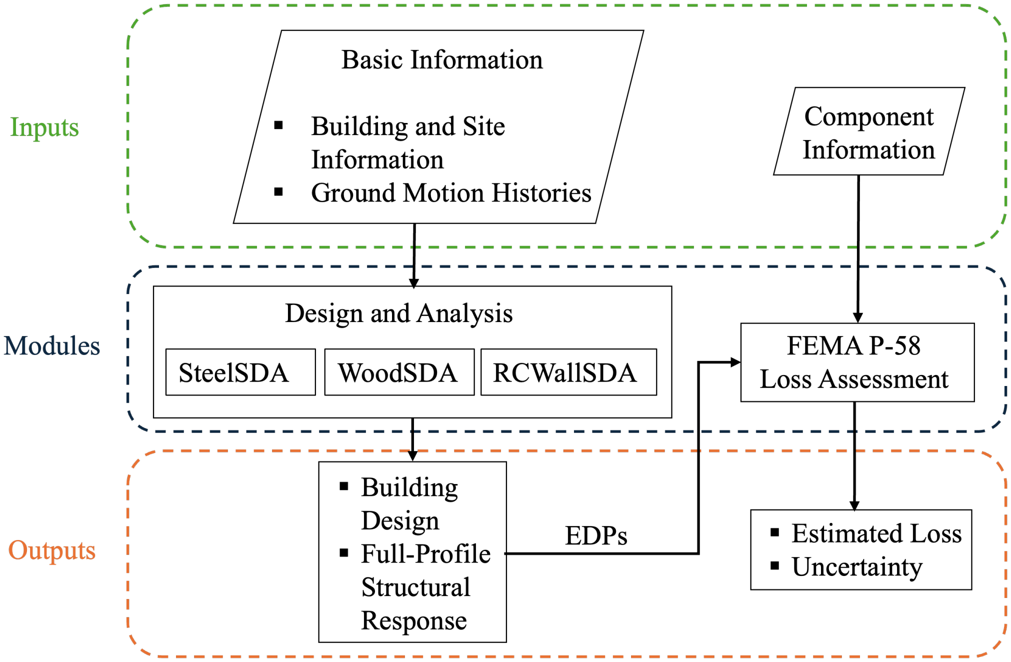


Figure 5.1: AutoSDA program Structure

5.2.1 Design Modules

The inputs required to run the design and analysis of a building depend on the LFRS type. The user creates a file or folder for each building with the corresponding inputs. In the top level, the user creates a csv file with the building ID, location, Site Class, Risk Category, and LFRS type. The program takes this information and initiates the appropriate sub-module based on the LFRS type. The buildings in this list can be run simultaneously using parallel processing. This initial input information is also used to perform a lookup of the seismic design parameters according to ASCE 7-16 [9] if the user does not want to input them manually. After the design and analyses are performed by the sub-modules, the results are stored in an output folder in the top level. The following sections briefly describe each sub-module. Figure 5.2 - Figure 5.4 outline the general inputs that are expected by each sub-module in order to execute.

5.2.1.1 SteelSDA

The SteelSDA module was developed by [53] and is originally named ”AutoSDA”. The name was changed to SteelSDA for this project to better identify the module and the name AutoSDA was repurposed to define the entire program and its modules. SteelSDA performs full code-conforming design and OpenSees analysis of steel moment frame buildings. The details of programming the design process are outlined in [53]. SteelSDA takes in information about the building geometry, loading conditions, and initial beam sizes to begin the design process of the frames. A summary of the required inputs is shown in Figure 5.2. Only three excel files are needed to run the program. The outputs include the code-conforming building design, connection design, component demands, and full-profile building responses.

SteelSDA determines the design forces and deformations using the Equivalent Lateral Force (ELF) method. Based on a set of pre-defined assumptions, the design is performed iteratively, checking demand limits specified in Table 12.2-1 of ASCE 7-16 [9], adjusting member sizes until the criterion is met. A structural model is then created and analyzed in OpenSees. An example archetype for a fourteen-story building with a five-bay steel moment frame LFRS is shown in Figure 5.7. Designs generated by steelSDA were validated in the corresponding publication by designs generated by Englekirk professional engineers as well as other researchers, reporting similar design layouts and story drifts. SteelSDA was also used to create a database [78] of steel moment frame (SMF) buildings and responses to ground motions at service, design, and maximum considered earthquake (MCE) levels that can serve as a repository for data-driven modeling.

5.2.1.2 WoodSDA

The WoodSDA module was developed by [36] to automate the design and analysis of wood framed buildings. WoodSDA requires more detailed information on the wall schedule and panel properties, so template building information files are provided that can be modified to aid in creating the necessary input files. Similar to the SteelSDA module, WoodSDA performs

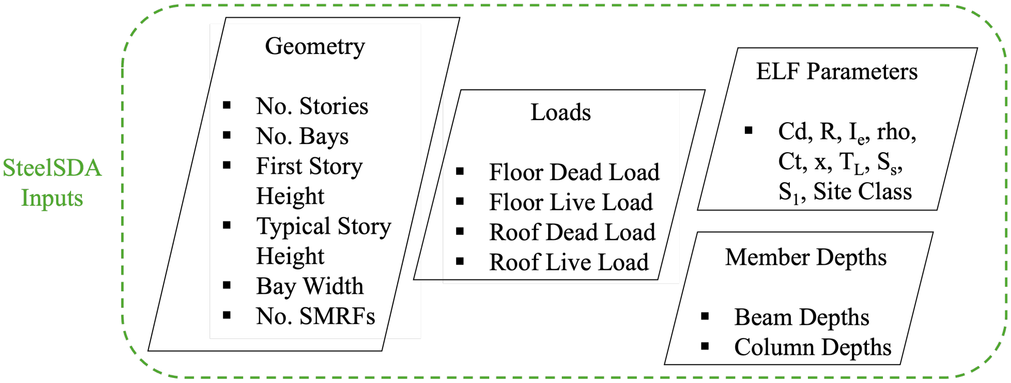


Figure 5.2: SteelSDA inputs

an iterative design process until demand limits are met and returns the full-detailed design before running the nonlinear analysis in OpenSees. The original Auto-WoodSDA included a loss assessment portion, but this was decoupled from WoodSDA and exists as a standalone module that can be applied to any LFRS type in AutoSDA. An example of the inputs for WoodSDA are shown in Figure 5.3. The inputs shown in the figure include general building geometry and properties, but WoodSDA does require more extensive effort to initialize due to the program structure. Loading and material properties, similar to those outlined in the figure must be manually defined for each grid of the building layout.

In order to simplify and generalize the building design sub-module, WoodSDA operates under certain assumptions in the design process. The assumptions are as follows: the building layout is simplified to a rectangular shape for modeling purposes, floor systems are initially treated as flexible (though this can be adjusted to rigid or semi-rigid behavior if needed), Solid shear walls without openings are used in the model to reduce complexity, and shear wall dimensions are kept constant throughout the building height. In the publication associated with WoodSDA, the model outputs were validated against a FEMA P-2139 project archetype and found to have good agreement.

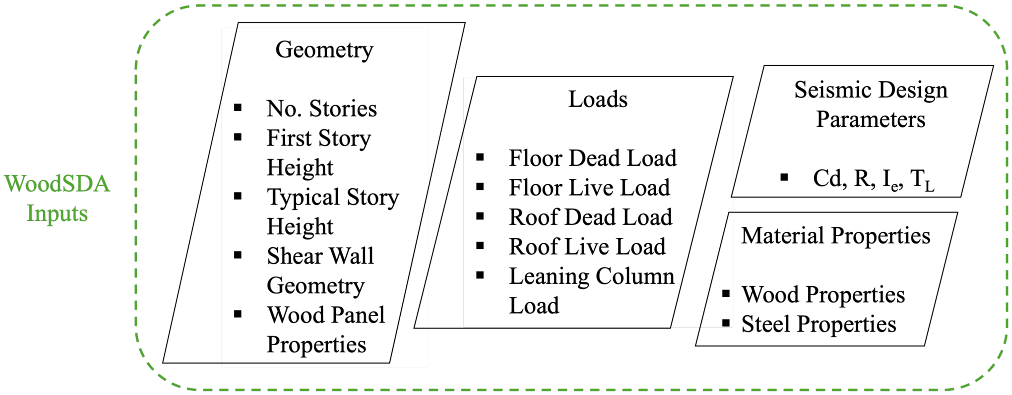


Figure 5.3: WoodSDA inputs

5.2.1.3 RCWallSDA

The RCWallSDA module was inherited from [5] as a python-based platform for the design and analysis of concrete shear wall buildings. Similar to the previous two modules, the code conforming design is performed iteratively, until the specified input criteria are satisfied. This tool streamlines the entire design process, from model construction to member sizing, significantly reducing the manual effort and time. The process begins with user-specified input parameters, including the building geometry, loads, site conditions, material properties, and initial wall dimensions. RCWall-SDA then implements the ASCE 7-16 equivalent lateral force procedure (with the option to use Response Spectrum Analysis, or RSA, instead) to determine seismic forces, and constructs an OpenSees model for linear elastic analysis to calculate story drifts. The WallDesign class performs design and analysis calculations according to ACI 318-19 provisions for special structural walls, evaluating the need for special boundary elements using a displacement-based approach. The iterative design process repeats until an adequate design meeting all code requirements is achieved. The necessary inputs to run RCWallSDA are shown in Figure 5.4.

5.2.2 Loss Assessment Module

The loss assessment module in AutoSDA uses the FEMA P-58 methodology to estimate potential losses for a particular building, and in the case of this project, a portfolio of buildings.

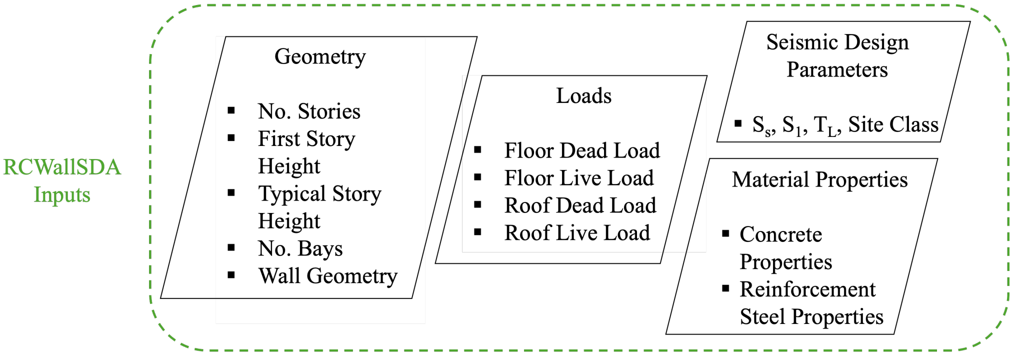


Figure 5.4: RCWallSDA inputs

Additional information regarding the component (structural and non-structural) population of a building is required to run this module. Aside from the components, the input necessary to perform the assessment are the EDP (i.e. peak story drift ratio and peak floor acceleration). The EDPs are converted to structural and non-structural component damage, which is used to compute the associated repair costs. Uncertainties in seismic hazard, structural response, damage, and consequences, are propagated throughout the methodology.

An overview of the PBEE methodology, as implemented in pelicun, is illustrated in Figure 5.5. First, the facility must be defined, D. This is where the building attributes such as the number of stories, occupancy type, any corresponding component definitions, are specified. Once this information is provided, the model proceeds to the probabilistic seismic hazard analysis (PSHA). The PSHA determines the probability distribution of one or more intensity measures at the site of interest. The result is a site hazard profile, $p[IM]$, characterizing the seismic intensity distribution at the location of interest. Next, nonlinear dynamic analysis or simplified nonlinear static analysis procedures are used to simulate how a structure will respond to seismic loading. Detailed structural models are developed to assess the response under a suite of ground motions. The structural response is evaluated with respect to the probabilistic site hazard obtained from the previous step. The outcome is the probability distribution of the structural response profile, $p[EDP]$.

Fragility assessments are then conducted to determine the probability of observing different damage states for each component conditioned on the seismic demand. This includes

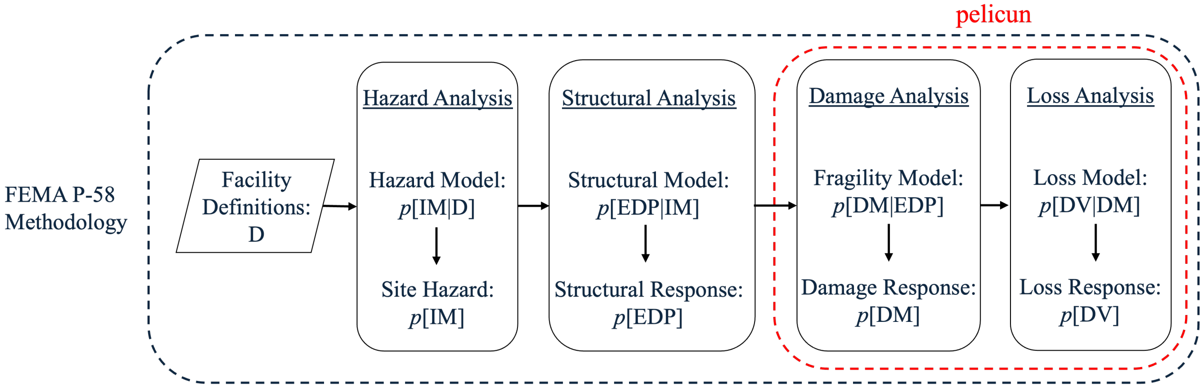


Figure 5.5: FEMA P-58 loss assessment procedure

both structural components such as beams and columns, and non-structural components like mechanical systems and cladding. These damage states are subsequently translated into economic losses, downtime, and casualties through consequence functions. These functions relate the extent of physical damage to repair costs, restoration time, and potential injuries or fatalities, also considering indirect losses such as business interruptions.

In the final phase, the potential losses associated with the damage are quantified using a loss model. This model calculates the decision variables (DV), which could include repair costs, downtime, or any other relevant metrics, given the damage measures (DM). The output of the FEMA P-58 methodology is typically expressed in probabilistic terms, providing metrics such as expected annual losses (EAL) and loss exceedance curves.

The pelicun package that is used in AutoSDA only performs the last two steps of this process. The structural response profile, $p[EDP]$, is defined by the structural analysis performed in the design and analysis modules and directly implemented into the damage analysis step, where the EDP distribution is combined with the component fragility information to obtain a probabilistic damage distribution.

5.2.2.1 Component Definitions

As part of the facility definition for the loss assessment, additional information on the components is required. This information must be provided in a CSV file formatted with details on the quantity, location, and demand direction of the component. Table 5.1 shows an example of a set of both structural and nonstructural components that were collected from the Seismic Performance Prediction Platform (SP3) [57], which also implements the FEMA 58 methodology. The table lists the component ID, description, type, location, the demand directions being considered, quantity units, and the quantity (θ_0). θ_0 refers to the quantity of the component in *each* direction at *each* location. As an example, for this 14-story building, there are 24 units of the B.10.31.001 component (bolted shear tab gravity connections) in 2 directions at 14 floors, for a total of 672 units.

The pelicun package in Python includes a database of component fragilities at installation. This database does not include all potential components, so the user has the option to manually input the fragility parameters for any additional components. Table 5.2 lists the fragility parameters for the components from the Table 5.1 and Figure 5.6 shows an example visualization of fragilities for two of the components, one for a component governed by drift demands (bolted shear tab gravity connections) and the other governed by acceleration (suspended ceiling). Each components fragility parameters define a lognormal cumulative distribution function. The generated example component table is an inexhaustive list of components generated by SP3 for a fourteen-story steel moment frame commercial office occupancy building. Once the component information is supplied, pelicun retrieves the appropriate fragility parameters, and estimates the per component loss based on the demand (EDPs). The component-level losses are aggregated to estimate the losses and repair times for a specified building.

