A DATA-DRIVEN COMPUTING FRAMEWORK FOR STRUCTURAL SEISMIC RESPONSE PREDICTION

A Dissertation

by

HUAN LUO

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Stephanie Paal
Committee Members,	Joseph Bracci
	Theodora Chaspari
	Stefan Hurlebaus
Head of Department,	Robin Autenrieth

December 2020

Major Subject: Civil Engineering

ABSTRACT

Accurate and rapid seismic response prediction of reinforced concrete (RC) structures in earthquake-prone regions is an important topic in structural and earthquake engineering. However, existing physics-based modeling approaches do not have a good compromise between predictive performance and computational efficiency. High-fidelity models have reasonable predictive performance but are computationally demanding, while more simplified models may be computationally efficient, but do not have as good of performance. The research presented herein aims to address this challenge by developing a novel data-driven computational paradigm via the coupling of machine learning (ML) methods and physics-based models. The ML methods can directly link the experimental data to nonlinear properties of target component, while the physical models meeting universal laws (e.g., Newton's law of motion) can be used to perform the seismic analysis. Additionally, in real-world scenarios, the dataset is most likely corrupted by outliers, contains missing values, and has sample bias due to the potentially small size. The performance of existing ML methods will be negatively affected by these data-related problems. Thus, novel computational methods to deal with these data-related problems are also developed to make the proposed data-driven framework robust under these circumstances. In sum, the contributions of this dissertation are the following:

1) Two RC column databases, one for rectangular and another for circular columns, were developed.
2) A new ML-based backbone curve model (ML-BCV) was developed by integrating a multi-output least squares support vector machine for regression (MLS-SVMR) with a

grid search algorithm for rapid prediction of the bi-linear cyclic backbone curve of RC columns.

3) A novel, locally-weighted ML model (LWLS-SVMR) was developed by combining LS-SVMR and a locally weighted learning algorithm for generalized drift capacity prediction of RC columns.
4) A new, component-level, data-driven framework was developed for generalized, accurate, and efficient seismic response history prediction of structural components subjected to both displacement-controlled cyclic loading and dynamic ground motions. The framework was illustrated for RC columns.
5) The component-level data-driven framework was extended to the system level by coupling it with the simplified, physics-based shear building model. The proposed system-level framework was illustrated for RC frames.
6) A novel, robust, locally-weighted ML model (RLWLS-SVMR) was developed by introducing a weight function into the reformulation of LWLS-SVMR to eliminate the negative effect induced by outliers.
7) A new multiple imputation (MI) method (SRB-PMM) was developed by using sequential regression and predictive mean matching to generate several candidates for imputing (filling in) each missing value while considering the uncertainty associated with the missing data.
8) A novel, regression-based, transfer learning model (DW-SVTR) was developed by coupling two weight functions with LS-SVMR to reduce the negative effect of sample bias due to small datasets.

ACKNOWLEDGEMENTS

First thanks go to my advisor, Dr. Stephanie Paal, for her continuous encouragement, provided guidance, showed patience, shared experience, technical and financial support during my Ph.D. studies. I would also like to thank my committee members, Dr. Joseph Bracci, Dr. Theodora Chaspari, and Dr. Stefan Hurlebaus, for their valuable guidance and support throughout the course of this research.

Thanks also go to my friends and colleagues, the department faculty and staff, and the other undergraduate and graduate students which I have had the opportunity to meet and collaborate with, for their kind help and making my time at Texas A&M University a great experience.

Finally, special thanks to my parents and sisters, who always encourage and support me.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professors Stephanie Paal (Chair), Joseph Bracci, Stefan Hurlebaus of the Zachry Department of Civil and Environmental Engineering and Professor Theodora Chaspari of the Department of Computer Science and Engineering.

All the work conducted for the dissertation was completed by the student, under the advisement of Professor Stephanie Paal of the Zachry Department of Civil and Environmental Engineering.

Funding Sources

Graduate study was supported by Texas A&M University faculty research initial grant.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

1.1 Motivation and Background

Reinforced concrete (RC) structures in earthquake-prone regions are exposed to high seismic collapse risk. Accurate and rapid structural response prediction for these RC structures under earthquake loads is an important step to quantify global collapse risk (Moehle and Deierlein 2004). Existing approaches to predict the seismic response of an RC structure include physical experiments and numerical modeling. Physical experiments are regarded as the most reliable approach and are typically performed by displacement-controlled quasi-static cyclic loading or dynamic shake table tests (Bracci et al. 1992; 1995; Moehle 2014). However, due to the extensive costs associated with large-scale structural testing, experimental validation is not always feasible. As an alternative, numerical modeling techniques (i.e., physics-based approaches such as the finite element method (FEM)) are often employed to predict the seismic response of an RC building by performing nonlinear time-history analyses (Chopra 2007; Deierlein et al. 2010; Moehle 2014). Nevertheless, existing physics-based approaches do not generally have a good compromise between predictive performance and computational efficiency. High-fidelity models (e.g., micro-scale FEM) have reasonable predictive performance but are computationally demanding, while more simplified models (e.g., shear building model) may be computationally efficient, but do not have as good of performance (Spacone et al. 2008; Taucer et al. 1991). Further, physics-based approaches do not have good generalization performance, where a given computational approach is suitable to one type of structure (e.g., distributed plasticity fiber model for a ductile structure) but may not be appropriate for another type of structure (e.g., non-ductile structure) (Deierlein et

al. 2010; Marini and Spacone 2006). These shortcomings are made even more evident when quantifying regional seismic risk. Urban areas, especially metropolitan areas, generally have a dense building population, which make the formulation of high-fidelity or even simplified models for all buildings in the region impractical due to the poor compromise between predictive performance and computational efficiency. Therefore, there is a strong need to provide a rapid and accurate approach to structural response prediction which can be extrapolated for regional assessments. Further, a novel computational methodology for generalized, efficient, and accurate seismic response prediction of RC structures is required.