5.3 Illustrative Example

In this section, a 14-story steel moment frame (SMF) archetype (Archetype 1188) from the database in [78] is used to illustrate loss module. The complete design and analysis are performed in steelSDA. A plan view of the building layout as well as an elevation of the moment frame are illustrated in Figure 5.7. A suite of 40 maximum considered earthquake (MCE) level ground motions are applied to the model for the dynamic analysis in AutoSDA. The

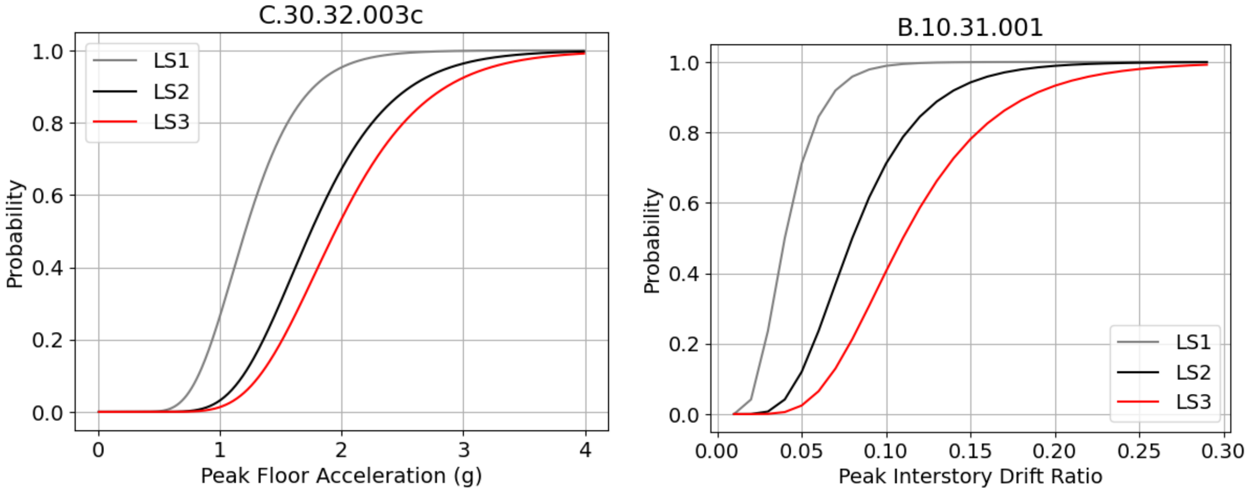
Table 5.1: Example component quantities

ID	Component Description	Type	Location	Direction	Units	θ_0
B.10.31.001	Bolted shear tabconnections	Structural	all	1,2	ea	24
B.10.31.011c	Steel Column Base Plates	Structural	1	1,2	ea	10
B.10.31.021c	Welded column splices	Structural	2,4,6,8, 10,12,14	1,2	ea	10
B.10.35.002	RBS connection w/ welded web, beam one side of column only	Structural	all	1,2	ea	4
B.10.35.012	RBS connection w/ welded web, beams both sides of column	Structural	all	1,2	ea	6
B.20.22.002	Curtain Walls - Generic Stick-Built Curtain wall	Exterior Cladding	all	1,2	ft^2	3000
C.20.11.001b	Prefabricated steel stair	Other	all	1,2	ea	1
C.30.11.001b	Wall Partition, Type: Gypsum + Wallpaper	Interior Finishes	all	1,2	ft	37.8
C.30.27.002	Raised Access Floor	Interior Finishes	2–14	0	ft^2	7500
C.30.32.003a	Suspended Ceiling, SDC D,E Area (A): A <250	Interior Finishes	2–14	0	ft^2	2250
C.30.32.003b	Suspended Ceiling, SDC D,E Area (A): 250 <A <1000	Interior Finishes	2–14	0	ft^2	2250
C.30.32.003c	Suspended Ceiling, SDC D,E Area (A): 1000 <A <2500	Interior Finishes	2–14	0	ft^2	2250

Table 5.2: Example component fragility parameters

ID	Demand Type	Demand Unit	LS1 θ	LS1 β	LS2 θ	LS2 β	LS3 θ	LS3 β	LS4 θ	LS4 β
B.10.31.001	PID	unitless	0.04	0.4	0.08	0.4	0.11	0.4	-	-
B.10.31.011c	PID	unitless	0.04	0.4	0.07	0.4	0.1	0.4	-	-
B.10.31.021c	PID	unitless	0.02	0.4	0.05	0.4			-	-
B.10.35.002	PID	unitless	0.03	0.3	0.04	0.3	0.05	0.3	-	-
B.10.35.012	PID	unitless	0.03	0.3	0.04	0.3	0.05	0.3	-	-
B.20.22.002	PID	unitless	0.021	0.45	0.024	0.45			-	-
C.10.11.001b	PID	unitless	0.01	0.3	0.013	0.3	0.018	0.3	-	-
C.20.11.001b	PID	unitless	0.005	0.6	0.017	0.6	0.028	0.45	-	-
C.30.11.001b	PID	unitless	0.0064	0.3	-	-	-	-	-	-
C.30.27.002	PFA	g	1.5	0.4	-	-	-	-	-	-
C.30.32.003a	PFA	g	1.6	0.3	1.95	0.3	2.07	0.3	-	-
C.30.32.003b	PFA	g	1.47	0.3	1.88	0.3	2.03	0.3	-	-
C.30.32.003c	PFA	g	1.21	0.3	1.75	0.3	1.95	0.3	-	-
C.30.32.003d	PFA	g	1.09	0.3	1.69	0.3	1.91	0.3	-	-
C.30.34.002	PFA	g	1.5	0.4	-	-	-	-	-	-

5 % damped acceleration spectra of the applied ground motions and the median spectrum are shown in Figure 5.8. Here, the suite of ground motions represents the site hazard characterization, $p[IM]$, in the hazard analysis stage of the FEMA P-58 methodology outlined in Figure 5.5. The distribution of EDPs to these ground motions represent the structural response characterization, $p[EDP]$. The individual and median responses are shown in Figure 5.9. In addition, pelicun requires the total building area and estimated replacement cost of the building for the loss assessment. For this example, the total building area is $315,000 ft^2$ and the estimated replacement cost ($\$323/ft^2$) and time (670 days) were extracted from SP3. With this information, pelicun can be used to perform the damage and loss analysis. The



(a)

(b)

Figure 5.6: Component fragilities for a) a suspended ceiling (C.30.32.003c) demand and b) bolted shear tab gravity connections (B.10.31.001)

distributional parameters of the full-profile response are sent to the loss assessment module, as well as component information. The component population is generated from SP3. The components in Table 5.1 are used for this example. The median and log standard deviation of the response are used to generate 10,000 Monte Carlo sample realizations of the structural response in the loss module and descriptive statistics of the loss and repair estimates are reported. Figure 5.10 shows the distribution of aggregated repair costs and times for the 10,000 response realizations. Table 5.3 shows the estimated aggregated loss statistics for this building.

In this example, the code conforming design and analysis of a 14-story steel special moment frame building generated by steelSDA was used to showcase the loss assessment module. Based on the distributional parameters of the response to MCE-level ground motions, the loss assessment provided estimates of the repair costs and repair times. This is a simple example of how the end-to-end framework that AutoSDA provides can be used to develop designs and anticipate potential losses in a single program. The loss module functions independently of the three design and analysis modules, and can be run with response data from

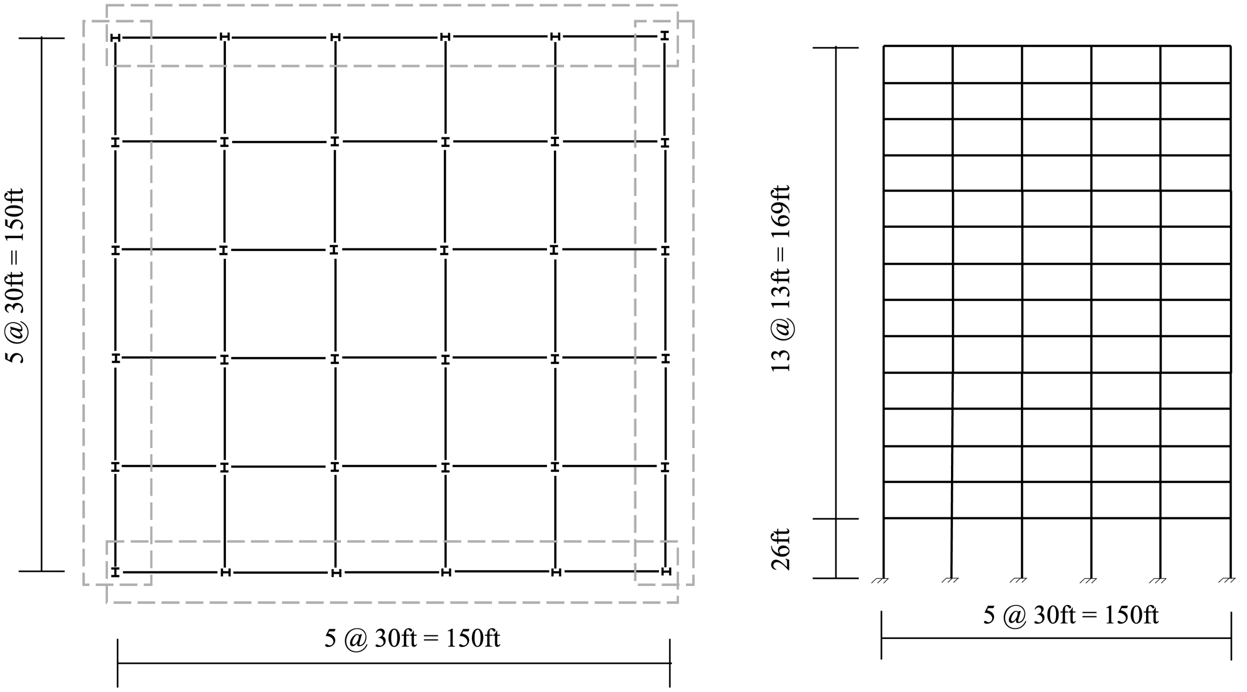


Figure 5.7: Archetype 1188: 14-story steel special moment frame

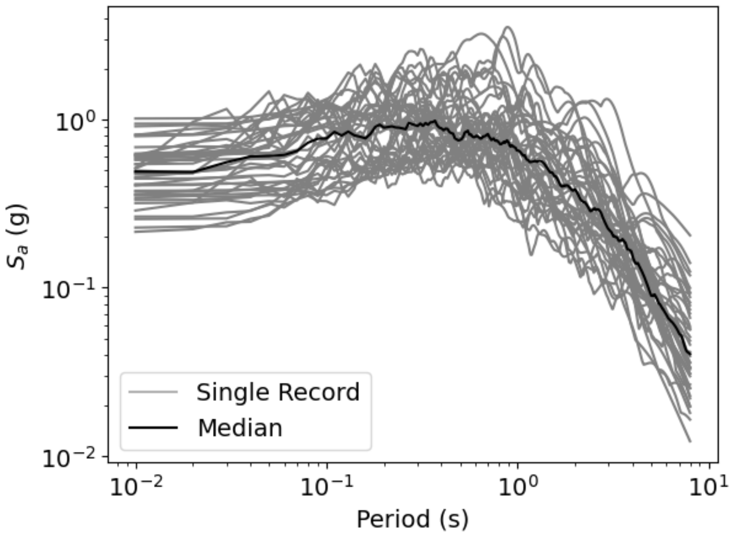
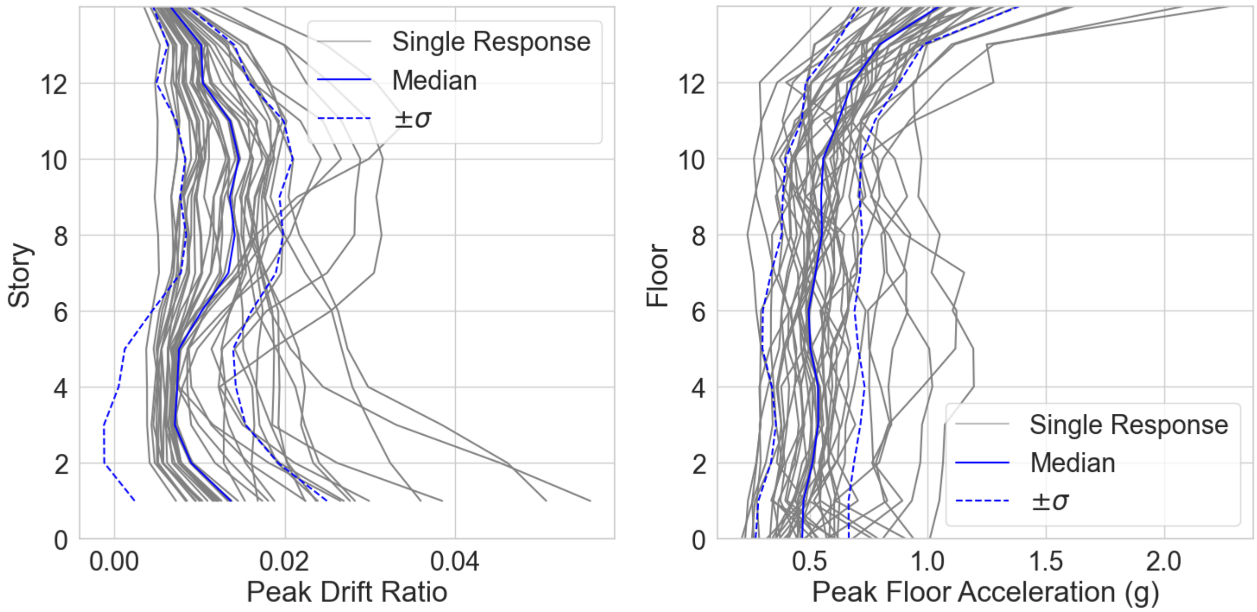


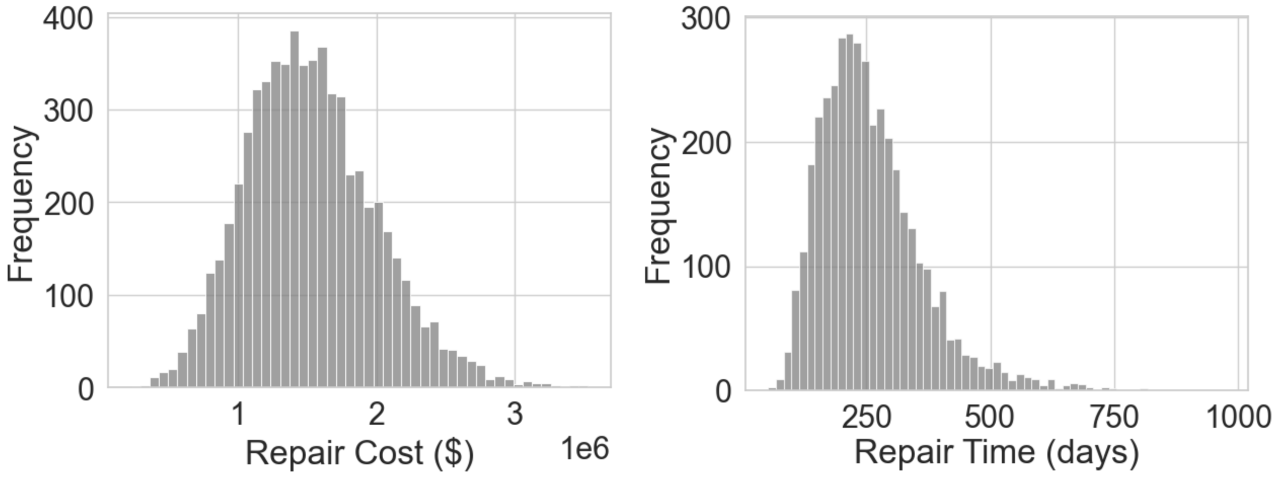
Figure 5.8: 5% damped spectra for the ground motions used in the structural response analysis



(a)

(b)

Figure 5.9: Full-profile structural responses including the median and log-standard deviation response for a) peak story drift ratio and b) peak floor acceleration



(a)

(b)

Figure 5.10: Estimated repair a) costs and b) times for the 10,000 realizations

external sources as well.

Table 5.3: Estimated repair cost and repair time statistics

	repair cost	repair time
	($)	(days)
count	10000	10000
mean	6.20E+07	403.40
std	4.92E+07	327.62
min	2.93E+05	0.17
25%	1.61E+06	0.87
50%	1.02E+08	670.00
75%	1.02E+08	670.00
max	1.02E+08	670.00

5.4 Summary

In this chapter, three previously developed automated design and analysis programs are unified, allowing for portfolio-level generation of building designs and full-profile analyses. An additional module that uses the pelicun-implemented FEMA P-58 methodology to perform loss assessment is also introduced. In addition to building-specific component information, the engineering demand parameters (EDPs) from the analysis modules are read into the loss module to perform the assessment.

Limitations of AutoSDA are inherited from each of its modules as well as in its intended use. A notable module-specific limitation is that WoodSDA requires substantial user inputs to be executed, especially when compared to the other modules. Templates are included in the top level folder of the program to help alleviate this, but deviations from the provided building layouts would require substantial user effort to initiate the WoodSDA module. The design and assessment of an entire region of buildings includes many simplifying assumptions, as the actual composition of a building portfolio would likely include significant irregularity. A major usability limitation of the program is the lack of a graphic user interface (GUI), which could make the program more accessible to those without a programming background. This would be especially useful for WoodSDA due to its extensive set of input files. This is a limitation that can be addressed in future versions. RCWall-SDA is limited in its displacement-based design approach, which is only applicable to continuous slender walls [5]. The loss assessment module requires an additional user input file containing the relevant component information. This module is also limited by the available component database provided in pelicun. Additional component fragility parameters must be manually provided by the user, which may not be feasible for large-scale assessments. FEMA P-58 provides a normative quantity tool that can be used to alleviate the limited component list in pelicun, but that tool is also limited as it does not include structural components. The loss module acts as a step towards portfolio-level loss estimation, but is limited by the pelicun implementation and databases.

The benefit of this unified program is that it allows for efficient multi-LFRS portfolio-scale design and assessments that can serve as an aid in determining where damage and losses have likely occurred during an event. This information could inform decisions on where to prioritize resources after an event, potentially reducing recovery times and downtime losses. The ability to generate portfolio-scale code-conforming designs and analyses may also be a powerful tool when limited instrumentation data is available. The following chapter shows an example use case, where the generated structural responses are combined with strong motion data to develop a generalized full-profile structural response prediction and loss estimation model for a portfolio of buildings.

CHAPTER 6

Development of a regional seismic impact assessment model for building inventories using hybrid data

6.1 Introduction

Structural health monitoring (SHM) has been shown to be an effective tool for enhancing civil infrastructure safety and resilience, particularly in seismically active regions. Developments in SHM have enabled real-time assessments of structural response during events and provide insights for engineers and policymakers in risk mitigation efforts. Often, traditional approaches to structural response reconstruction modeling rely solely on analytical models or on limited sensor data (e.g. partially instrumented buildings or regions of sparsely instrumented buildings). More recently, significant developments have been made to advance data-driven methodologies for structural response reconstruction. Many of these approaches take advantage of the predictive power of machine learning techniques, offering an alternative to conventional analytical models. However, existing models still face limitations in effectively capturing the complex and nonlinear dynamics exhibited by structures subjected to seismic excitation.

A previous publication [2], outlined in chapter 3, presented a dual model formulation to improve the predictive performance of a structural response reconstruction (SRR) model using only building strong motion response data. This model leveraged the spatially correlated peak ground acceleration (PGA) with Kriging to estimate the values at locations of uninstrumented buildings. The spatially interpolated PGA was then included as a feature in the ML model, improving predictions for the maximum peak floor acceleration (maxPFA), while maintaining performance for the maximum peak story drift ratio (maxPSDR). This was an alternative to another ML approach, performed in [110], that developed a cross-building response reconstruction (CBRR) model. It used a modified ground motion model (GMM) to estimate median structural responses combined with kriging on the within-event residuals. The model performed well overall, but was limited by its event-dependent approach. Despite the limitations, both of these studies were steps towards developing generalized response prediction models that can aid in rapid post-event assessment, which is crucial in natural hazards recovery efforts.

Many response reconstruction studies typically rely on a single type of data, either measured or simulated, while the use of hybrid data that integrates both measured and simulated responses to address data gaps remains less common. Existing literature on this hybrid framework is limited, often focusing on reconstructing ground motion data rather than structural responses. [127] developed a hybrid identification method for estimating structural parameters and ground motions in multi-story buildings. Similarly, [124] introduced a hybrid approach, which improved signal-to-noise ratio and data recovery efficiency. Recently, a study by [52] proposed a two-step hybrid method for reconstructing seismic responses in partially instrumented buildings, utilizing a shear-flexural beam model calibrated with data from instrumented floors, and Gaussian Process regression to model residuals. While these studies enhanced our understanding of hybrid-data response reconstruction, they either focused on ground motion parameter (not response) reconstruction or the measurement data in the hybrid set was obtained from simplified models (as opposed to strong motion data). Moreover, all of the prior studies in this area focused on individual buildings and none of them incorporated performance-based assessment.

This study introduces an approach using data augmentation to address some of the shortcomings of existing methodologies by integrating simulated structural response data with measured responses from instrumented buildings. An Extreme Gradient Boosting (XGBoost) model is used to develop a full-profile reconstruction of building responses at the regional scale. Augmenting the measured data with simulated response data enables the

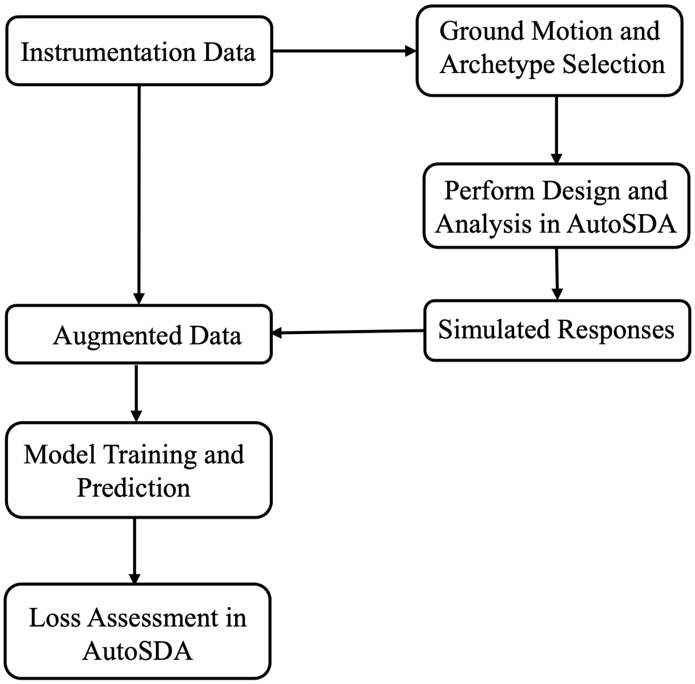

Instrumentation Data
Ground Motion and Archetype Selection
Perform Design and Analysis in AutoSDA
Simulated Responses
Augmented Data
Model Training and Prediction
Loss Assessment in AutoSDA

Figure 6.1: Hybrid data-driven structural response reconstruction and regional impact assessment procedure

model to learn from a broader spectrum of structural behaviors, resulting in a more robust model. Moreover, by incorporating regional seismicity data, the model can account for variations in ground motion characteristics, further improving its predictive capabilities. This approach to structural response reconstruction facilitates more accurate damage assessment than models based only on simulated or measured data alone. The study culminates in using the structural responses from the predictive model to perform a regional seismic impact assessment using the repair costs associated with earthquake-induced damage as the performance metric. The overall procedure for the hybrid data-driven SRR model and regional impact assessment is outlined in Figure 6.1.

.

6.2 Description of the Data

The data set used in this study consists of strong motion recordings for buildings across the state of California from various historical events, as well as simulated responses from nonlinear structural models using OpenSees. Ground motions from a subset of the event set are applied to the simulated models to obtain the response parameters of interest. The augmented nature of the data is an important aspect in this study that compensates for the inadequacies of the measured data alone. In general, buildings are only partially instrumented, typically containing sensors at the ground and roof levels, and few intermediate levels, if any. Of course, this varies building to building, but none of the buildings in this study are fully instrumented, and this is true in most cases. This means that other techniques need to be employed in order to obtain full-profile responses. Interpolation and machine learning methods are commonly used to fill in this information gap. The current study attempts to alleviate this issue by augmenting the data set with simulated structural response data, where the full-profile responses of analytical models are available. The ultimate goal is to leverage this expanded set to make predictions on the full-profile responses of uninstrumented or partially-instrumented buildings, which can be used as inputs into the loss assessment.

6.2.1 Event Set and Instrumented Buildings

The data used in this study consists of instrumentation data from buildings across the state of California, subjected to nearly 40 different events spanning from 1984 to 2020 with $\mathbf{M} \geq 4$. Information on the events, their locations, and the number of records for each event can be found in chapter 2. The simulated set uses a subset of these events with $\mathbf{M} \geq 6$, which are highlighted in Table 6.1. The instrumented recordings were collected from the Center for Engineering Strong Motion Data (CESMD) database [25]. The ground motions applied to the nonlinear structural models in OpenSees were collected from various station recordings from the PEER NGA-West2 database [15]. The events in bold are the events whose ground motions were applied to the OpenSees models. Eight of the events with **M** ≥ 6 were selected from the set and ground motions from these events. Five ground motion recording stations were selected for each event, resulting in a total of 40 ground motions applied to each structural model. Figure 6.2 shows the 5% damped acceleration spectra of the applied ground motions, as well as the median spectrum. Figure 6.3 shows a map of station locations and the recorded ground motions for the Northridge Earthquake that are applied to the simulated building models to illustrate the ”placement” of simulated buildings. The event epicenter is indicated by the red dot, and the blue dots are the station locations. As the source-to-site distance will be a feature in the predictive model, it was important to obtain ground motion recordings at various distances from the event.

For each event, the data collected includes the magnitude, location, and depth of the epicenter, shown in Table 6.1. For each building that has recordings for any of the events in the set, the location, number of stories, story heights, Lateral Force Resisting System (LFRS), ASCE 7-16 estimated period, and response history data are collected. These features will be important in the predictive model for reconstructing the full-profile responses. After data selection, the combined data set contains 718 unique data points that are included in the analysis.

6.2.2 Nonlinear Structural Models

The structural models used to develop the simulated data set were generated using the Automated Seismic Design and Analysis (AutoSDA) platform, which is powered by OpenSees. AutoSDA is a compilation of three previously developed design and analysis programs (steelSDA, woodSDA, and rcwallSDA, [53, 36, 5]), created for their respective lateral force resisting system types. An additional module for performing loss assessment was developed and implemented into the platform.

For each LFRS type in AutoSDA, multiple archetypes were used to represent various building configurations. Each archetype was subjected to the ground motions collected from PEER-NGA. The seismic parameters were estimated based on the station locations where

Table 6.1: Summary of events for which ground motions are used in the nonlinear response history analysis

Name	Depth (Km)	M	Date	Latitude	Longitude
Morganhill	9	6.2	4/24/84	37.32	-121.68
Whittier	9.5	6.1	10/1/87	34.06	-118.07
LomaPrieta	18	7	10/17/89	37.04	-121.88
BigBear	1	6.5	6/28/92	34.20	-116.83
Landers	1.1	7.3	6/28/92	34.22	-116.43
Northridge	19	6.4	1/17/94	34.21	-118.55
SanSimeon	4.7	6.5	12/22/03	35.71	-121.10
Parkfield	7.9	6	9/28/04	35.81	-120.37

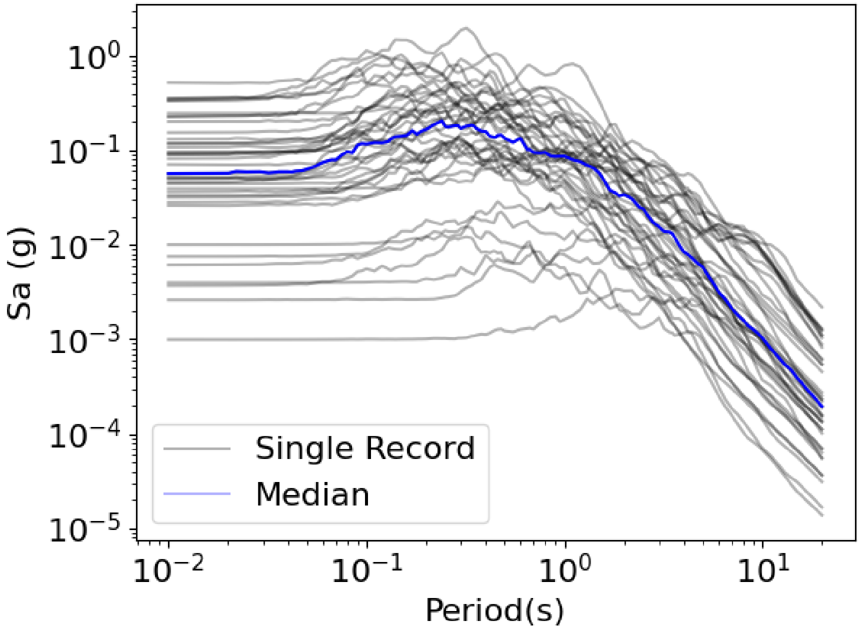


Figure 6.2: Spectra of 40 ground motions used in the nonlinear response history analyses

the ground motions were recorded. After applying all ground motions to each archetype, the peak responses were recorded. When ”moving” an archetype to a new station location, this is considered a new data point in the set. This not only refers to the seismic parameters, but the source-to-site distance is based on the station location as well. The archetypes are effectively placed at the locations of the ground motion recording stations. For example, ground motions from eight of the events in the set are applied to the simulated models. Five stations from each event are selected for the ground motions records. So, there are 40 ground motions applied to each model (80 when considering both directions). For the steel moment frame buildings, 5 archetypes are created, resulting in 40 ground motions x 5 archetypes = 200 data points for just this LFRS type. After all data points are generated for each LFRS type, only a subset is used in the final augmented set. Some responses may be too small. Also, in rare cases, the simulation may detect collapse, and in the final case, the buildings are selected such that the feature distributions of the simulated and instrumented sets are similar. Additionally, it was decided that the number of simulated points should not exceed the number of instrumented as the simulations are only meant to aid the predictions of uninstrumented or partially instrumented buildings, not create a model that is good at predicting simulated building responses. The goal is, again, to develop a generalized response prediction model that will better capture the full-profile response of buildings than an instrumentation-only set. The loss assessment, which utilizes FEMA P-58 methodology also depends on demand data at each floor and story of a building to estimate potential losses and repair times. Missing data is a limitation for this type of assessment, such as with partially instrumented buildings.

6.2.2.1 Steel moment frame archetypes

The steel frame building models are based on building designs generated and stored in the database developed in [78]. The database of 621 code-conforming designs consists of 1, 5, 9, 14, and 19-story steel moment frame buildings with various configurations. One building from each class of number of stories was randomly selected from the database to be used as

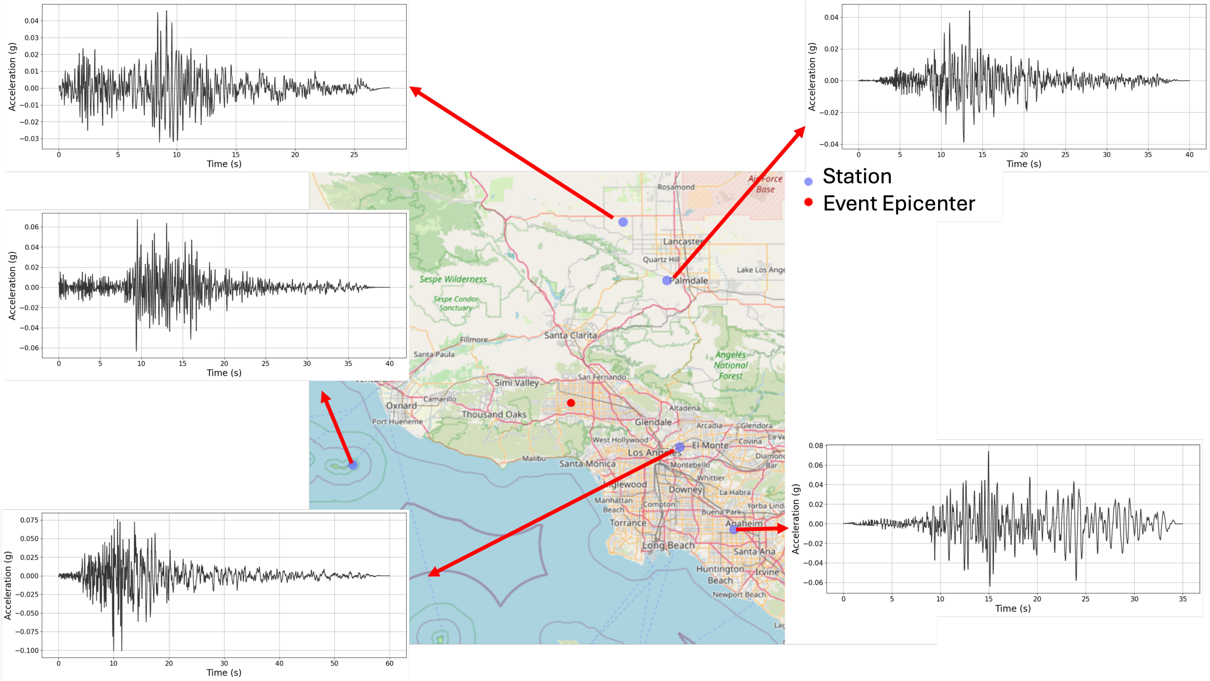


Figure 6.3: Map showing the stations where ground motions for the Northridge earthquake were recorded

archetypes for generating the structural models and responses in AutoSDA. Figure 6.4 shows a visual representation of one of the steel moment frame archetypes used in the study. A summary of the selected steel moment frame archetype geometries are outlined in Table 6.2. The inputs include information about the building geometry and the loading conditions. The loading conditions for each archetype can be found in Table C.1

6.2.2.2 Woodframe archetypes

The woodframe buildings in the set were obtained from the Applied Technology Council (ATC) 116 project archetypes [11]. The archetypes are for 1, 2, and 4 story woodframe buildings. Basic inputs for these archetypes are shown in Table 6.3, but it should be noted that woodSDA requires significantly more input information than the other modules in order to be executed. Extensive information on the panel properties and loading are also necessary. To aid in the creation of the input files, there are templates included in WoodSDA, also based

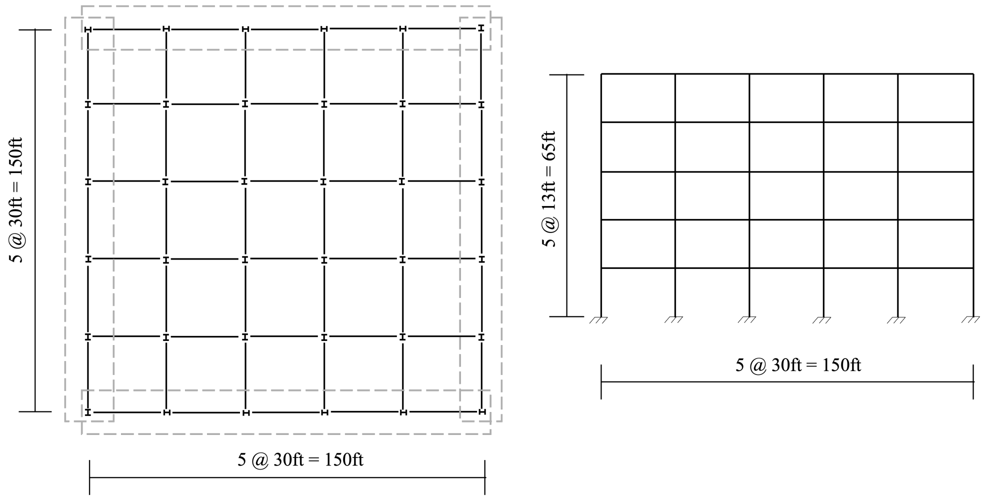


Figure 6.4: Archetype 279: Five Story Steel Moment Frame

Table 6.2: Steel moment frame archetype building geometry

Archetype ID	No. Stories	No. X Bays	No. Z Bays	1st Story Height	Typical Story Height	X Bay Width	Z Bay Width	No. X LFRS	No. Z LFRS
22	1	3	3	26	13	20	20	2	2
279	5	5	5	13	13	30	30	2	2
765	9	5	5	13	13	30	30	2	2
1188	14	5	5	26	13	20	20	2	2
1566	19	5	5	13	13	20	20	2	2

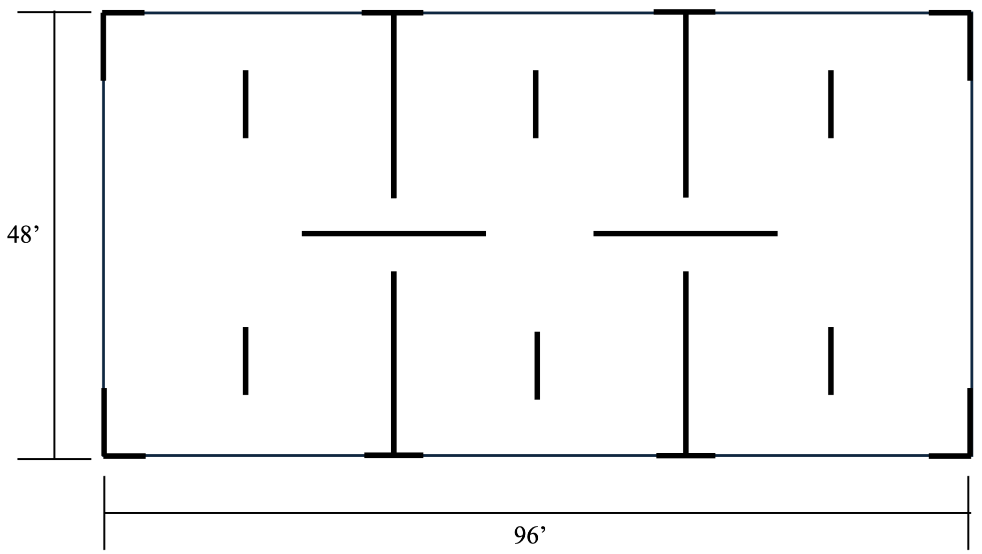


Figure 6.5: Floor plan for woodframe building Archetype 196481

on the ATC-116 archetypes. The floor plan for one of the woodframe archetypes is shown in Figure 6.5. Information on the building loads are listed in Table C.2.