Typically, high-fidelity models involve less theoretical assumptions at the expense of computational efficiency, while simplified models are the opposite. Both high-fidelity and simplified models are based on the use of two types of equations. The first one, of axiomatic character, is related to universal laws that are recognized as epistemic (e.g., force equilibrium based on Newton's law of motion and geometric compatibility based on kinematics), while the second one is composed of empirical models that researchers have derived from collected experimental data (e.g., constitutive equations for materials and structural components). The first one does not suffer any empiricism or uncertainty but the second does. In general, the material constitutive models are used by high-fidelity models for microscale computation, while the component constitutive equations are employed by simplified models for macroscale computation. The computational cost for the former cannot be further reduced in terms of the mathematical formulations. But the prediction performance for the latter can be significantly improved when the parameters that reflect the nonlinear properties are accurately defined for target RC structures. However, the existing component constitutive equations used to define these parameters are based primarily on empirical relations, which cannot fully capture the underlying patterns in the

experimental data. In such a way, though simplified models are computational efficient they may not be able to capture the experimentally observed behavior, leading to incompatibility between the predicted results and observed values. With the recent push towards real-world, big data, many disciplines, both in engineering and science, have been driven towards advancements in data science. In data science-type approaches, the physical behavior is derived directly from real-world big data (Solomatine et al. 2008). Empirical relations that are used to define the parameters in the physics-based approaches will be less informative than those directly reflected in the data. In data science, knowledge is extracted from the data (also called the training data) by using advanced artificial intelligence (AI) techniques or statistical learning approaches (e.g., machine learning (ML)) without any human assumptions or inference. This knowledge is typically expressed in a specific mathematic model which is then employed to directly relate the input predictors to the output responses with high generalization performance and computational efficiency without worrying too much about the underlying physical processes. In the past several decades, many researchers have published relevant physical experiments (e.g., displacement-controlled, quasi-static cyclic loading and shake table tests) for both RC structural components (e.g., beam, column, and wall) and systems (e.g., frame, frame-wall). The source of the real-world physical experimental data provides valuable information to develop data-driven computing approaches for structural seismic response prediction. With the help of this physical experimental data and ML techniques, the observed physical behavior for new RC structural components and systems of interest may be efficiently reproduced. In this direction, the errors produced by empirical relations can be minimized.

1.2 Research Objectives

This research is focused on developing a novel data-driven computational paradigm to predict the seismic response of RC structures in a more generalized, robust, scientific, and efficient way, actuating next-generation modeling approaches along the way. The proposed approach can directly link the experimental data to nonlinear properties of target RC structures, while still employing universal laws (i.e., equilibrium and compatibility can be enforced). Therefore, the proposed approach can potentially minimize the modeling errors characteristic of empirical models (i.e., component constitutive equations). The main objectives of this research are:

(1) To collect a large number of experimental test specimens for RC structural components under reversed cyclic loading. For each specimen, the information collected will include the structural features such as geometry, material properties, and design details and the experimental force-displacement data. Based on the collected information, a database will be developed for the use in this research.

(2) To develop new machine learning (ML) models for hardening and softening behavior prediction of the target structural components subjected to cyclic loading reversals based on the collected database.

(3) To formulate a novel hysteretic modeler for the target structural components based on the collected database and the developed ML model.

(4) To couple the developed hysteretic modeler with simplified physics-based modeling approaches at the component and system levels, forming novel data-driven frameworks for seismic response prediction of the target structural component and system. The hysteretic modeler can directly link the experimental data to the nonlinear properties of target structural components.

Simplified physics-based approaches that meet universal laws can be employed for the seismic analysis at the component and system levels in a computationally efficient way.

(5) To develop new computational methods addressing the data-related problems, such as a dataset that is small in size and contains outliers or missing data in order that the data-driven frameworks are robust under such circumstances.

1.3 Research Scope

In order to validate the proposed methodology, the aim of the research proposed here has been limited as described below. These constraints have been established based on the availability of physical experiment data and to maintain consistency with the research objectives.

- Structural Systems: reinforced concrete (RC) frame buildings
- Structural Components: RC columns
- External Loads: quasi-static cyclic loading and ground motions
- Seismic Response: force-displacement and time-response quantity relations
- Data-Related Problems: outliers, small dataset, and missing data

Although the selection of appropriate ground motions is important in order to quantify the seismic collapse risk of RC structures located at a specific site, this dissertation does not deal with such problems and only focuses on the development and validation of a novel data-driven computational paradigm for seismic response prediction of RC structures. The efforts in this dissertation are fully focused on the development of novel approaches regarding the ML, data-driven frameworks at the component and system levels, as well as solutions to data problems.

1.4 Outline of the Dissertation

This dissertation is organized into seven chapters with the following contents:

Chapter II provides a literature review of the current state of knowledge in structural seismic response prediction for RC columns and frames using both traditional physics-based modeling and machine learning (ML) approaches, along with pertinent research efforts in this domain. This chapter concludes with the definition of the existing gaps in these knowledge areas which will be addressed in this research work.

Chapter III presents the development of the column datasets and the validation and assessment methods of the ML models. All the columns in the datasets were tested under reversed cyclic loading. For each column specimen, the information collected includes the structural features such as geometry, material properties, and design details and the experimental force-displacement data. A modified three-parameter hysteretic model and a hybrid optimization algorithm are proposed to extract the backbone curve and hysteretic parameters from the experimental force-displacement data. By pairing the collected column features and the extracted parameters, the RC column datasets thus can be formed. Then, the commonly used validation methods and assessment metrics for quantifying the performance of the ML models are introduced.

Chapter IV describes the formulation of the novel component-level, data-driven framework. First, new ML models are developed to predict the hardening and softening behavior of RC columns subjected to cyclic loading reversals based on the column datasets presented in Chapter III. Then, a novel hybrid-ML-physics-based data-driven framework is proposed for generalized seismic response prediction of RC columns subjected to both quasi-static cyclic loadings and ground motions. An ML model is used to directly link the experimental data to the nonlinear properties of RC columns and a physics-based model that meets universal laws is used

to perform the seismic analysis. Two data-driven seismic response solvers are developed to implement the proposed framework. Numerical experiments are designed to validate the performance of the proposed method and results are also discussed.

Chapter V presents the extension of the component-level data-driven framework to seismic response prediction of structural systems with emphasis on RC frames. The extension is achieved by coupling the proposed component-level framework with the multi-story shear building model. One major advantage of the novel approach is that the hysteretic property of each column in each story is determined by the proposed component-level framework based on the column dataset developed in Chapter III, while the shear building model can efficiently perform the seismic analysis. Therefore, this methodology can achieve a good compromise between prediction performance and computational efficiency. Two system-level data-driven seismic response solvers are developed to implement the proposed framework. The performance of the proposed method is assessed and validated by comparing the numerical results with experimental data and results obtained by widely-used distributed plasticity fiber approaches.

Chapter VI details the novel computational methods for addressing data-related problems which have been developed in this work. These problems include cases where the dataset is corrupted by outliers, the size of the dataset is small (and thus, large sample bias), and where the dataset contains missing values. The majority of existing ML methods can be negatively affected by outliers, resulting in misleading predictions. A small dataset that has large sample bias can lead to a fully trained ML model that has large bias. Almost all existing ML methods fail to deal with such a problem. Addressing missing data is one of the most important problems in ML since inappropriate treatment may result in loss of important information, which in turn, decreases the generalization performance of the ML model trained on the incomplete dataset. The investigation

of the effect of these data-related problems on the performance of proposed ML methods is presented in this chapter. Moreover, new computational approaches to solve these data problems are developed and presented. These computational approaches are effective at generalizing the proposed data-driven frameworks such that they are robust under such circumstances. The numerical validation of the proposed methods is presented and the results are also discussed.