6.2.2.3 Concrete shear wall archetypes

Archetypes developed by [5] were used to create the concrete shear wall models. The archetypes have the same floor plan, but vary based on the number of stories in the building as well as the length of the shear walls. A typical plan for the RCWall building type is shown in Figure 6.6. The building geometry for each archetype is outlined in Table 6.4 and the loading conditions are presented in Table C.3. Because the archetypes all share the same plan configuration, a few more archetypes are selected for the RCWall buildings to introduce more variation into the set for this building type.

6.2.3 Component Population

For the loss assessment module, information such as the total replacement cost of a building and component definitions are needed. The component populations were generated using the Seismic Performance Prediction Platform (SP3) [57], which relies on a database of building

Table 6.3: Woodframe archetype building geometry

Archetype ID	No. Stories	Length (ft)	Width (ft)	Floor Areas (ft^2)	Floor Max X Dimension (in)	Floor Max Z Dimension (in)	Story Height (in)	No. X Wood Panels	No. Z Wood Panels
148321	1	48	32	1536	576	384	120	13	8
196481	1	96	48	4608	1152	576	120	17	16
248321	2	48	32	1536	576	384	120	13	8
296481	2	96	48	4608	1152	576	120	18	9
496481	4	96	48	4608	1152	576	120	16	16

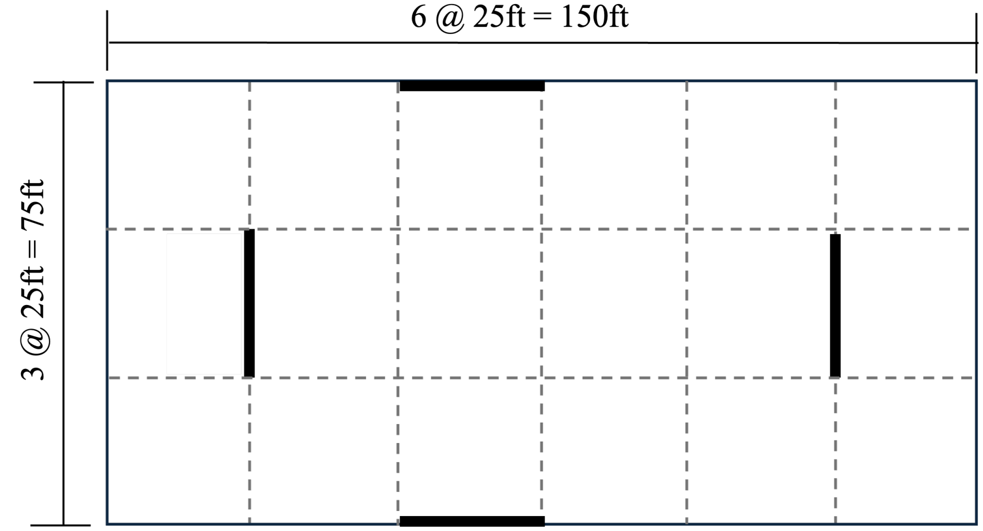


Figure 6.6: Typical floor plan for reinforced concrete shear wall building archetypes

Table 6.4: Summary of geometric information for RC shear wall archetypes

Archetype ID	No. Stories	Building length (ft)	Building width (ft)	No. Parallel Bays	No. Perpendicular Bays	1st Story Height (ft)	Typical Story Height (ft)	Wall Length (in)
1	2	75	150	3	6	12	12	144
2	2	75	150	3	6	12	12	120
3	4	75	150	3	6	12	12	288
5	4	75	150	3	6	12	12	192
7	8	75	150	3	6	12	12	576
11	8	75	150	3	6	12	12	288
12	12	75	150	3	6	12	12	864
16	12	75	150	3	6	12	12	432

and component information collected from FEMA. In order to generate the potential losses for a portfolio of buildings, as is the aim of this study, unique building models were created and used to generate likely components based on attributes such as the LFRS-type, occupancy type and number of stories. The program outputs a table that lists the components as well as their locations within the building and fragility parameters. The components, building area, estimated costs per square foot and median collapse capacity and dispersion in terms of Sa_{T_1} for the building models are the only information being used from the program. The median collapse capacity, which are used to generate a global collapse fragility, is estimated in SP3 as the spectral acceleration at which the building has a 50% probability of collapse. This is calculated using FEMA P-154 [49], which provides a simplified scoring procedure.

The building occupancies for the instrumented set from CESMD were matched using the station name. Many of the names included common phrases such as ”Hospital”, ”School”, or ”Office Building”, which made assigning their occupancies straight forward. FEMA P-

58 uses occupancy designations (Warehouse, Healthcare, Education, Hospitality, Single and Multi-Unit Residential, and Commercial Office) to classify buildings by use. The simulated set of buildings were randomly assigned one of these occupancies. The unique combinations of LFRS-type, occupancy type and number of stories were extracted from the data, resulting in approximately 90 unique component population models created in SP3. After each model was created, the component information was compiled into csv files to be used in the loss assessment. When the loss assessment is run, the building is matched to one of the population models, so that the appropriate component information csv file is selected for the assessment. An example of the component information as well as component fragilities that can be collected from SP3 are outlined in chapter 5.

6.3 Description of Machine Learning Model

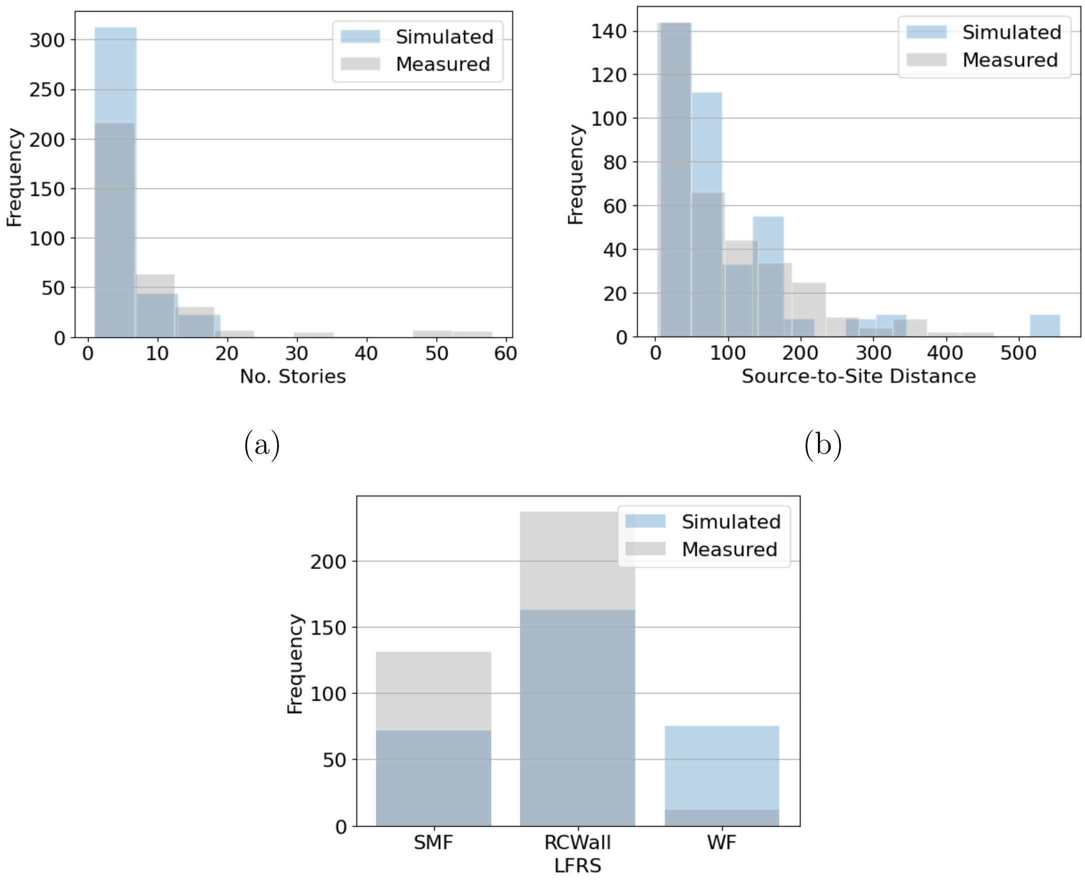
For this study, an extreme gradient boosting (XGBoost) model is used to perform the full-profile EDP predictions. A detailed description of XGBoost models is provided in chapter 3. For model training, a 70-30 train-test split is used with the augmented set (instrumentation + simulated data). Uncertainty quantification is not native to XGBoosting models, but will be necessary to propagate to the loss assessment. To alleviate the lack of native support, bootstrapped boosting is performed over 1000 repetitions, and the median and logarithmic standard deviation of the predictions will be used for the loss assessment. The simulated data serves two purposes: in training, it provides full-profile response data that the instrumentation set lacks, and in testing, it acts as a full-profile response prediction validation set. Two separate models are trained for comparison. One model is trained on the measured data only, and the other is trained with both simulated and instrumented data.

Features used in model training include the building height (H_j for building j), first mode period (T_1), source-to-site distance (R), event magnitude (M), Base Area of the building (A), one-hot encoded LFRS-type, and the story height ratio (SHR). The SHR, shown in Equation 6.1, is essentially a normalized story height, which is used in lieu of the absolute story height (h_i at story i) value, as it is expected that the relative position in the building is more useful in terms of potential correlational relationships with the response. This feature will also allow for more apt comparison of model performance with a data set that has a wide range of building heights and number of stories. The simulated building set is selected such that the number of simulated data points did not overwhelm the measured data, leaving approximately half of the training points as instrumented and half as simulated. The feature distributions of the buildings across the two groups are also similar. In Figure 6.7, the distributions for the number of stories, source-to-site distance, and LFRS types are shown. The primary goal of this model is to assess how the data augmentation process may provide a more generalized model when predicting full-profile structural responses. The performance at various SHRs will be evaluated and compared between the two models to directly address this point.

$$SHR_{ij} = \frac{h_i}{Hj} \tag{6.1}$$

6.4 Full-Profile Response Prediction

Figure 6.9 shows the observed vs predicted points with a red line indicating perfect agreement. These plots show, similarly for PFA and PSDR that the measured-only training set does not perform as well as the augmented set. This is likely due to the additional information on intermediate SHR values for buildings in the augmented set. To assess the fit of the models to the training and testing sets, the R^2 values are shown in Table 6.5. The measured-only model fits well to the training data, but does not perform well on new data, as shown by the significantly lower R^2 for the test set. This is not surprising, given the model is being expected to predict full-profile structural responses with only partially-instrumented building data for training. The augmented model shows similar fits to the training and testing set, meaning it generalizes well for predictions at new data points. Again, this follows expectation as this model has full-profile response data in its training set, and therefore, is better able to make full-profile response predictions.



(a)

(b)

(c)

Figure 6.7: Comparison of building characteristics in the simulated versus measured data sets: a) number of stories, b) source-to-site distance, and c) LFRS type

Table 6.5: R^2 values for both models

	Augmented		Measured	
	Train	Test	Train	Test
PFA	0.99	0.93	0.99	0.25
PID	0.99	0.90	0.99	0.38

$$NRMSE = \frac{1}{Y_{max} - Y_{min}} \sqrt{\frac{\sum_{i=1}^{N} (Y_i - \hat{Y}_i)^2}{N}} \tag{6.2}$$

The normalized root mean squared errors (NRMSE) of the two models are compared for the overall models as well as at binned SHR levels. In Equation 6.2, Y_{max} and Y_{min} are the maximum and minimum values of the observed target, and Y_i is the observed target for data point i. $\hat{Y}_i$ is the predicted target for data point i. The NRMSE is 0.191 and 0.108 for PSDR, for the measured and augmented models, respectively. This shows that the error is nearly twice as large for the measured-only model. Similarly, for PFA, the values are 0.130 and 0.062. In addition, Table 6.6 and Table 6.7 describe the performance at various SHRs in the buildings. The augmented model outperforms the measured model at every binned level. An important detail to point out is that at the floor, roof, and mid levels of the buildings, the measured model performs best. This is consistent with our intuition because these are often locations at which sensors are placed. The model performance decreases at the intermediate levels where sensors are less often placed. Figure 6.8 shows a histogram of the SHRs where sensors are located in the instrumented buildings. From this figure, the most common locations are the ground level, roof level, and a slight peak around an SHR value of 0.5, which corroborates the findings.

The superior predictive performance of the model trained on augmented data can be seen in Figure 6.11 and Figure 6.12, where the full profile response predictions for a 12-story simulated building and 10-story instrumented building in the test set are visualized. The augmented model better captures the shape of the full-profile response of the simulated

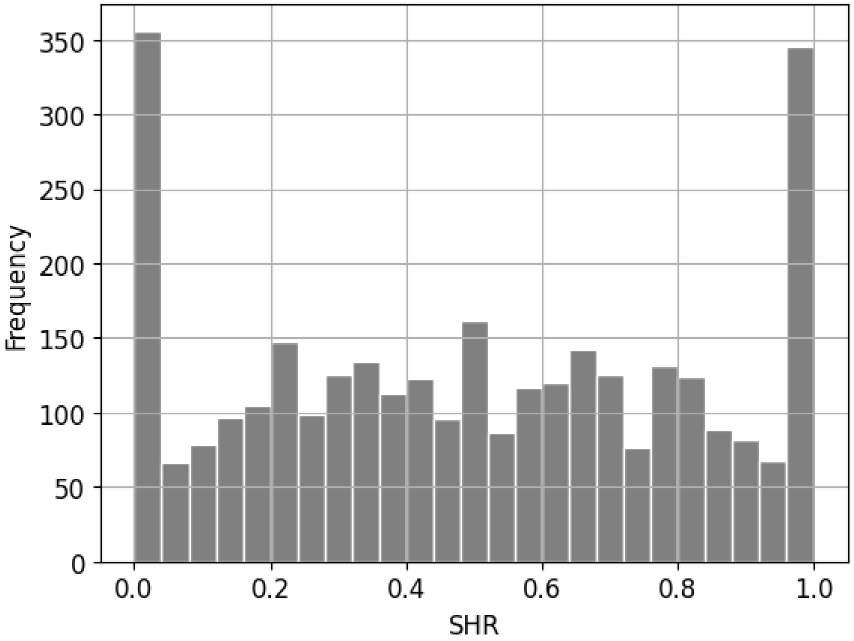
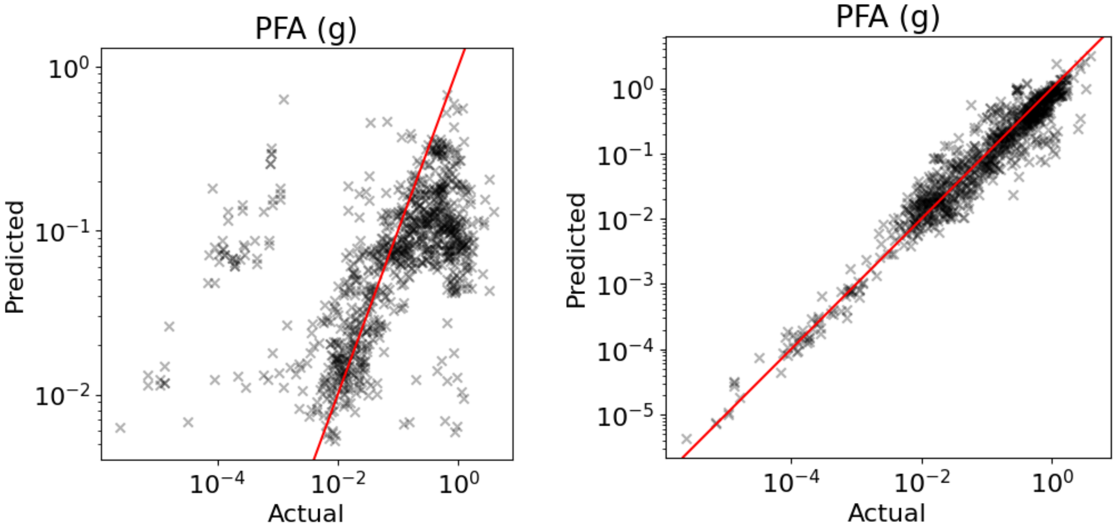


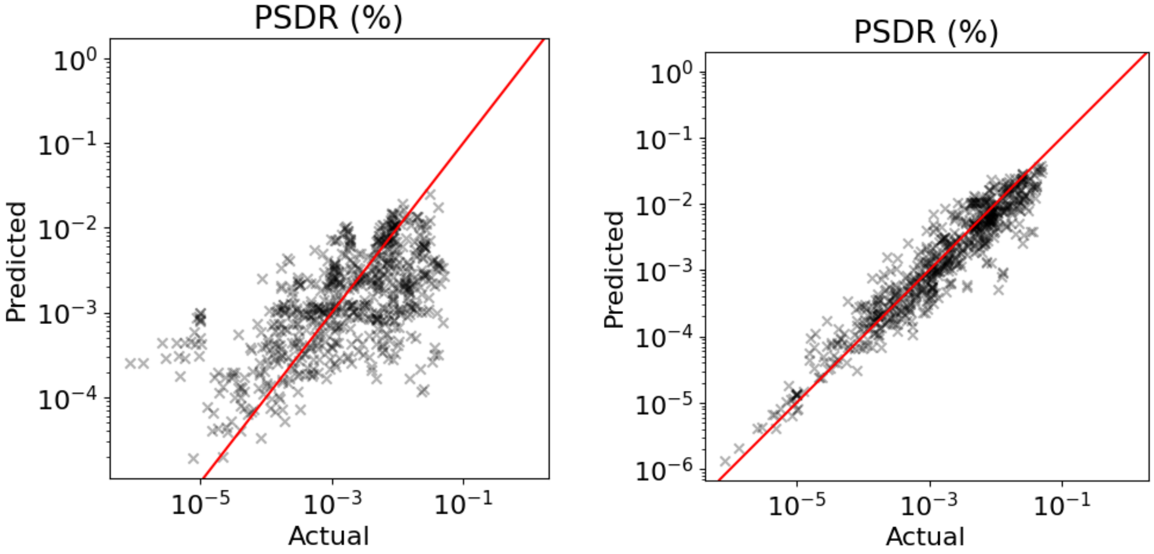
Figure 6.8: Histogram of story height ratios in the instrumented set



(a)

(b)

Figure 6.9: Predicted vs Actual for PFA a) measured-only trained model and b) augmented-trained model



(a)

(b)

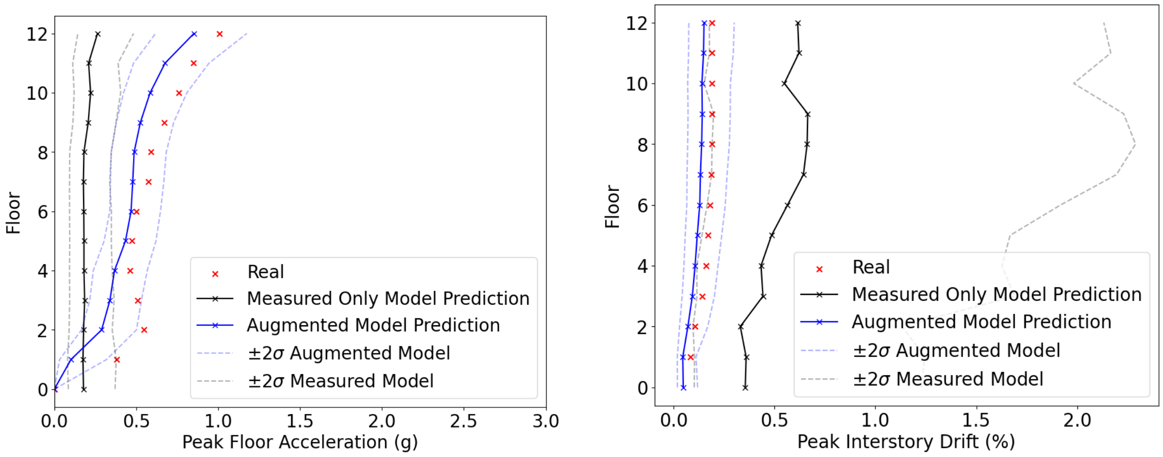
Figure 6.10: Predicted vs Actual for PSDR a) measured-only trained model and b) augmented-trained model

Table 6.6: NRMSE for PFA

SHR	Augmented	Measured
0	0.083	0.129
(0,0.2]	0.144	0.342
(0.2,0.3]	0.131	0.301
(0.3,0.4]	0.147	0.332
(0.4,0.6]	0.121	0.251
(0.6,0.8]	0.126	0.284
(0.8,1]	0.131	0.329
1	0.070	0.155

Table 6.7: NRMSE for PSDR

SHR	Augmented	Measured
0	-	-
(0,0.2]	0.163	0.246
(0.2,0.3]	0.153	0.231
(0.3,0.4]	0.109	0.224
(0.4,0.6]	0.100	0.162
(0.6,0.8]	0.099	0.207
(0.8,1]	0.154	0.278
1	0.118	0.221



(a)

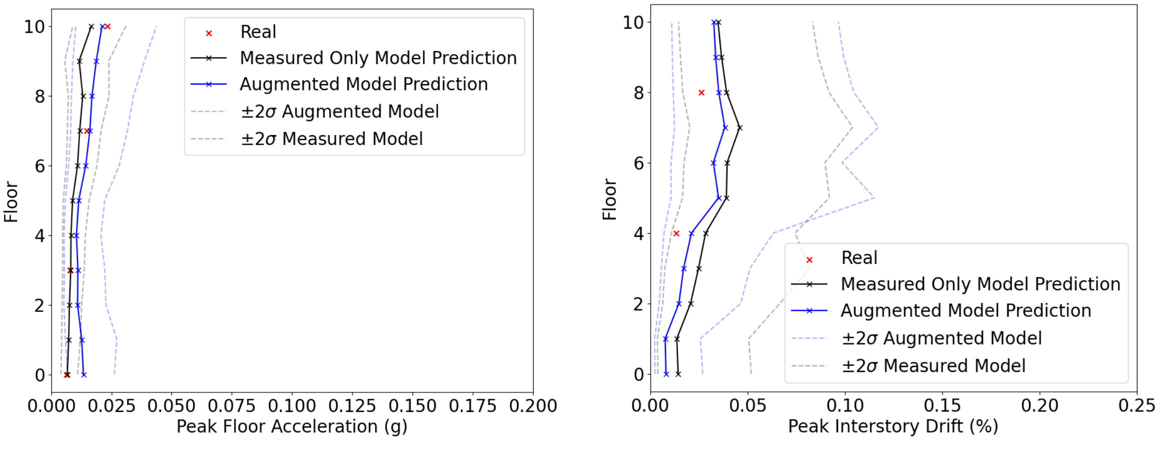
(b)

Figure 6.11: Full-profile response predictions of a 12-story simulated building in the test set

building, with narrower confidence bands. In the second example, showing the response of a partially instrumented building, the performance is more similar with respect to the information available on the buildings response. This is promising because the simulated dataset appears to at least meet the performance of a measured-only trained model predicting responses of instrumented buildings. The ability to better predict the full-profile response of a building reduces reliance on the instrument location for rapid response reconstruction, though it does still introduce uncertainty. The uncertainty in these predictions will be propagated to the loss assessment, replacing what would traditionally be the median and logarithmic standard deviation of the demand obtained through nonlinear response history analysis, with the median and logarithmic standard deviation of the predicted demand from the ML model.

6.5 Assessment of Potential Losses

In traditional approaches, the distributional demand parameters that would be used in the loss assessment would be obtained by applying a suite of ground motions to a structural



(a)

(b)

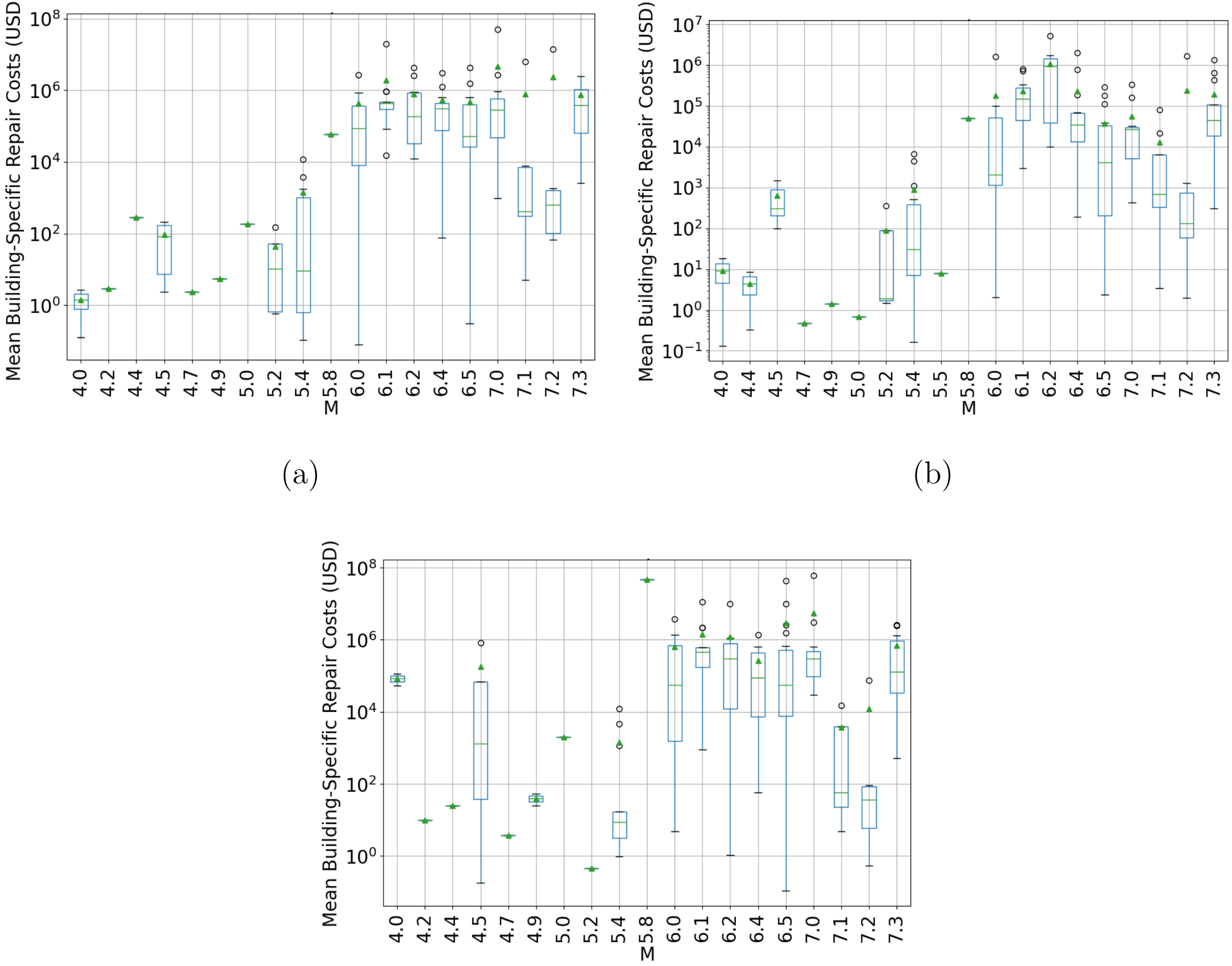
Figure 6.12: Full-profile response predictions of a 10-story instrumented building in the test set

model and calculating the median and logarithmic standard deviation of the demand. Without generating a structural model for each instrumented building in the set, we are unable to perform the necessary analyses to do this for these buildings. The current implementation seeks to perform this loss assessment incorporating real demands of the instrumented buildings by predicting the full-profile response with an ML model. Based on the predictions and the model uncertainty obtained in the model application, the FEMA P-58 loss assessment in pelicun is run. The predictions and uncertainty from the model are used as the demand median and the log-standard deviation that define the lognormal sampling distribution for the assessment. Each building is matched to a component population file, as outlined in 6.2.3. For comparison to the real responses in the test set, and as an estimate of the uncertainty in the measured responses, the maximum allowable accelerometer noise of $0.03mg$, as outlined in the CSMIP guidelines for instrumentation [22], was assumed as an approximation of the dispersion for the loss assessment. For accelerometers, noise is often described in terms of Root-Mean-Square (RMS) noise. As the noise type was not specified, this was assumed. The equation for RMS is shown in Equation 6.3, which is very similar to the standard deviation calculation for data centered around 0. A brief example illustrating the loss assessment on an individual building level is outlined in chapter 5. This process is performed on each building in the set in this study, but the results are aggregated for a regional-scale assessment.

$$X_{RMS} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} x_i^2} \tag{6.3}$$

Figure 6.13 and Figure 6.14 show box plots of the mean building-specific repair costs and repair times, separated by event magnitude for each of the 2 model predictions and the test set. The plots show a general upward trend in repair costs as the magnitude increases. The wider box plot whiskers in the estimate based on measured model predictions reflect the higher uncertainty. Again, the demand realizations are generated from the uncertainty in the ML model predictions. It is important to note that the test set contains missing data, as the non-instrumented floors in the building do not have recordings. These values are left as NaN in the analysis, and do not contribute to the damage and loss calculations. Given the limited response range of the data set, the loss analysis returned that some buildings did not experience any damage in any of the 10,000 realizations of the response. The loss estimates based on the predictions from the augmented model returned 33 undamaged buildings, the measured model returned 29, and the test returned 38. The box plots in the mentioned figures omit the buildings that were not assessed to experience damage for visualization purposes. The distributions of aggregated repair costs for each event are shown in Figure 6.15. The figures show similar distribution of the repair costs across the different models. It is notable that the test set often underestimates the repair costs, in some cases, reporting no repair costs, where the other models report minor or moderate costs. This is expected behavior given that the trained models have more information on the intermediate floors.

Figure 6.16 shows a positive trend for the mean building-specific repair cost against the individual building demand, with respect to Sa_{T_1}, for all 3 conditions. While the models following the expected general trends is promising, in a more practical setting for rapid response and recovery, the loss would be estimated at an event-level. Even with the augmented data, the size of the dataset is still limited relative to a regional-scale assessment, but the example provided in Figure 6.3 illustrates a potential use for the model implementation, especially

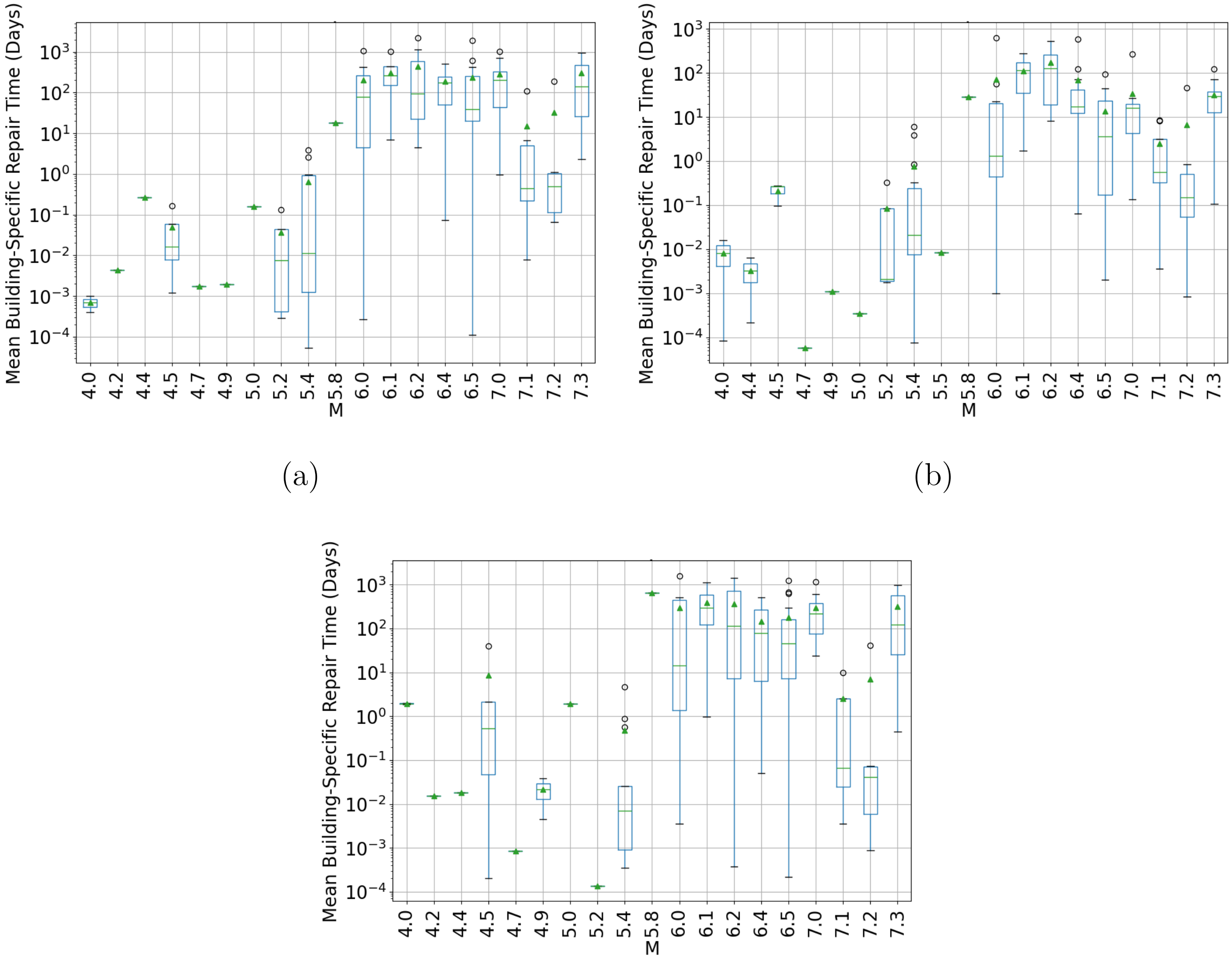


(a)

(b)

(c)

Figure 6.13: Box plots showing the relationship between the magnitude of mean building-specific repair cost for buildings experiencing damage for the a) augmented model predictions, b) measured model predictions, and c) test set