Chapter VII summarizes the findings and contributions of this research. It also describes the limitations of the present study and provides suggestions for future research work.

CHAPTER II

LITERATURE REVIEW

2.1 Overview

This chapter explores various existing methods for predicting the structural response of reinforced concrete (RC) columns and frames under seismic loads. It begins with a review of traditional physics-based methods, focusing primarily on the strength, deformation capacity, and entire response history (e.g., hysteretic curve and relations of time-response quantities). Existing artificial intelligence (AI)-based techniques (i.e., machine learning (ML) methods) for predicting the structural seismic are subsequently presented. Further, the effect of data-related problems on the performance of ML methods is also introduced (e.g., a dataset that is corrupted by outliers, has large sample bias due to little data availability, or a dataset that contains missing values). The advantages and disadvantages of both physics- and ML-based methods as well as the significant performance deterioration of ML methods due to the data-related problems are discussed.

2.2 Existing Physics-Based Methods in Seismic Response Prediction

Traditional physics-based approaches for predicting the seismic response can be divided into models for structural components and those for the entire system. At the component level, specifically for RC columns, many researchers have proposed various simplified formulas (either empirical or semi-empirical) to estimate the seismic capacity, which are quantified by the strength and deformation capacity. At the system level, both finite element methods (FEM) and simplified models are employed to predict the seismic response history. The peak seismic response quantities of interest (e.g., peak inter-story drift) can be extracted from the seismic analysis results. Relevant representatives of these methods are introduced in this section.

2.2.1 Lateral strength estimation

Existing approaches to predict the lateral strength of RC columns can be separated into two categories according to the type of RC column: one for flexural strength prediction of flexure-critical columns and another for shear strength prediction of shear- and flexure-shear-critical columns. The most commonly used method for rapid flexural strength estimation is the rectangular stress block method (Whitney 1937). This approach requires the estimation of two coefficients ($\alpha_1$and β_1) to constitute the sectional rectangular stress block. Two common ways of calculating these coefficients are via the ACI code (ACI 318-14-22) and the approach proposed by Ozbakkaloglu and Saatcioglu (2004). Another more accurate method is the fiber model, in which the various material constitutive relations (i.e., cover and confined concrete and longitudinal reinforcement) are pre-defined at the column section level. Different from the rectangular stress block method, the stress distribution along the section in the fiber model is not necessarily rectangular, which more reasonably reflects the actual flexural behavior. However, the premise of both the rectangular stress block method and fiber model is that the actual deflected section of the

column must meet the plane section assumption. Nevertheless, the actual deflected section is not necessarily plane, and the errors produced by these two approaches will be amplified when the actual deflected section is far away from the plane.

In contrast to the approaches for flexural strength estimation, there is still no unified method for predicting the shear strength of RC shear- and flexure-shear-critical columns. Several research efforts have focused on the development of semi-empirical shear strength models. Priestley et al. (1994) proposed a shear strength model consisting of three independent components: (1) the concrete component for which the magnitude depends on the concrete compressive strength, displacement ductility, effective concrete area (which is a function of gross section area), and longitudinal reinforcement content; (2) the axial load component for which the magnitude depends on the column aspect ratio and the applied axial load; and, (3) the truss component for which the magnitude depends on the transverse reinforcement content and stirrup spacing to effective depth ratio. This model has been verified and can provide significantly improved correlation with experimental results when compared to the earlier models developed by Ghee et al. (1989), ASCE-ACI 426 (1973), and Watanabe and Ichinose (1991) in predicting the shear strength of RC shear- and flexure-shear-critical columns. Sezen and Moehle (2004) developed a model via a statistical regression analysis of 51 test columns from previous experiments reported in the literature. This model incorporates the axial load component in the concrete component and introduces a ductility-related factor for both concrete and transverse reinforcement. The numerical results were validated by comparison of the predicted results with experimentally observed data and values calculated according to ACI 318 (2002), FEMA 273 (1997), and the model proposed by Priestley et al. (1994). However, both the Priestley et al. (1994)

and Sezen and Moehle (2004) methods require an accurate definition of the displacement ductility factor, where the actual value (not design value) is unknown before an experimental test.

2.2.2 Deformation capacity estimation

For the estimation of drift capacity, Pujol et al. (1999) and Elwood and Moehle (2005) are the most popularly used empirical models. Pujol et al. (1999) proposed a conservative model developed based on a database including 92 columns. This model identified that the predictors most affecting the drift capacity are column aspect ratio, concrete compressive strength, longitudinal and transverse reinforcement content, and maximum normalized shear stress (which is a function of the maximum shear force). Elwood and Moehle (2005) also developed a drift capacity model that has been validated and demonstrated as superior to that developed by Pujol et al. (1999). In addition to some of the predictors used in Pujol et al. (1999), the axial load ratio is considered in this model. The drift capacity model developed by Elwood and Moehle (2005) can provide a better estimation of the drift capacity at shear failure. However, both of these models include the maximum shear force of the column as a predictor variable, and the actual value of maximum shear force is an unknown parameter prior to testing of the column.

2.2.3 Response history prediction

For seismic response history prediction of RC columns and frames, one of the commonly used methods is nonlinear time-history analysis. A traditional time-history analysis for an RC column or frame relates the established, detailed, as-designed, structural model—which is based on available design parameters such as material properties (e.g., concrete and reinforcement strength), structural member dimensions (e.g., width, depth, and length), and boundary conditions (e.g., fixed at the column base)—to the input ground motions. Two frequently employed ways to build the numerical model are lumped and distributed plasticity approaches. A lumped plasticity model is

established by means of zero-length nonlinear springs located at the structural member ends, representing plastic hinges (Gilberson 1967; Filippou and Issa 1988). Hysteretic force-displacement relations representing the members' nonlinear cyclic responses are assigned to these plastic hinges. This method simplifies the modeling procedure, reduces the computational cost, and improves the numerical stability of the computations. A distributed plasticity fiber model more accurately describes the inelastic behavior of RC members, as the material nonlinearity can be accounted for at any element section (Spacone et al. 1996a; 1996b). At the section level, material constitutive relationship models such as those for concrete and reinforcement can be defined via discretizing the section into a series of fibers representing cover, core, or confined concrete, and reinforcement. Then, the well-defined section is assigned to the element (i.e., force- or displacement-based beam-column elements) by designating a number of integration points to simulate the nonlinear behavior of structural members along their lengths. The structural model is developed by assembling these elements to formulate the structural system. The fiber elements can capture cracking, onset of yielding, and the spread of plasticity throughout the cross-section as well as along the element length (Haselton et al. 2009; Deierlein et al. 2010).