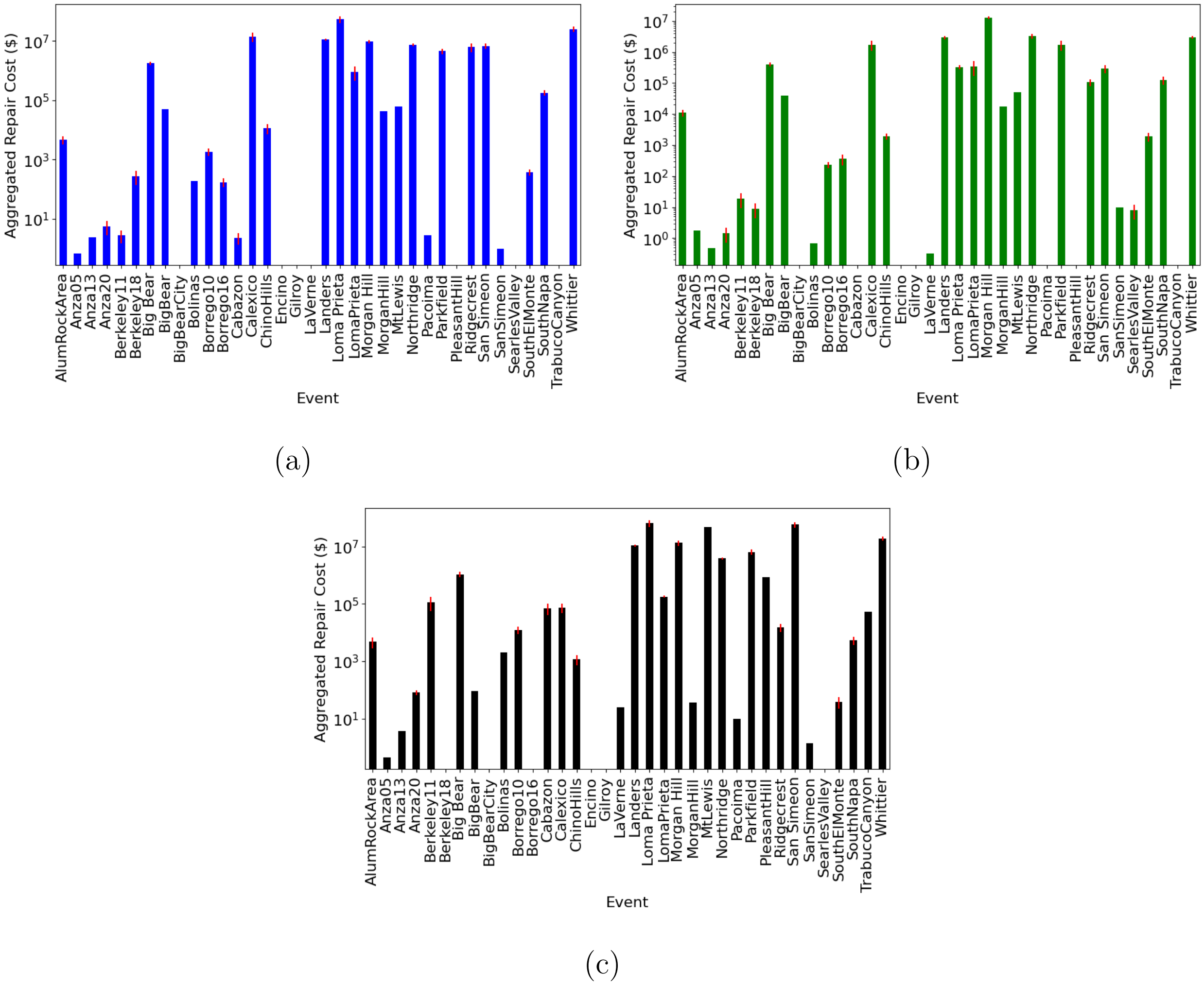


(a)

(b)

(c)

Figure 6.14: Box plots by magnitude of mean repair time for buildings experiencing damage for a) augmented model predictions, b) measured model predictions, and c) test set



(a)

(b)

(c)

Figure 6.15: Aggregated repair costs by event for a) augmented model, b) measured model, and c) test set

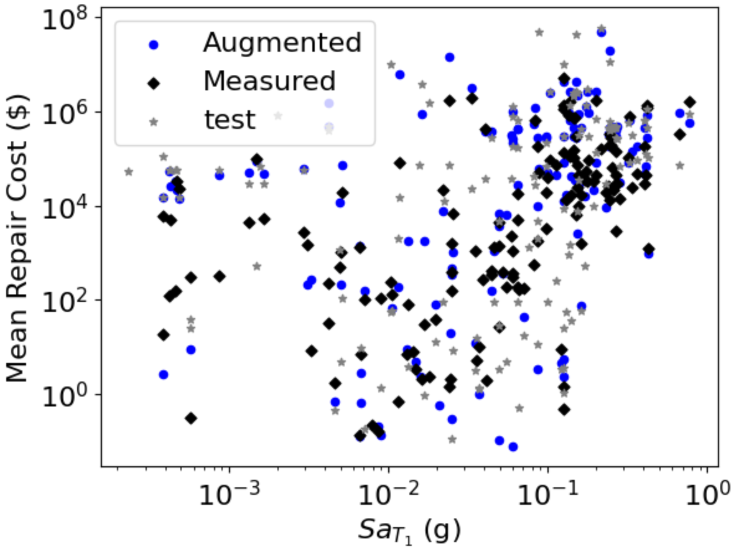
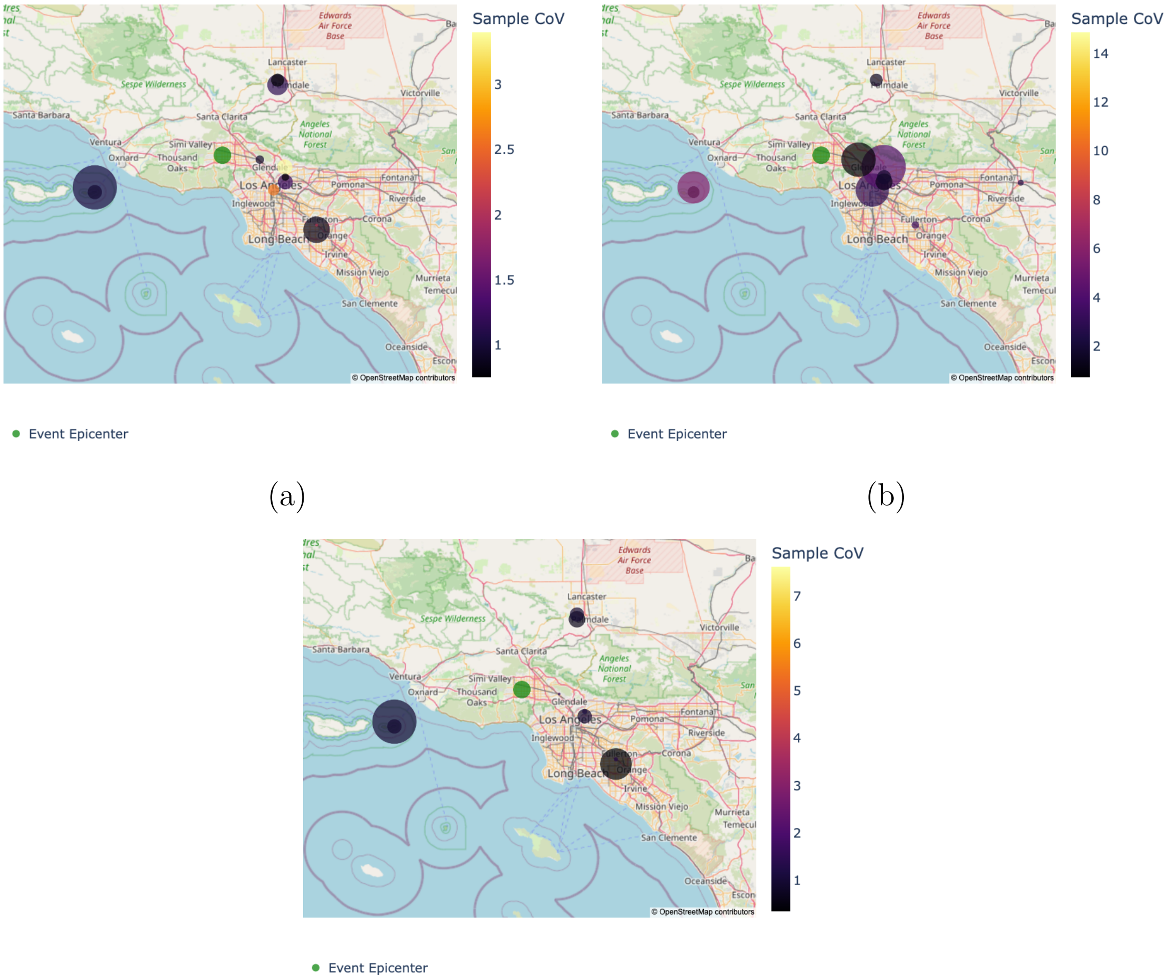


Figure 6.16: Mean repair cost vs median demand (Sa_{T_1})

given a much larger and more robust dataset. The Northridge earthquake was selected from the set, as it is one of the events with the most recordings. The buildings in the test set corresponding to this event are plotted on a map with the event epicenter. The individual repair cost for each building is normalized by its replacement cost, and the coefficient of variation of the Monte Carlo samples generated for a building are scaled by color to compare the uncertainties in the estimates. The map generated based on losses from the augmented model predictions show that there is, in general, lower uncertainty in the estimates compared to the measured and test set models. It appears that the model trained on measured-only data makes larger predictions with large uncertainty. This may be explained by the fact that instruments are often placed in locations in buildings that are expected to experience larger responses. So, the predicted building responses are based on training data that are generally larger in magnitude. These maps are shown as a way to compare the loss output from the various models. Figure 6.18 shows a map with an expanded set by combining the augmented model predictions and the training set, in order to better illustrate what a more realistic map with more data may look like. These maps are similarly scaled based on size and color by the normalized repair cost/time and coefficients of variation.

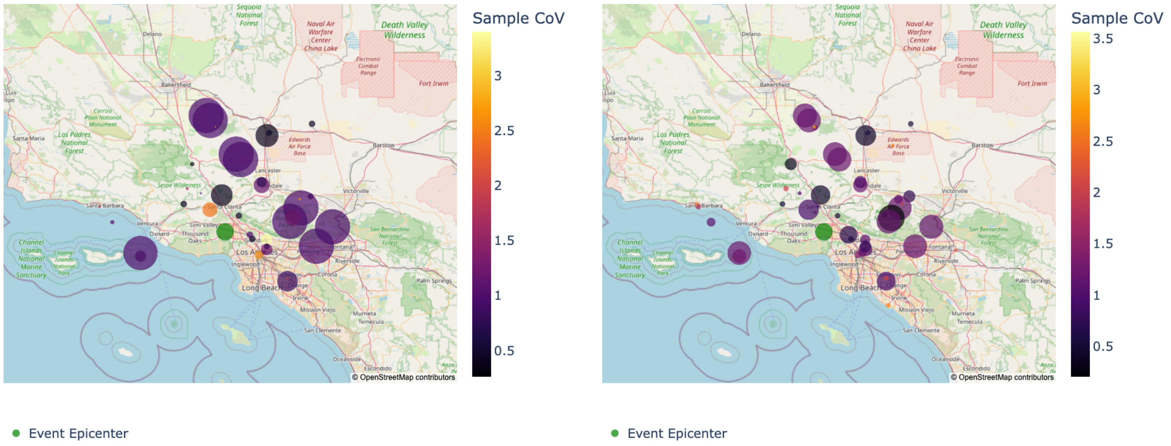


(a)

(b)

(c)

Figure 6.17: Map showing building locations, color coded by median-normalized repair cost for a) augmented model, b) measured model, and c) test set



(a) (b)

Figure 6.18: Map showing building locations of the combined training set and augmented model predictions with color scaled by mean and size scaled by standard deviation for a) repair cost and b) repair time

6.6 Conclusions

In this chapter, a hybrid data approach to regional response reconstruction and loss estimation is introduced. Strong motion data collected from buildings across the state of California are combined with simulated responses to create a more generalized training set for predicting full-profile structural responses. An Extreme Gradient Boosting (XGBoost) model is trained on the hybrid data (measured + simulation) and used to predict the full-profile responses on a hold out set. The model trained with augmented data significantly outperformed the one trained on only measured data, better capturing the profile of the test set responses. This can be attributed to the larger training set and the full-profile response data (the latter is not included in the measured-only model). The normalized root mean squared error (NRMSE) was compared across the two models for overall performance, as well as at various story height ratios (SHR). The model trained on the hybrid data had an overall NRMSE of 0.108 and 0.062 for PSDR and PFA, respectively. The predictions from the measured-only model were 0.191 and 0.130. This shows that the data-augmentation doubled the predictive performance for this metric. Also, the hybrid data model performed better at every level of the binned SHR values. An instrumented building and a simulated building from the test set are used as an example to further highlight the performance, with the hybrid data model predictions better capturing the response shape with lower uncertainty.

The median and log-standard deviations of the predictions, obtained from bootstrapped sampling, are used as input into the loss assessment performed using the pelicun package in Python. The estimated repair costs and times obtained from each model are compared and mapped to highlight spatial patterns in the predictive performance. The overall trends in the augmented, measured, and test model repair costs/times are similar, with the latter two models consistently showing higher degrees of uncertainty in the estimates. This larger uncertainty of the measured model compared to the augmented model is explained by the propagated uncertainty from the response predictions. The larger uncertainty in the test set relative to the augmented set is likely due to the missing data. The instrumented buildings in the test set do not have the full-profile responses, as sensors are not placed at every floor. So, the responses at these floors are omitted from the assessment, increasing the uncertainty in the estimates.

The current implementation of this hybrid data model for regional assessments has various limitations. While the approach to utilize the uncertainty in the predictions from the machine learning model for the loss assessment allows for defining a demand distribution of real responses that would otherwise require additional nonlinear response history analysis, the trade-off is that the loss output depends on model selection/performance, so this should be carefully considered. Also, if the full design and analysis are to be run, it requires substantial computational expense and time. Running the entire design and analysis for every simulated building takes extensive computational power before even reaching the predictive modeling stage. In practical settings, hybrid models would benefit from a suite of previously analyzed structural models. The structural models can be selected such that the feature composition matches the needs of the engineer or researcher for a particular region being analyzed. This would remove the need to perform the design and/or analysis for every simulated building that is attached to the set, and the EDPs can be directly extracted, making it more feasible for rapid post-event impact assessment. Future studies could utilize the full extent of databases, such as that developed in [78]. Finally, the loss assessment is highly dependent on the resolution of the information provided. In this study, for example, unique component population models were created in SP3 based on the LFRS type, number of stories, and occupancy. Every building that is paired with a specific population model has the same component distribution, and only components available in the fragility database provided with pelicun are used. The analysis also did not use the ”blocking” feature in pelicun to consider damage states in series, and assumed components/floors were damaged in parallel.

The prediction model presented in this study offers a novel implementation of using hybrid data for structural response reconstruction and regional impact assessment. It comes with limitations related to assumptions in the information provided, given specifying detailed information for each building may be infeasible on large scales. Despite these limitations, it presents a promising avenue in rapid post-earthquake impact assessment utilizing strong motion data.

CHAPTER 7

Conclusions

7.1 Summary

The studies outlined in this dissertation sought to improve upon and develop new methods for rapid post-earthquake impact assessment suing strong motion data. In chapter 2, a well-organized database of strong motion building response data was developed, creating an easily accessible repository for future research studies. The database addressed limitations found in other seismic data repositories. Some of these databases are limited to a single building type, whereas the one presented has a wide range of structural systems. Another issue is often related to accessibility. Some users may not have the necessary background for extracting the database information, an example of this is the need for targeted querying. A python tool that allows users to extract and visualize data without manually writing queries is provided to help mitigate accessibility issues. The database is also expandable to include more buildings and events, and even potentially data from other hazards.

In chapter 3, a novel dual-model formulation for predicting event-agnostic maximum building responses was created that improves on previous similar models. The dual model used kriging to leverage the spatial relationship of peak ground acceleration (PGA) with the event and site characteristics in a machine learning model. This phase successfully improved the predictive performance of an extreme gradient boosting (XGBoost) algorithm for the engineering demand parameters (EDPs) of interest. The dual model was particularly effective in improving maximum peak floor acceleration (maxPFA) predictions.

Chapter 4 evaluated the effectiveness of several ground motion intensity measures that are candidates for input features into the response prediction models. Sufficiency, efficiency, and a series of causal inference methods were used to evaluate the performance of the various ground intensity measures (IM) using building strong motion response data. This contrasts with prior IM evaluation studies which have all utilized simulated structural responses. The findings showed that cumulative absolute velocity (CAV) generally outperformed the other considered IMs, which is consistent with prior studies based on simulation data. PGA was generally the least effective, especially with respect to estimating the peak story drift ratio (PSDR). Some unexpected findings were related to the spectral IMs under-performing and all IMs being insufficient with respect to source-to-site distance, which was likely due to the small sample size, as sufficiency is documented as being sensitive to this.

In chapter 5, a platform for generating code-conforming designs and performing nonlinear response history analyses was developed, building on three previously developed LFRS-specific platforms. The combination of these programs with the addition of a module for performing loss assessment using FEMA P-58 methodology provides for an end-to-end multi-LFRS framework for design, analysis, and performance assessments on the regional scale. An single example building, designed and analyzed in AutoSDA, was used to illustrate the end-to-end process that can be aggregated over a portfolio for regional impact assessments, as shown in chapter 6. In this chapter, the platform was used to generate a suite of simulated building responses, based on the ground motions from events collected in chapter 2. The instrumentation data from the database was then combined with these simulated responses to create a hybrid data training set for a structural response reconstruction model. This data augmentation, or hybrid dataset, provided for a more generalized prediction model that better characterized the full-profile response of buildings in the test set compared to the instrumentation data alone. This was exhibited in the better overall predictive performance, especially when stratified by the normalized floor height of the building, as well as the reduced uncertainty of these predictions. This technique served as a way to address the limitation of partial building instrumentation which is a common research problem in the literature. It is important to note that the benefits of this model formulation may not only be attributed to the full-profile responses provided by the simulated set, but also by the fact that there is simply more data available in model training. It helps to address two forms of partial instrumentation: (1) within a single building and (2) across a portfolio of buildings where some are entirely uninstrumented, but it also can be seen as a way to mitigate issues of small sample size, in general. The study then proceeds to a loss assessment, showcasing the potential use of this model formulation for rapid-response resource allocation. The uncertainty from the response predictions were propagated to the loss assessment, and the potential benefit of this model formulation was showcased using loss maps for an inventory of buildings subjected to an event. The uncertainties in the loss assessment were generally lower for the estimates generated based on the augmented model predictions compared to the measured-only model predictions and the test set.

7.2 Limitations and Future Work

The studies presented in this dissertation are not without their limitations, which can serve as a source of inspiration for future research.

The database presented in chapter 2 is primarily composed of structural response data that remains in the linear-elastic range, which limits its applicability for studies on nonlinear behavior in structures. The database was developed with the intention of supporting future expansion. Future work can focus on integrating a more robust building response range into the database, as well as expanding it to a multi-hazard data repository.

Chapter 3 developed a predictive model for reconstructing maximum seismic building responses. The study had several limitations. The dataset is limited in size, primarily consisting of events with magnitudes between 4 and 6 and buildings under 20 stories located within 100 km of the earthquake source. Additionally, it inherits the limitation from the database, representing mostly linear-elastic building responses, limiting the model's ability to generalize, particularly for damage detection. The two-phase model structure introduces error and uncertainty during the initial kriging phase used to estimate Peak Ground Acceleration (PGA), although it improves overall performance. However, PGA is not well-correlated with displacement- or deformation-based engineering demand parameters (EDPs), which resulted in insignificant effects on model performance for drift predictions. Given the model's focus on only maximum responses, it also doesn't capture localized damage. A later chapter implements a similar model that addresses this limitation, predicting full-profile peak responses.