However, neither approach has good generalization performance. Lumped plasticity models require prior assumptions for the determination of spring parameters. The selected parameters must be capable of representing the experimental hysteretic behavior of target RC members (Taucer et al. 1991). Distributed plasticity models fail to fully represent the strength and stiffness degradation as well as the pinching effects of the hysteretic loops (Haselton et al. 2009; Deierlein et al. 2010). Additionally, high computational costs are required to conduct these analyses, especially for the distributed plasticity approach, where for each load step or time instant, the section, element, and structure stiffness matrix are updated iteratively. Many studies have been

conducted to construct reduced nonlinear models that can be used more conveniently in practice (i.e., low computational cost, but reasonable estimation). These simplified models ignore some degrees of freedom (DOFs) that do not significantly influence the results. For the multi-story frames, the most commonly used simplified approach is the shear building model, where the model only considers the lateral DOFs at each floor due to the assumption that the beam stiffness is infinite in axial and flexure. Therefore, the shear building model still maintains the MDOF properties, but can significantly reduce the computational cost. The prediction performance of the shear building model heavily relies on the definition of lateral force-displacement properties for each story.

Decanini et al. (2004) proposed a simplified procedure for evaluating seismic demand and performance of RC frames subjected to severe ground motions based on an equivalent shear-building model. In this approach, the yield strength to define the lateral force-deformation relation for each story is determined by analysis with an inverted-triangular static force pattern. The accuracy of this simplified method is validated by comparing the results of 6- and 12-story full-frame models of nonlinear time-history analyses. Hajirasouliha and Doostan (2010) developed a simplified analytical model for seismic response prediction of concentrically braced frames. An as-designed multi-story frame model is first reduced to an equivalent shear-building model. Then, the shear-building model is improved by introducing supplementary springs to account for the displacements induced by flexural deformation in addition to shear displacements. The nonlinear force-deformation relation for each story of the modified shear-building model is determined by performing a static pushover analysis considering P-Delta effects on the full-frame models. The adequacy of the proposed simplified model has been examined via nonlinear time-history analyses on full-frame models of 5-, 10-, and 15-story concentrically braced frames. The results show that

the proposed simplified model not only significantly reduces the computational time, but is also accurate enough for practical applications in seismic design and performance assessment. However, the lateral force-displacement properties of the shear building model were routinely determined according to static pushover analyses, which ignore the cyclic strength deterioration and depend on the nonlinear properties of the full-frame model. Further, the determination of the nonlinear properties of the full-frame model is based primarily on empirical or semi-empirical models (e.g., material or structural component constitutive models), which may not be able to capture the underlying patterns in the experimental data of the materials (e.g., concrete, steel) and structural components (e.g., RC columns).

Another family of simplified methods uses a static pushover analysis in place of the dynamic time-history analysis to efficiently estimate the peak seismic response quantities (e.g., peak inter-story ratio). In a traditional pushover analysis, a monotonically increasing lateral load pattern, with an invariant height-wise distribution, is applied to the RC frame until a target displacement is reached. Both the force distribution and target displacement are based on the assumption that the response is controlled by the fundamental mode and that the mode shape remains unchanged after the structure yields (Chopra and Goel 2002). Therefore, the conventional pushover analysis cannot consider the higher mode effect (Chopra and Goel 2002; Krawinkler and Seneviratna 1998) and also do not account for changes in lateral load patterns (Gupta and Kunnath 2000; Amini and Poursha 2018). To take the higher mode effect into consideration, modal pushover analysis (Chopra and Goel 2002) and its variants (Chopra and Goel 2004; Poursha et al. 2009) have been developed. To consider the inelastic effects that influence the height-wise distribution of inertia forces, the adaptive force distribution procedure was first proposed by Bracci et al. (1997), where the lateral load patterns are progressively updated to account for changes in

the dynamic properties of the RC frame in the inelastic region. Later, many researchers proposed variants to solve such problems (Gupta and Kunnath 2000; Antoniou, S., & Pinho, R. 2004; Amini and Poursha 2018). However, all of these methods can only estimate the peak seismic response and cannot predict the entire response history. Further, they also cannot consider the effect of cyclic strength deterioration, which is important for seismic collapse risk assessment (Deierlein et al. 2010).

2.3 Machine Learning-Based Techniques in Structural Engineering

In contrast to the aforementioned physics-based approaches, machine learning (ML)-based techniques give a completely different perspective. In ML, data is the most important component. Given a data set which is high-quality and sufficiently large in size, ML methods are typically used to fit the data set without involving the laws of physics, forming a mathematical model that can closely capture any underlying patterns. With the recent rapid development of ML approaches in the engineering and science domains, many researchers in structural and earthquake engineering have also employed advanced ML approaches to address problems that have not been solved adequately via traditional physics-based approaches (Xie et al. 2020). However, few efforts have focused on RC columns and buildings. In this section, the most relevant literature regarding the strength prediction of structural components and the response history prediction of frames using ML methods are reviewed. This literature review is focused on those approaches that employ experimental data for model development; those that use simulated data are not considered in this dissertation. For a more detailed review of application of ML in structural and earthquake engineering, refer to a recently published survey paper (Xie et al. 2020).

2.3.1 ML methods in strength prediction

Jeng and Mo (2004) carried out research regarding the quick seismic response estimation of a pre-stressed concrete bridge under earthquake excitation of various magnitudes along various directions using artificial neural networks (ANNs). Although ANNs are capable of capturing the nonlinear mapping between independent and dependent variables and can obtain desirable results (Cheng and Cao, 2015; Guler, 2014), their implementation is subject to several drawbacks. One of the major disadvantages is that the ANN training process is reached via a gradient descent

algorithm on the error space which can be more complex and may contain many local minima values. Moreover, trial and error processes are required to establish the optimal network structure.

Jeon et al. (2014) proposed a novel set of probabilistic joint shear strength models using a multiple linear regression method and advanced ML methods including multivariate adaptive regression splines (MARS) and symbolic regression (SR). Experimental databases comprising reinforced and unreinforced concrete beam-column joint tests were established to obtain high-fidelity regression models with reduced model error and bias. The comparison among simulated results by these approaches indicated that the MARS method is the best estimation method. Meanwhile, the predicted accuracy using MARS compared to existing joint shear strength relationships showed more accurate agreement. Alipour et al. (2017) adopted decision trees and random forests (RF) (Breiman 2001) to evaluate the load-capacity rating of bridge populations. Over 40,000 concrete slab bridge data sets were used. The analytical results were compared with a number of existing judgment-based strategies, showing that the proposed method can aid in determining which posted bridges should be further examined for both possible load restriction and restriction removal.

Pal and Deswal (2011) adopted support vector machines for regression (SVMR) (Vapnik 1995) to predict the shear strength of reinforced and prestressed concrete deep beams. The results predicted by the SVMR were compared with those obtained from ANNs and three empirical equations (for the reinforced concrete deep beams), and one empirical equation (for the prestressed concrete deep beams). The results illustrate an improved performance in terms of prediction capabilities by the SVMR when compared to the ANNs and empirical equations. Chou et al. (2014) also proposed an ML model to predict the shear strength of RC deep beams. The proposed ML model consists of a smart artificial firefly colony algorithm (SFA) and least squares-SVMR (LS-

SVMR) (Suykens et al. 2002). The model performance was validated by comparing results with those obtained by SVMR and formula-based approaches, showing that the proposed model is superior to others in predicting the shear strength of RC deep beams. Vu and Hoang (2016) established a hybrid ML model to predict the ultimate punching shear capacity of FRP-reinforced slabs. LS-SVMR and the firefly algorithm (FA) were adopted to discover the mapping between the influencing factors and the slab punching capacity. The predicted results from the proposed model had better agreement when compared with experimental data than those calculated by formula-based and ANN-based approaches.

2.3.2 ML methods in response history prediction

Many ML approaches have also been proposed to model the hysteretic behavior of structural components (Farrokh and Joghataie 2013; Farrokh et al. 2015; Yun et al. 2008). The hysteretic behavior is represented by the nonlinear force-displacement relationship. These approaches used the experimental force-displacement data of a specific structural component to compose the training set, where the predictor is the displacement and the response is the force. All of the existing approaches are based on variants of ANNs, which are used to learn the nonlinear relationship exhibited by the experimental training set and then model the hysteretic behavior for the same structural component. However, these strategies do not relate the structural features to the hysteretic behavior and cannot capture the behavior variation when some structural features change. Moreover, the hysteretic behavior of RC structural components changes significantly when some structural features change (e.g., reducing the reinforcement ratio and concrete compressive strength or changing the geometry of the RC structural component). Therefore, these strategies are not appropriate for the scope of this dissertation.

Similar strategies were proposed to model the response history of structural systems subjected to earthquake excitations. Zhang et al. (2019) proposed a deep learning-based approach to model the nonlinear seismic response of a structural system. In this method, a training set is developed where the predictors are the ground motion-related information (e.g., ground acceleration, velocity, and displacement), and the response variables are the structural response-related information (e.g., story acceleration, velocity, and displacement). This training set is used to train a deep learning model. Then, the well-trained deep learning model can be used to predict the structural seismic response given a new ground motion. However, this method requires prior knowledge about the structural response under multiple ground motions, where both structural response and ground motions will be grouped as training sets. Therefore, this method is only valid for a structure where the response and corresponding ground motions are known in advance and may produce significant errors when predicting for another structure where the prior knowledge about the training set is unknown. This is because the training set does not relate any structural features (e.g., structural geometry, material properties, and reinforcement details that can define a structural system) to the structural response. Once some structural features vary (e.g., structural geometry, material strength or reinforcement details are changed), the training sets will no longer be valid for the new structure. Thus, the well-trained deep learning model does not have predictive capabilities for a new structure where prior knowledge is unknown. A similar procedure was also used by Guarize et al. (2007) for seismic response prediction of marine structures, by Lagaros and Papadrakakis (2012) for seismic response prediction of 3D buildings, and by Wu and Jahanshahi (2018) for seismic response prediction of a three-story steel frame. Therefore, the disadvantages of these methods are consistent with the one proposed by Zhang et al. (2019), and thus they are not suitable for the scope of this dissertation.

2.4 Effect of Data-Related Problems on the Performance of ML Methods

By the review presented in Section 2.3, it can be concluded that the ML methods can provide better prediction than the traditional physics-based approaches in structural and earthquake engineering. However, one important premise for those standard ML methods is that the experimental data set used is high-quality and sufficiently large in size. If a data set has problems, such as being corrupted by outliers, or having large sample bias due to small size, or containing missing values, the conclusions made by these models will no longer be valid. This is because standard ML methods are vulnerable to these data-related problems. Significant performance deterioration will occur for these standard ML methods when a training set is subjected to those mentioned data-related problems. This section reviews the existing methods to deal with these data problems.

2.4.1 Methods for addressing outliers

Typically, standard ML approaches are able to fit and generalize the input data well and can produce extremely good prediction capabilities if the input data is high quality and reasonably large in size. However, if the input data is corrupted by outliers, these ML methods (e.g., methods for regression), especially those that are sensitive to outliers, will yield unreliable prediction. Some of them even break down when the data is contaminated by extreme outliers. Outliers are those observations that are far away from all other observations due to misplaced decimal points, recording or transmission errors, or exceptional phenomena. These are all common occurrences in real-world data (Rousseeuw and Leroy 1987). In general, there are two commonly employed ways to deal with outliers for regression problems (Rousseeuw and Leroy 1987). The first is to use robust approaches while the second method is to construct outlier diagnostics. A robust approach first fits a regression model that adequately addresses the normal data points and then discovers the outliers as those points having large residuals estimated from the robust regression model

(Hampel et al. 2011; Rousseeuw 1984; Rousseeuw and Yohai 1984). On the contrary, outlier diagnostics first identify the outliers and then remove them and fit the remaining normal data points (Rousseeuw and Hubert 2011; Mu and Yuen 2015; Yuen and Mu 2011; Yuen and Ortiz 2017). In some applications, both methods yield exactly the same result. However, outlier diagnostics may result in outliers which are not entirely detected, leading to biased results, while robust regression does not pose such a risk. Robust regression approaches include least absolute deviations (LAD), least trimmed squares (LTS), M-estimators, etc., which were proposed to address the fact that the least squares (LS) method is easily affected by outliers (Rousseeuw and Leroy 1987). These robust methods were originally developed for parametric regression (e.g., linear regression). Recently, many efforts have been made to incorporate these regression approaches into the reformulation of ML methods to enhance their robustness.

LTS has been integrated into backpropagation neural networks (BPNNs) to replace the mean squared error (MSE) as the minimization criterion (Rusiecki 2007), and LAD has been applied in random forests (RFs) (Roy and Larocque 2012) to replace the original LS, leading to model reformulations. As introduced in Section 2.3, LS-SVMR (Suykens et al. 2002) is one of the more frequently used ML methods in structural and earthquake engineering. LS-SVMR is a reformulation of SVMR (Vapnik 1995), which uses the sum of squared errors (SSE) as the loss function and equality constraints in place of inequality constraints to greatly simplify the SVMR formulation. Due to this, LS-SVMR solves a linear system problem instead of the complex quadratic programming (QP) problem, leading to greater computational efficiency. However, the use of SSE as the loss function in the formulation of LS-SVMR leads to a non-robust property. To overcome this problem, the weighted SSE has been adapted as the loss function by Suykens et al. (2002) to substitute the original SSE for the reformulation of LS-SVMR, which resulted in a new

LS-SVMR variant called WLS-SVMR that is robust to outliers. The weight used in WLS-SVMR is a function of residuals estimated by LS-SVMR, where the potential outliers tend to have larger residuals. The points which have larger residuals in the training set will then be assigned smaller weights to reduce their associated negative influence. However, WLS-SVMR breaks down under non-Gaussian noise distribution with heavy tails (i.e., extreme outliers) (Brabanter et al. 2009). To solve this problem, De Brabanter et al. (2009) proposed an iterative version of WLS-SVMR (IWLS-SVMR), where the weights are updated in each iteration to reduce the negative influence of extreme outliers until convergence criteria is reached.