Chapter 4 attempts to address one of the limitations presented in the previous chapter, where an IM was well-correlated with one demand parameter, but poorly correlated with another. This chapter focused on assessing the effectiveness of various IMs with building response. Notably, in this study, the structural characteristics of the buildings are imbalanced, with most being low- to mid-rise structures and very few high-rises. This imbalance, combined with the small dataset size, prevented the study from accounting for heterogeneous effects, such as varying performance across different building heights. The limited data size also restricted the effectiveness of some of the causal analysis methods. Future research can overcome these limitations by expanding the dataset to include a broader range of demand levels, potentially through a hybrid set combining recorded and simulated data (similar to chapter 6). This would allow for better balance in structural characteristics and enable more comprehensive causal analyses and moment balancing. Additionally, as the study in this chapter sought to provide insights on how to better implement the IMs in a model such as in chapter 3, future studies should consider the uncertainty in the prediction of these IMs. This would help to answer whether the trade-off of increased average predictive performance is worth the additional quantified uncertainty that is introduced in the first phase of the model. Although some IMs may be determined to be more effective predictors for a particular EDP, they may not be well-correlated spatially, possibly eliminating any presumed advantage of the first phase of the model.

Building upon techniques and limitations of previous chapters, chapter 6 sought to introduce a hybrid data model to alleviate the issue of partial instrumentation. Structural responses obtained from models designed and analysed in an automated seismic design and analysis (AutoSDA) platform were combined with instrumentation data from the developed database. The addition of the simulated structural model responses improved the models ability to reconstruct full-profile responses. This directly addresses a major limitation of chapter 3, but the model selection/performance still introduces uncertainty. Another source of uncertainty lies in the assumptions that were necessary in order to execute the study. It is assumed that the simulated structural models accurately depict real building responses, however, it would be interesting to quantify the validity of this assumption. Future studies could attempt to measure the difference in simulated models and their instrumented building counterparts and include this in the uncertainty quantification portion of the framework. A recent study [52] that implemented a hybrid method did this by using a calibrated shear-flexural beam model and calculating the residuals between measurements and the beam models predictions. It trained a Gaussian Process model on these residuals as a second phase to make a final prediction. This effectively quantified the error associated with simulated models and included it in the predictions for building response, but for a single building. This study offers a potential way to improve the implementation of the hybrid data framework in the dissertation. The loss assessment depends heavily on the resolution of available information. The component populations were generated based on LFRS-type, occupancy, and number of stories in the building. The Python package that implements the loss assessment methodology was limited in the available component fragility parameters. Moreover, individually selecting each component for a building, using engineering judgement or external sources to estimate the parameters was not feasible given the size of the inventory, so not all of the available components were used. Additionally, the median collapse capacity used in the loss assessment was based on the estimates provided by SP3, which is based on a simplified procedure outlined in FEMA P-154. A better alternative would be to obtain the collapse capacity from explicit simulations. Future work could consider developing an extended component fragility database. Additionally, a database such as the one developed in [78], but with a more diverse set of buildings (i.e. other LFRS types, seismic design characteristics, etc) would greatly benefit the hybrid data models. Specifically, it would reduce the computational expense associated with generating the full design and structural responses for the building set. A search algorithm based on structural attributes could be used to select the buildings and responses to be implemented into the model.

APPENDIX A

Database Tables

Table A.1: Building information table

Attribute	**Database Variable**	**Data Type**	**Example**
LightCyan Building ID	ID	INT	23285
Height (ft)	height	INT	804
No. of Stories Above Ground	num_stories_above	INT	5
No. of Stories Below Ground	num_stories_below	INT	1
No. of Channels	num_channel	INT	10
X Channels	X_channel	VARCHAR(255)	3(-1);5(5);9(2)
Y Channels	Y_channel	VARCHAR(300)	1(-1);4(5);6(5);7(5);8(2);10(2)
Z Channel	Z_channel	VARCHAR(255)	2(-1)
Level Height	story_height	VARCHAR(400)	-156;0;180;336;492;648;804
Lateral Force Resisting System	LFRS	VARCHAR(255)	Shear Wall
Instrumentation Year	instrumentation_date	YEAR	1976
Design Year	design_date	YEAR	1968
Base Dimensions	base_dimensions	VARCHAR(255)	261’-7 1/2”x 138’-8”
Green Site ID	site_ID	VARCHAR(255)	SITE_23285

Table A.2: Earthquake information table

Attribute	**Database Variable**	**Data Type**	**Example**
LightCyan Earthquake ID	ID	VARCHAR(255)	Ridgecrest
Earthquake Name	name	VARCHAR(255)	Ridgecrest
Earthquake Date	date	DATE	7/05/2019
Epicenter Longitude	epicenter_longitude	DECIMAL(7,4)	-117.5993
Epicenter Latitude	epicenter_latitude	DECIMAL(6,4)	35.7695
Magnitude	magnitude	DECIMAL(2,1)	7.1
Fault Type	fault_type	VARCHAR(255)	strike-slip
Focal Depth	focal_depth	DECIMAL(3,1)	8
No of Recordings	num_recordings	INT	67

Table A.3: Earthquake-Building junction table

Attribute	**Database Variable**	**Data Type**	**Example**
LightCyan Earthquake-Building Pair ID	ID	VARCHAR(255)	Ridgecrest_23285
Green Building ID	building_ID	INT	23285
Green Earthquake ID	earthquake_ID	VARCHAR(255)	Ridgecrest

Table A.4: Site information table

Attribute	**Database Variable**	**Data Type**	**Example**
LightCyan Site ID	ID	VARCHAR(255)	site_23285
Site Geology	geology	VARCHAR(255)	Deep Alluvium
Site Class	site_class	VARCHAR(255)	D
Vs30	Vs30	VARCHAR(255)	337 (inferred)
Latitude	latitude	DECIMAL(6,4)	34.1822
Longitude	longitude	DECIMAL(7,4)	-117.3242
Elevation (m)	elevation	INT	459
City	city	VARCHAR(255)	San Bernardino

Table A.5: Response history table

Attribute	**Database Variable**	**Data Type**	**Example**
LightCyan Response History ID	ID	VARCHAR(255)	Ridgecrest_23285_Chn3
Sensor Reading Interval (s)	sensor_interval	DECIMAL(5,4)	0.01
Displacement (cm)	displacement	MEDIUMTEXT	-0.0001479,-0.0001462,-0.0001446,-0.0001429...
Velocity (cm/s)	velocity	MEDIUMTEXT	0.0001636,0.0001648,0.0001661,0.0001674...
Acceleration (cm/s/s)	acceleration	MEDIUMTEXT	0.000124,0.000123,0.000126,0.000124,...
Green Earthquake-Building Pair ID	ID	VARCHAR(255)	Ridgecrest_23285

APPENDIX B

MySQL Schema

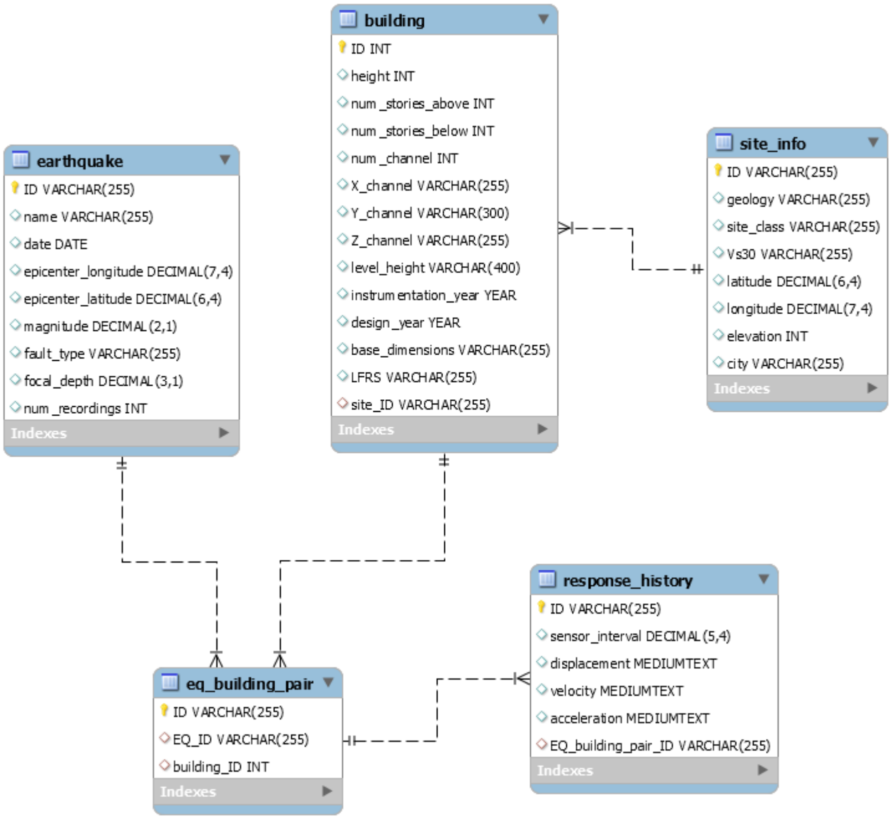


Figure B.1: Schema implementation in MySQL

APPENDIX C

Archetype Loading Conditions

Table C.1: Steel Moment Frame Archetype Loading Conditions

Archetype	Floor	floor weight (lb)	floor dead load (psf)	floor live load (psf)	beam dead load (plf)	beam live load (plf)	leaning column dead load (kips)	leaning column live load (kips)
22	Typical	1800	80	50	800	500	900	562.5
279	Typical	1800	80	50	1200	750	900	562.5
	Roof	450	20	50	300	300	225	562.5
765	Typical	1800	80	50	1200	750	900	562.5
	Roof	450	20	50	300	300	225	562.5
1188	Typical	1125	50	50	500	500	562.5	562.5
	Roof	450	20	50	200	200	225	562.5
1566	Typical	1125	50	50	500	500	562.5	562.5
	Roof	450	20	50	200	200	225	562.5

Table C.2: Wood Frame Archetype Loading Conditions

Archetype	Floor	floor weight (kips)	interterio wall weight (psf)	live load (psi)
148321	Typical	51	11	0.345
196481	Typical	141	11	0.345
248321	Typical	70	11	0.345
	Roof	53	11	0.345
296481	Typical	182	11	0.345
	Roof	144	11	0.345
496481	Typical	237	11	0.345
	Roof	149	11	0.345

Table C.3: Reinforced Concrete Shear Wall Archetype Loading Conditions

Archetype	Floor	floor dead load	floor live load
1	Typical	125	40
2	Typical	125	40
3	Typical	125	40
5	Typical	125	40
7	Typical	125	40
11	Typical	125	40
12	Typical	125	40
16	Typical	125	40

Bibliography

[1] Osama Abdeljaber, Onur Avci, Mustafa Serkan Kiranyaz, Boualem Boashash, Henry Sodano, and Daniel J Inman. 1-d cnns for structural damage detection: Verification on a structural health monitoring benchmark data. *Neurocomputing*, 275:1308–1317, 2018.

[2] Eusef Abdelmalek-Lee and Henry Burton. A dual kriging-xgboost model for reconstructing building seismic responses using strong motion data. *Bull Earthquake Eng*, 2023.

[3] Eusef Abdelmalek-Lee, Tricia Jain, Sebastian Madero, Han Sun, Henry Burton, and John Wallace. Relational database for building strong motion recordings used for seismic impact assessments, 2022.

[4] Eusef Abdelmalek-Lee, Tricia Jain, Sebastian Madero, Han Sun, Henry Burton, and John Wallace. Relational database for building strong motion recordings used for seismic impact assessments. *Earthquake Spectra*, 2023.

[5] Muneera Aladsani and Henry Burton. Rcwall-sda. Reference to a developed software without publication.

[6] Muneera A Aladsani, Henry Burton, Saman A Abdullah, and John W Wallace. Explainable machine learning model for predicting drift capacity of reinforced concrete walls. *ACI Structural Journal*, 119(3), 2022.

[7] Engin Burak Anil, Burcu Akinci, James H Garrett, and Ozgur Kurc. Information requirements for earthquake damage assessment of structural walls. *Advanced Engineering Informatics*, 30(1):54–64, 2016.

[8] ASCE. *Minimum design loads and associated criteria for buildings and other structures.* Number 1. American Society of Civil Engineers, 2016.

[9] ASCE. Minimum design loads and associated criteria for buildings and other structures. *American Institute of Steel Construction, Chicago-Illinois American Society of Civil Engineers*, 2017.

[10] ASCE. Report card for california's infrastructure, 2021.

[11] ATC. *Procedures for Post-earthquake building safety evaluation procedures.* Number 1. Applied Technology Council, 1995.

[12] S. A. Babichev, J. Ries, and A. I. Lvovsky. Quantum scissors: teleportation of single-mode optical states by means of a nonlocal single photon, 2002. Preprint at https://arxiv.org/abs/quant-ph/0208066v1.

[13] M Bagheri Bodaghabadi. Is it necessarily a normally distributed data for kriging? a case study: soil salinity map of ghahab area, central iran. *Desert*, 23(2):284–293, 2018.

[14] M. Beneke, G. Buchalla, and I. Dunietz. Mixing induced CP asymmetries in inclusive B decays. *Phys. Lett.*, B393:132–142, 1997.

[15] Yousef Bozorgnia, Norman A. Abrahamson, Linda Al Atik, Timothy D. Ancheta, Gail M. Atkinson, Jack W. Baker, Annemarie Baltay, David M. Boore, Kenneth W. Campbell, Brian S.-J. Chiou, Robert Darragh, Steve Day, Jennifer Donahue, Robert W. Graves, Nick Gregor, Thomas Hanks, I. M. Idriss, Ronnie Kamai, Tadahiro Kishida, Albert Kottke, Stephen A. Mahin, Sanaz Rezaeian, Badie Rowshandel, Emel Seyhan, Shrey Shahi, Tom Shantz, Walter Silva, Paul Spudich, Jonathan P. Stewart, Jennie Watson-Lamprey, Kathryn Wooddell, and Robert Youngs. Nga-west2 research project, August 2014.

[16] Brendon A Bradley, Rajesh P Dhakal, Gregory A MacRae, and Misko Cubrinovski. Prediction of spatially distributed seismic demands in specific structures: Ground motion and structural response. *Earthquake engineering structural dynamics*, 39(5):501–520, 2010.

[17] Scott J Brandenberg, Paolo Zimmaro, Jonathan P Stewart, Dong Youp Kwak, Kevin W Franke, Robb ES Moss, K Önder Çetin, Gizem Can, Makbule Ilgac, John Stamatakos, et al. Next-generation liquefaction database. *Earthquake Spectra*, 36(2):939–959, 2020.

[18] M. Broy. Software engineering—from auxiliary to key technologies. In M. Broy and E. Denert, editors, *Software Pioneers*, pages 10–13. Springer, New York, 1992.

[19] Henry Burton. Causal inference on observational data: Opportunities and challenges in earthquake engineering. *Earthquake Spectra*, 39(1):54–76, 2022.

[20] Henry Burton. Evaluating the effectiveness of ground motion intensity measures for structural response simulation using statistical and causal inferencing. Master's thesis, UCLA, 2022. Retrieved from https://escholarship.org/uc/item/8qw9s32k.

[21] Henry V Burton and Jack W Baker. Evaluating the effectiveness of ground motion intensity measures through the lens of causal inference. *Earthquake Engineering & Structural Dynamics*, 2023.

[22] California Strong Motion Instrumentation Program. *System Requirements- Central Recording MultiChannel Accelerograph System*, 9 2023.

[23] S. L. Campbell and C. W. Gear. The index of general nonlinear DAES. *Numer. Math.*, 72(2):173–196, 1995.

[24] M Çelebi, A Sanli, M Sinclair, S Gallant, and D Radulescu. Seismic monitoring instrumentation needs of a building owner and the solution: A cooperative effort. In *2003 ASCE/SEI Structures Congress and Exposition: Engineering Smarter*, pages 359–360, 2003.

[25] CESMD. Center for engineering strong motion data, 2023.

[26] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794, 2016.

[27] Tianqi Chen, Tong He, Michael Benesty, Vadim Khotilovich, Yuan Tang, Hyunsu Cho, Kailong Chen, et al. Xgboost: extreme gradient boosting. *R package version 0.4-2*, 1(4):1–4, 2015.

[28] Jianye Ching, James L Beck, Keith A Porter, and Rustem Shaikhutdinov. Bayesian state estimation method for nonlinear systems and its application to recorded seismic response. *Journal of Engineering Mechanics*, 132(4):396–410, 2006.

[29] S. T. Chung and R. L. Morris. Isolation and characterization of plasmid deoxyribonucleic acid from streptomyces fradiae, 1978. Paper presented at the 3rd international symposium on the genetics of industrial microorganisms, University of Wisconsin, Madison, 4–9 June 1978.

[30] Computers and Inc. Structures. Etabs: Integrated analysis, design, and drafting of building systems, 2024.

[31] Computers and Inc. Structures. Sap2000: Integrated software for structural analysis and design, 2024.

[32] Paul P Cordova, Gregory G Deierlein, Sameh SF Mehanny, and C Allin Cornell. Development of a two-parameter seismic intensity measure and probabilistic assessment procedure. In *The second US-Japan workshop on performance-based earthquake engineering methodology for reinforced concrete building structures*, volume 20, page 0, 2000.

[33] International Code Council. International building code. *International Building Code*, 2024.

[34] Gemma Cremen and Jack W. Baker. Quantifying the benefits of building instruments to fema p-58 rapid post-earthquake damage and loss predictions. *Engineering Structures*, 176:243–253, 2018.

[35] CSMIP. California strong motion instrumentation program, 2021.

[36] Laxman Dahal, Henry Burton, and Zhengxiang Yi. An end-to-end computational platform to automate seismic design, nonlinear analysis, and loss assessment... *ResearchGate*, June 2022.

[37] Kaoshan Dai, Dan Lu, Songhan Zhang, Yuanfeng Shi, Jiayao Meng, and Zhenhua Huang. Study on the damping ratios of reinforced concrete structures from seismic response records. *Engineering Structures*, 223:111143, 2020.