Both WLS-SVMR and IWLS-SVMR are robust, global data-driven regression models, meaning their solution requires the fitting of the entire training set. However, in many cases, the performance of global models can be further improved by local models (Menzies et al. 2011; Hand and Vinciotti 2003; Bottou and Vapnik 1992; Vapink and Bottou 1993; Vapnik 1992). As introduced in Bottou and Vapnik (1992), local learning algorithms attempt to locally adjust the capacity of the training system to the properties of the training set in each area of the input space. This results in a local model that only requires the fitting of a subset of the training data nearby (relevant to) the query point and can overcome the potential negative influence of irrelevant points. Therefore, a robust, local model may provide an improvement under these circumstances when compared to the robust, global models. This is because a robust, local approach can yield a model that both overcomes the negative interference of outliers and avoids the potential negative influence of irrelevant points, achieving a suitable trade-off between the capacity of the learning system and the number of training data points. The majority of existing local models are based on polynomial regression, which means that the local models are polynomial functions (Cleveland 1979; Cleveland and Devlin 1988; Atkeson et al. 1997a; 1997b). These local models have the prior

assumption of polynomial functions within local regions. Thus, the class of such local models are out of the scope of this dissertation. Motivated by these existing solutions introduced previously, a novel, robust version of the local ML model will be proposed in this dissertation to address the problem associated with non-robustness to outliers.

2.4.2 Methods for addressing missing data

An incomplete dataset involves observations with missing values, as shown in Table 2.1. Table 2.1 shows an example where three explanatory variables (or features/predictors) are partially observed and have missing values (represented by 'NAN' values), making this dataset incomplete. Given an incomplete dataset, existing ML approaches fail to directly construct an appropriate data-driven model, as their original analysis procedures are only valid for complete datasets and are not designed to handle missing data (Rubin 1976; Little and Rubin 1987). The most common way to deal with this missing data problem is to simply discard every incomplete observation or case, transforming the incomplete dataset into a reduced, but complete, dataset. Then the reduced, complete dataset can be employed along with any existing ML approaches. Nevertheless, considering all observations in the original incomplete dataset are from realistic cases, this strategy involves throwing away a potentially large amount of useful information, leading to biased inference, and finally misinterpreted conclusions (Rubin 1976; Little and Rubin 1987). Further, this strategy is not always applicable. In specific, in post-earthquake structural evaluations, deleting the data associated with any damaged buildings with critical structural information missing means that further structural analyses of these damaged buildings are not feasible, and thus, the global collapse risk for these damaged buildings will remain unknown, posing a substantial, potential threat.

In addition to simply removing any observations associated with missing data, another effective way to deal with an incomplete dataset is to impute the missing values with plausible candidates, resulting in an imputed, complete dataset. In this way, this type of imputation approach maintains the size of the original incomplete dataset without risking the loss of useful information. The most direct imputation method is single imputation, which is performed by filling in a valid candidate for each missing value, such as imputing each missing value with a fixed value (e.g., the mean of partially observed explanatory variables) or a single value estimated by regression predictions (Batista and Monard 2003). However, single imputation is statistically incorrect, as it implies that those missing values are certain when in fact the missing values have not been observed (Rubin 1976; 1996; 2004). Thus, analyses of the *imputed*, complete dataset by single imputation methods fail to account for the uncertainty due to the missing data.

Table 2.1 Schematic format of an incomplete dataset, where '*NAN*' represents a missing value, and missing values only exist in the partially observed explanatory variables $Z_{(1)}$, $Z_{(2)}$, and $Z_{(3)}$.

Observations	$\boldsymbol{X_1}$	$\cdots$	$\boldsymbol{X_p}$	$\boldsymbol{Z_{(1)}}$	$\boldsymbol{Z_{(2)}}$	$\boldsymbol{Z_{(3)}}$	$\boldsymbol{y}$
$(\boldsymbol{x_1}, \boldsymbol{y_1})$	x_{11}	$\cdots$	x_{1p}	$x_{1(p+1)}$	*NAN*	*NAN*	y_1
$(\boldsymbol{x_2}, \boldsymbol{y_2})$	x_{21}	$\cdots$	x_{2p}	$x_{2(p+1)}$	$x_{2(p+2)}$	*NAN*	y_2
$(\boldsymbol{x_3}, \boldsymbol{y_3})$	x_{31}	$\cdots$	x_{3p}	*NAN*	*NAN*	*NAN*	y_3
$(\boldsymbol{x_4}, \boldsymbol{y_4})$	x_{41}	$\cdots$	x_{4p}	$x_{4(p+1)}$	$x_{4(p+2)}$	$x_{4(p+3)}$	y_4
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$(\boldsymbol{x_n}, \boldsymbol{y_n})$	x_{n1}	$\cdots$	x_{np}	$x_{n(p+1)}$	$x_{n(p+2)}$	*NAN*	y_n

As an alternative, a multiple imputation (MI) method was developed by Rubin (Rubin 2004) to address this drawback. The method of MI has become an extremely popular means for handling incomplete datasets in statistical analyses. The MI approach involves filling in each missing value with several plausible candidates, creating *several imputed*, complete datasets for analyses. Each

dataset is analyzed independently using techniques designed for the complete dataset, and then, the analyzed results are combined in such a way that the uncertainty due to missing data may also be incorporated into the analyses (Rubin 1996; 2004). Two popularly used approaches to create multiple plausible candidates for MI include joint modeling (*JM*) of a multivariate imputation model specification (Schafer 1997; Schafer and Yucel 2002) for all of the partially observed explanatory variables (e.g., $\boldsymbol{Z}_{(1)}$, $\boldsymbol{Z}_{(2)}$, and $\boldsymbol{Z}_{(3)}$ in Table 2.1) and fully conditional specifications (*FCS*) of a series of univariate imputation models (Buuren and Groothuis-Oudshoorn 2010) for each partially observed explanatory variable, conditional on all the other variables.

JM involves specifying a multivariate distribution for the missing data and drawing plausible candidates from the corresponding posterior predictive distributions via a Markov chain Monte Carlo (MCMC) approximation (Schafer 1997; Schafer and Yucel 2002). The *JM* methodology is attractive when the specified multivariate distribution is a reasonable representation of the population distribution of the data. The commonly used multivariate distributions specified by JM techniques for imputation include the multivariate normal model, the multinomial log-linear model, and the general location model for mixed continuous and discrete variables (Liu and Rubin 1998). All of the mentioned *JM* techniques are discussed in greater detail in Schafer (1997). However, it is often challenging to specify a correct multivariate distribution for the missing data (Buuren and Groothuis-Oudshoorn 2010). As an alternative to *JM*, *FCS* specifies the multivariate imputation model on a variable-by-variable basis by a set of conditional densities, one for each partially observed explanatory variable (Buuren and Groothuis-Oudshoorn 2010). Given starting values, *FCS* draws plausible candidates by iterating throughout all conditional densities. Compared to *JM*, the use of *FCS* is much more flexible. This is because, for each partially observed explanatory variable (e.g., continuous or discrete variable), an appropriate

univariate model can be selected. This strategy is more attractive than *JM* in cases where there is no evident, appropriate multivariate distribution for the data. Nevertheless, *FCS* also has a drawback, that is, the conditional densities may be incompatible. This means that there may not exist a joint density such that the conditional densities for each of the partially observed explanatory variables are fully conditional (e.g., the iterations cannot reach convergence). Additionally, both *JM* and *FCS* produce plausible candidates for missing values in terms of simulation. The candidates obtained by simulation may be outside the observed data range due to the model misspecification of either *JM* or *FCS*, leading to meaningless imputation results (Little and Rubin 1987).

In Bayesian parameter estimation, a joint distribution can be factored as a product of conditional and marginal distributions (Hoff 2009; Raghunathan et al. 2001). By appropriately specifying the univariate distribution for each partially observed explanatory variable as either a marginal or conditional distribution, the joint distribution for the entire set of explanatory variables with missing values can be achieved. Motivated by this, this dissertation will propose a novel MI approach to create multiple plausible candidates for imputing each missing value with consideration of the uncertainty due to missing data.

2.4.3 Methods for addressing small datasets

Small datasets are an extremely challenging problem in the ML realm, and in specific, in regression scenarios, as the lack of relevant data can lead to ML models that have large bias. This is because, when a dataset is small, it may lead to a biased sample. This means that the sample points in the small dataset cannot accurately represent the distribution of a target domain and cannot reflect the underlying patterns in the target domain data (Quinonero Candela et al. 2009), leading to large bias in the final, fully-trained ML model for prediction in the target domain. Transfer learning (TL)

aims to address the problems with sample bias induced by small datasets by transferring ML models trained with a relevant large data set to improve prediction (Pan and Yang 2009; Torrey and Shavlik 2010; Weiss et al. 2016). In this dissertation, the small dataset is from the "target domain" and the large dataset is from the "source domain". The target and source domains are typically assumed to be somewhat different but related to each other in many TL approaches (Pan and Yang 2009), and thus, the well-trained ML models from the source domain(s) can be applied to the prediction on the target domain. This seems to deviate from the default assumption in many standard ML settings, where the training and test datasets are independently and identically distributed (i.i.d), as the dataset is shifted (Cortes and Mohri 2014; Huang et al. 2007; Gretton et al. 2009; Quinonero Candela et al. 2009). Mathematically speaking, dataset shift happens when two datasets are drawn from two different distributions (Quinonero Candela et al. 2009). Specifically, given the distributions of the source and target domains, one can sample the training dataset $\{(\boldsymbol{x}_j^S, y_j^S)\}_{j=1}^n$ from the source domain distribution $p^S(\boldsymbol{x}, y)$ and the test dataset $\{(\boldsymbol{x}_k^T, y_k^T)\}_{k=1}^m$ from the target domain distribution $p^T(\boldsymbol{x}, y)$, where $\boldsymbol{x} \in R^p$ and $y \in R$. A dataset shift is when $p^S(\boldsymbol{x}, y) \neq p^T(\boldsymbol{x}, y)$.

However, the majority of existing TL approaches are designed for classification problems (Li and Chaspari 2019; Feng and Chaspari 2019; Gao and Mosalam 2018; Pan and Yang 2009), but less attention has been paid on regression problems (Pardoe and Stone 2010; Salaken et al. 2019). The main difference between classification and regression problems is that the response variable for classification problems is discrete while that for regression problems is continuous (James et al. 2013). This difference strictly restricts the direct use of some existing TL approaches for addressing regression problems (i.e., some TL methods for classification must be modified for their use in regression settings, e.g., the work in Pardoe and Stone 2010). Besides, the existing

regression-based TL methods generally assume that the target and source domains are related to each other (Garcke and Vanck 2014; Pardoe and Stone 2010). Therefore, these TL methods may work well for the regression problems when the source and target domains are related but may work poorly when two domains are unrelated. The relevance is represented by the joint distributions of two domains (Huang et al. 2007). According to Bayes rule, the joint distribution can be written as $p(\boldsymbol{x}, y) = p(\boldsymbol{x}|y)p(y) = p(y|\boldsymbol{x})p(\boldsymbol{x})$. The equation $p(\boldsymbol{x}, y) = p(\boldsymbol{x}|y)p(y)$ is called the generative model, while the equation $p(\boldsymbol{x}, y) = p(y|\boldsymbol{x})p(\boldsymbol{x})$ is called the discriminative model (Garcke and Vanck 2014; Quinonero Candela et al. 2009). The majority of existing TL approaches focus on the discriminative approach. Thus, $p^S(y|\boldsymbol{x})p^S(\boldsymbol{x}) \neq p^T(y|\boldsymbol{x})p^T(\boldsymbol{x})$ (i.e., the source and target domains are different) is achieved by either different marginal distributions, i.e., $p^S(\boldsymbol{x}) \neq p^T(\boldsymbol{x})$ (also called covariate shift) (Quinonero Candela et al. 2009) or different posterior distributions, i.e., $p^S(y|\boldsymbol{x}) \neq p^T(y|\boldsymbol{x})$ or both.

As mentioned previously, the majority of these approaches have been used to deal with classification problems, and only a few recent research efforts have focused on regression problems. Pardoe and Stone (2010) modified two existing boosting-based classification TL models, ExpBoost (Rettinger et al. 2006) and TrAdaBoost (Dai et al. 2007), to form two TL models called ExpBoost.R2 and Two-stage TrAdaBoost.R2 for regression problems. Both of these TL models are based on AdaBoost.R2 (Drucker 1997), where the reweighting of instances (i.e., data points) that have larger residuals predicted by a learner (i.e., ML model) is achieved by normalizing errors into adjusted errors within the range [0, 1] in each boosting iteration. The proposed boosting-based transfer regression models are validated effectively by numerical experiments. Garcke and Vanck (2013) proposed two approaches for inductive transfer regression based on importance weighting. These two methods are to estimate a weight which is a density ratio of the target and

source data. The first one relies on the prediction performance of an ML model learned from the data in the source domain, while the second one minimizes the Kullback-Leibler divergence (Sugiyama et al. 2008) between two distributions of the target and source data. Numerical experiments are performed, and results indicate that the former is better than the latter. A seed-based TL model for regression problems was proposed by Salaken et al. (2019). In this approach, each sample point in the target domain is regarded as a seed for initiating the transfer of the source data. An auto-encoder deep learning technique is used to transform the source data into an abstracted feature space, where the number of features for the data in the source domain matches that in the target domain. Then a k-means clustering algorithm, with the number of clusters equal to the number of sample points in the target domain, is applied to cluster the source domain data, and each target domain sample point is appended with a relevant cluster by minimizing the Euclidean distance. The effectiveness of this method is verified by numerical results.