[38] Héctor Dávalos and Eduardo Miranda. Filtered incremental velocity: A novel approach in intensity measures for seismic collapse estimation. *Earthquake Engineering & Structural Dynamics*, 48(12):1384–1405, 2019.

[39] Somayajulu LN Dhulipala, Adrian Rodriguez-Marek, Shyam Ranganathan, and Madeleine M Flint. A site-consistent method to quantify sufficiency of alternative ims in relation to psda. *Earthquake Engineering & Structural Dynamics*, 47(2):377–396, 2018.

[40] Yongtao Dong, Ruiqiang Song, Helen Liu, et al. Bridges structural health monitoring and deterioration detection-synthesis of knowledge and technology. Technical report, Alaska. Dept. of Transportation and Public Facilities, 2010.

[41] Ao Du and Jamie E Padgett. Entropy-based intensity measure selection for site-specific probabilistic seismic risk assessment. *Earthquake Engineering & Structural Dynamics*, 50(2):560–579, 2021.

[42] Laura Eads, Eduardo Miranda, and Dimitrios G Lignos. Average spectral acceleration as an intensity measure for collapse risk assessment. *Earthquake Engineering & Structural Dynamics*, 44(12):2057–2073, 2015.

[43] Laura Eads, Eduardo Miranda, and Dimitrios G. Lignos. Average spectral acceleration as an intensity measure for collapse risk assessment. *Earthquake Engineering amp;amp; Structural Dynamics*, 44(12):2057–2073, 2015.

[44] Mahdi Ebrahimian and Maria I Todorovska. Structural system identification of buildings by a wave method based on a layered timoshenko beam model. In *Health Monitoring of Structural and Biological Systems 2014*, volume 9064, pages 385–397. SPIE, 2014.

[45] A Elkady and D G Lignos. Two-dimensional OpenSEES numerical models for archetype steel buildings with special moment frames, 2019.

[46] Kalil Erazo and Eric M Hernandez. Uncertainty quantification of state estimation in nonlinear structural systems with application to seismic response in buildings. *ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering*, 2(3):B5015001, 2016.

[47] Mohsen Zaker Esteghamati, Jeonghyun Lee, Matthew Musetich, and Madeleine M Flint. Inssept: An open-source relational database of seismic performance estimation to aid with early design of buildings. *Earthquake Spectra*, 36(4):2177–2197, 2020.

[48] C. R. Farrar and K. Worden. An introduction to structural health monitoring. *Transactions of the Royal Society A: Mathematical, Physical & Engineering Sciences*, 365(1851):303–315, 2006.

[49] Federal Emergency Management Agency (FEMA). Seismic performance assessment of buildings, 2012.

[50] Christian Fong, Chad Hazlett, and Kohsuke Imai. Covariate balancing propensity score for a continuous treatment: Application to the efficacy of political advertisements. *The Annals of Applied Statistics*, 12:156–177, 03 2018.

[51] K. O. Geddes, S. R. Czapor, and G. Labahn. *Algorithms for Computer Algebra*. Kluwer, Boston, 1992.

[52] Farid Ghahari, Daniel Swensen, Hamid Haddadi, and Ertugrul Taciroglu. A hybrid model-data method for seismic response reconstruction of instrumented buildings. *Earthquake Spectra*, 40(2):1235–1268, 2024.

[53] Xingquan Guan, Henry Burton, and Thomas Sabol. Python-based computational platform to automate seismic design, nonlinear structural model construction and analysis of steel moment resisting frames. *Engineering Structures*, 224:111199, 2020.

[54] Jens Hainmueller. Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis*, 20(1):25–46, 2012.

[55] C. Hamburger. Quasimonotonicity, regularity and duality for nonlinear systems of partial differential equations. *Ann. Mat. Pura. Appl.*, 169(2):321–354, 1995.

[56] Z. Hao, A. AghaKouchak, N. Nakhjiri, and A. Farahmand. Global integrated drought monitoring and prediction system (gidmaps) data sets, 2014. figshare https://doi.org/10.6084/m9.figshare.853801.

[57] LLC Haselton Baker Risk Group.

[58] J. He, X. Guan, and Y. Liu. Structural response reconstruction based on empirical mode decomposition in time domain. *Mechanical Systems & Signal Processing*, 28(4):348–366, 2012.

[59] David V. Hinkley. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):582–584, 1980.

[60] Seong-Hoon Hwang and Dimitrios G. Lignos. Nonmodel-based framework for rapid seismic risk and loss assessment of instrumented steel buildings. *Engineering Structures*, 156:417–432, 2018.

[61] Kosuke Imai and Marc Ratkovic. Covariate Balancing Propensity Score. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 76(1):243–263, 07 2013.

[62] Fatemeh Jalayer, JL Beck, and F Zareian. Analyzing the sufficiency of alternative scalar and vector intensity measures of ground shaking based on information theory. *Journal of Engineering Mechanics*, 138(3):307–316, 2012.

[63] Nirmal Jayaram and Jack W Baker. Correlation model for spatially distributed ground-motion intensities. *Earthquake Engineering & Structural Dynamics*, 38(15):1687–1708, 2009.

[64] William B Joyner and David M Boore. 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*, 71(6):2011–2038, 1981.

[65] AK Kazantzi, D Vamvatsikos, and E Miranda. The effect of damping on floor spectral accelerations as inferred from instrumented buildings. *Bulletin of Earthquake Engineering*, 18(5):2149–2164, 2020.

[66] Hamid Khodabandehlou, Gökhan Pekcan, and M Sami Fadali. Vibration-based structural condition assessment using convolution neural networks. *Structural Control and Health Monitoring*, 26(2):e2308, 2019.

[67] Wenjie Liao, Xingyu Chen, Xinzheng Lu, Yuli Huang, and Yuan Tian. Deep transfer learning and time-frequency characteristics-based identification method for structural seismic response. *Frontiers in Built Environment*, page 10, 2021.

[68] Dimitrios G Lignos and Eduardo Miranda. Estimation of base motion in instrumented steel buildings using output-only system identification. *Earthquake engineering & structural dynamics*, 43(4):547–563, 2014.

[69] MP Limongelli. Seismic health monitoring of an instrumented multistory building using the interpolation method. *Earthquake engineering & structural dynamics*, 43(11):1581–1602, 2014.

[70] Dan Lu, Jiayao Meng, Songhan Zhang, Yuanfeng Shi, Kaoshan Dai, and Zhenhua Huang. Damping ratios of reinforced concrete structures under actual ground motion excitations. In *Dynamics of Civil Structures, Volume 2*, pages 259–268. Springer, 2020.

[71] Xiao Lu, Xinzheng Lu, Hong Guan, and Lieping Ye. Comparison and selection of ground motion intensity measures for seismic design of super high-rise buildings. *Advances in Structural Engineering*, 16:1249–1262, 07 2013.

[72] Nicolas Luco and C. Allin Cornell. Structure-specific scalar intensity measures for near-source and ordinary earthquake ground motions. *Earthquake Spectra*, 23(2):357–392, 2007.

[73] Jerry Lynch and Kenneth Loh. A summary review of wireless sensors and sensor networks for structure health monitoring. *Shock Vib. Dig*, 38:91–128, 2006.

[74] Stephen Mahin, Patxi Uriz, Ian Aiken, Caroline Field, and Eric Ko. Seismic performance of buckling restrained braced frame systems. In *13th World Conference on Earthqauke Engineering*, 2004.

[75] N. M. Maia, R. A. Almeida, A. P. Urgueira, and R. P. Sampaio. Damage detection and quantification using transmissibility. *Mechanical Systems & Signal Processing*, 25(7):2475–2843, 2011.

[76] Silvia Mazzoni, Tadahiro Kishida, Jonathan P Stewart, Victor Contreras, Robert B Darragh, Timothy D Ancheta, Brian SJ Chiou, Walter J Silva, and Yousef Bozorgnia. Relational database used for ground-motion model development in the nga-sub project. *Earthquake Spectra*, 38(2):1529–1548, 2022.

[77] Scott M.H. et al. McKenna F., Fenves G.L. Open system for earthquake engineering simulation, 2000.

[78] Xingquan Guan M.EERI, Henry Burton M.EERI, and Mehrdad Shokrabadi. A database of seismic designs, nonlinear models, and seismic responses for steel moment-resisting frame buildings. *Earthquake Spectra*, 37(2):1199–1222, 2021.

[79] Sameh Samir Fahmy Mehanny. *Modeling and assessment of seismic performance of composite frames with reinforced concrete columns and steel beams.* Stanford University, 2000.

[80] Eduardo Miranda. Use of probability-based measures for automated damage assessment. *The structural design of tall and special buildings*, 15(1):35–50, 2006.

[81] Jack Moehle and Gregory G Deierlein. A framework methodology for performance-based earthquake engineering. In *13th world conference on earthquake engineering*, volume 679. WCEE Vancouver, 2004.

[82] Sifat Muin and Khalid M. Mosalam. Cumulative absolute velocity as a local damage indicator of instrumented structures. *Earthquake Spectra*, 33(2):641–664, 2017.

[83] Sifat Muin and Khalid M Mosalam. Structural health monitoring using machine learning and cumulative absolute velocity features. *Applied Sciences*, 11(12):5727, 2021.

[84] MySQL. Open source database, 2020.

[85] F. S. Naeim, A. Alimoradi, and E Miranda. Automated post-earthquake damage assessment of instrumented buildings. In S. T. Wasti and G. Ozcebe, editors, *Advances in earthquake engineering for urban risk reduction*, pages 117–134. Springer, Berlin, 2006.

[86] Farzad Naeim, Scott Hagie, Arzhang Alimoradi, and Eduardo Miranda. Automated post-earthquake damage assessment of instrumented buildings. In S. Tanvir Wasti and Guney Ozcebe, editors, *Advances in Earthquake Engineering for Urban Risk Reduction*, pages 117–134, Dordrecht, 2006. Springer Netherlands.

[87] Peng Ni, Limin Sun, Jipeng Yang, and Yixian Li. Multi-end physics-informed deep learning for seismic response estimation. *Sensors*, 22(10):3697, 2022.

[88] Morolake Omoya, Itohan Ero, Mohsen Zaker Esteghamati, Henry V Burton, Scott Brandenberg, Han Sun, Zhengxiang Yi, Hua Kang, and Chukuebuka C Nweke. A relational database to support post-earthquake building damage and recovery assessment. *Earthquake Spectra*, 38(2):1549–1569, 2022.

[89] Jamie E. Padgett, Bryant G. Nielson, and Reginald DesRoches. Selection of optimal intensity measures in probabilistic seismic demand models of highway bridge portfolios. *Earthquake engineering structural dynamics*, 37(5):711–725, 2008.

[90] J. Pearl. *Causality.* Cambridge University Press, Cambridge, 2019.

[91] Fabio Pioldi and Egidio Rizzi. Assessment of frequency versus time domain enhanced technique for response-only modal dynamic identification under seismic excitation. *Bulletin of Earthquake Engineering*, 16(3):1547–1570, 2018.

[92] Keith Porter, Judith Mitrani-Reiser, and James L. Beck. Near-real-time loss estimation for instrumented buildings. *The Structural Design of Tall and Special Buildings*, 15(1):3–20, 2006.

[93] Ellen M Rathje, Clint Dawson, Jamie E Padgett, Jean-Paul Pinelli, Dan Stanzione, Ashley Adair, Pedro Arduino, Scott J Brandenberg, Tim Cockerill, Charlie Dey, et al. Designsafe: New cyberinfrastructure for natural hazards engineering. *Natural Hazards Review*, 18(3):06017001–06017001, 2017.

[94] Hugo A. Rojas, Christopher Foley, and Shahram Pezeshk. Risk-based seismic design for optimal structural and nonstructural system performance. *Earthquake Spectra*, 27(3):857–880, 2011.

[95] Milad Roohi, Kalil Erazo, David Rosowsky, and Eric M Hernandez. An extended model-based observer for state estimation in nonlinear hysteretic structural systems. *Mechanical Systems and Signal Processing*, 146:107015, 2021.

[96] Milad Roohi and Eric M Hernandez. Performance-based post-earthquake decision making for instrumented buildings. *Journal of Civil Structural Health Monitoring*, 10:775–792, 2020.

[97] Milad Roohi, Eric M Hernandez, and David Rosowsky. Nonlinear seismic response reconstruction and performance assessment of instrumented wood-frame buildings—validation using neeswood capstone full-scale tests. *Structural Control and Health Monitoring*, 26(9):e2373, 2019.

[98] Milad Roohi, Eric M Hernandez, and David Rosowsky. Nonlinear seismic response reconstruction in minimally instrumented buildingsâ€" validation using neeswood capstone full-scale tests. *Structural Health Monitoring 2019*, 2019.

[99] Olivier Roustant, David Ginsbourger, and Yves Deville. Dicekriging, diceoptim: Two r packages for the analysis of computer experiments by kriging-based metamodeling and optimization. *Journal of statistical software*, 51:1–55, 2012.

[100] D.B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):422–331, 2005.

[101] Donald B. Rubin. Using multivariate matched sampling and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association*, 74(366):318–328, 1979.

[102] R. S. Seymour, editor. *Conductive Polymers*. Plenum, New York, 1981.

[103] Mehrdad Shokrabadi and Henry V Burton. Ground motion intensity measures for rocking building systems. *Earthquake spectra*, 33(4):1533–1554, 2017.

[104] Nilesh Shome. *Probabilistic seismic demand analysis of nonlinear structures*. Stanford University, 1999.

[105] DA Skolnik, WJ Kaiser, and JW Wallace. Instrumentation for structural health monitoring: measuring interstory drift. In *Proceedings of the 14th world conference on earthquake engineering*, pages 12–17, 2008.

[106] M. K. Slifka and J. L. Whitton. Clinical implications of dysregulated cytokine production. *J. Mol. Med.*, 78:74–80, 2000.

[107] S. E. Smith. Neuromuscular blocking drugs in man. In E. Zaimis, editor, *Neuromuscular junction. Handbook of experimental pharmacology*, volume 42, pages 593–660, Heidelberg, 1976. Springer.

[108] Jerzy Splawa-Neyman, D. M. Dabrowska, and T. P. Speed. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1923.

[109] B. Stahl. deepSIP: deep learning of Supernova Ia Parameters. Astrophysics Source Code Library, Jun 2020.

[110] Han Sun, Henry Burton, Jonathan Stewart, and John Wallace. Development of a generalized cross-building structural response reconstruction model using strong motion data. *Journal of Structural Engineering*, 148, 03 2022.

[111] Han Sun, Henry Burton, and John Wallace. Reconstructing seismic response demands across multiple tall buildings using kernel-based machine learning methods. *Structural Control and Health Monitoring*, 26(7):e2359, 2019.

[112] Han Sun, Henry Burton, and John Wallace. Reconstructing seismic response demands across multiple tall buildings using kernel-based machine learning methods. *Structural Control and Health Monitoring*, 26(7):e2359, 2019. e2359 STC-18-0189.R2.

[113] Han Sun, Henry Burton, Yu Zhang, and John Wallace. Interbuilding interpolation of peak seismic response using spatially correlated demand parameters. *Earthquake Engineering & Structural Dynamics*, 47(5):1148–1168, 2018.

[114] Han Sun, Henry V Burton, Jonathan P Stewart, and John W Wallace. Development of a generalized cross-building structural response reconstruction model using strong motion data. *Journal of Structural Engineering*, 148(6):04022053, 2022.

[115] Bentley Systems. Ram structural system: Comprehensive building analysis and design software, 2024.

[116] Maria I Todorovska and Mihailo D Trifunac. Earthquake damage detection in the imperial county services building i: The data and time–frequency analysis. *Soil Dynamics and Earthquake Engineering*, 27(6):564–576, 2007.

[117] Polsak Tothong and Nicolas Luco. Probabilistic seismic demand analysis using advanced ground motion intensity measures. *Earthquake Engineering & Structural Dynamics*, 36(13):1837–1860, 2007.

[118] Stefan Tübbicke. Entropy balancing for continuous treatments. *Journal of Econometric Methods*, 11(1):71–89, 2022.

[119] HR Tavakoli, A Rashidi, and S Akbarpoor. Effect of lateral force resisting system on seismic performance of steel moment frames under progressive collapse. *Sharif Journal of Civil Engineering*, 31(4.2):101–108, 2016.

[120] USGS. United states geological survey, 2021.

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

[122] Yu Xie, Jennie E. Brand, and Ben Jann. Estimating heterogeneous treatment effects with observational data. *Sociological Methodology*, 42(1):314–347, 2012. PMID: 23482633.

[123] S. Xu, J. Dimasaka, Wald, D.J., and et al. Seismic multi-hazard and impact estimation via causal inference from satellite imagery. *Nature Communications*, 13(7793), 2022.

[124] Dong Zhang, Yatong Zhou, Hanming Chen, Wei Chen, Shaohuan Zu, and Yangkang Chen. Hybrid rank-sparsity constraint model for simultaneous reconstruction and denoising of 3D seismic data. *Geophysics*, 82(5):V351–V367, September 2017.

[125] Ruiyang Zhang, Zhao Chen, Su Chen, Jingwei Zheng, Oral Büyüköztürk, and Hao Sun. Deep long short-term memory networks for nonlinear structural seismic response prediction. *Computers & Structures*, 220:55–68, 2019.

[126] Ruiyang Zhang, Yang Liu, and Hao Sun. Physics-guided convolutional neural network (phycnn) for data-driven seismic response modeling. *Engineering Structures*, 215:110704, 2020.

[127] X. Zhao, Y.L. Xu, J. Chen, and J. Li. Hybrid identification method for multi-story buildings with unknown ground motion: Experimental investigation. *Engineering Structures*, 27(8):1234–1247, 2005.

[128] Yi Zou, LPSG Tong, and Grant P Steven. Vibration-based model-dependent damage (delamination) identification and health monitoring for composite structures—a review. *Journal of Sound and vibration*, 230(2):357–378, 2000.

[129] Adam Zsarnoczay, Pouria Kourehpaz, and kuanshi. NHERI-SimCenter/pelicun: v3.0, 2022.