Although these mentioned regression-based TL approaches can reduce the effect of small sample bias and thus improve prediction performance for small datasets, such capabilities may be limited to the transfer between two related domains, as validated in the numerical experiments. If the source and target domain data are far apart and unrelated, these methods may no longer be valid because these approaches may not be able to extract the shared information from two unrelated domains. To alleviate this limitation, this dissertation will propose a novel TL approach for regression problems.

2.5 Summary

This chapter reviewed the relevant existing work on seismic response prediction using both traditional physics-based approaches and emerging ML-based techniques as well as reviewed the existing methods to reduce the negative effect of data-related problems on the performance of ML-based techniques. It can be concluded that, for physics-based approaches, the methods to define the nonlinear properties of RC columns depend primarily on empirical and semi-empirical models (e.g., material and structural component constitutive equations), which may not be able to capture the underlying patterns in the experimental data. For ML-based techniques, the existing methods do not relate the structural features to the seismic response history, which cannot capture the changes in nonlinear behavior due to changes in structural features. Also, the existing applications of ML-based techniques in structural and earthquake engineering require that the dataset used is high-quality and sufficiently large in size. However, real-world datasets are likely subjected to outliers, large sample bias due to small size, and missing values, where all can significantly degrade the prediction capability of standard ML methods.

In sum, although many researchers have currently applied various ML methods to address the limitations of existing physics-based approaches in structural and earthquake engineering, there are still many existing challenges which remain unaddressed. The development of solutions to these problems is critical to determine if data-driven computing methods can be employed in structural seismic response prediction, and therefore, they will be addressed in this dissertation.

CHAPTER III

DATASETS, VALIDATION, AND ASSESSMENT

3.1 Overview

This chapter presents the development of two RC column databases including 262 rectangular RC columns and 160 circular RC columns covering flexure-, shear-, and flexure-shear failure modes. The specimens in both of the RC column databases are subjected to displacement-controlled quasi-static cyclic loading. A modified hysteretic model and a hybrid optimization algorithm are developed to extract the critical parameters from the experimental force-displacement data. Finally, the column features and corresponding extracted critical parameters are paired to form the RC column datasets, which will be used for the research work in this dissertation. Additionally, the validation methods for machine learning approaches are also introduced along with the assessment metrics used to quantify the performance of the ML models created in this work.

3.2 Material and Geometric Properties of RC Columns Databases

For the rectangular RC column database, 208 specimens were recorded as flexure failures, 18 specimens were classified as shear failures, and the remaining 36 specimens were recorded as flexure-shear failures. The first 194 of the 208 flexure-critical columns, the 18 shear-critical columns, and the 36 flexure-shear-critical columns were extracted from the database compiled by Berry et al. (2004). Among the last 14 of the 208 flexure-critical columns, four columns were from Eom et al. (2014), six columns were from Verderame et al. (2008), and the last four columns were from Xie et al. (2014). The force-displacement data for all of the columns in the database have been modified and treated as cantilever cases to remain consistent with those tests extracted from Berry et al. (2004). The original number of columns with rectangular sections classified as flexure failures developed by Berry et al. (2004) is 199. The five columns excluded in this dissertation (Nos. 147, 148, 149, 181, and 182) did not contain all the necessary information (i.e., loss of force-displacement data) required in this study. The specimen numbers are compatible with the naming conventions in the original references.

Table 3.1 Statistical range of material and geometric properties for the rectangular RC column database.

Property	Minimum	Maximum	Mean	Std.Dev
Shear span to effective depth ratio a/d	1.08	8.40	3.84	1.57
Stirrup spacing to effective depth ratio s/d	0.11	1.14	0.32	0.21
Concrete compressive strength f_c (MPa)	16	118	50.40	28.72
Longitudinal reinforcement yield stress f_{yl} (MPa)	318	635	437.58	65.88
Transverse reinforcement yield stress f_{yt} (MPa)	249	1424	486.91	217.57
Longitudinal reinforcement ratio $p_l=A_{sl}/bh$	0.01	0.06	0.02	0.01
Transverse reinforcement ratio $p_t = A_{st}/bs$	0.0006	0.03	0.008	0.005
Axial load ratio (P/A_gf_c)	0	0.9	0.26	0.19

Table 3.1 presents the statistical ranges of the properties for the columns in this database. The full database with the input parameters and all response variables considered throughout this work is presented in Appendix A. For the circular RC column database, 98 specimens were recorded as flexure failures, 32 specimens were classified as shear failures, and the remaining 30 specimens were recorded as flexure-shear failures. All of these specimens were extracted from the database compiled by Berry et al. (2004). The force-displacement data for all of the columns in the database recorded in the tests have also been modified and treated as cantilever cases by Berry et al. (2004). The original number of columns with circular sections classified as flexure and flexure-shear failures developed by Berry et al. (2004) are 99 and 32, respectively. The three columns excluded in this dissertation (No. 38 for flexure and Nos. 27 and 29 for flexure-shear) did not contain all the necessary information required in this study. Table 3.2 presents the statistical ranges of the properties of the columns in this database. The full database with the input parameters and all response variables considered throughout this work is presented in Appendix B.

Table 3.2 Statistical range of material and geometric properties for the circular RC column database.

Property	Minimum	Maximum	Mean	Std.Dev
Shear span to effective depth ratio a/d	1.18	10.49	3.64	2.13
Stirrup spacing to effective depth ratio s/d	0.00	0.73	0.17	0.11
Concrete compressive strength f_c (MPa)	18.9	90	37.16	14.39
Longitudinal reinforcement yield stress f_{yl} (MPa)	240	565.4	415.52	62.58
Transverse reinforcement yield stress f_{yt} (MPa)	0.00	1000	411.62	154.22
Longitudinal reinforcement ratio $p_l=A_{sl}/A_g$	0.0046	0.0558	0.0265	0.0104
Transverse reinforcement ratio $p_t = 4A_{st}/(D-2c)s$	0.00	0.0427	0.0099	0.0075
Axial load ratio (P/A_gf_c)	0.00	0.74	0.15	0.15

3.3 Development of RC Column Datasets

This section introduces a modified hysteretic model and a hybrid optimization algorithm, which are used to extract the critical parameters that govern the shape of the hysteretic loops in the experimental force-displacement data. The column features introduced in **Section 3.2** and corresponding extracted critical parameters for each column specimen in the databases are paired to form the RC column datasets used throughout this research work. The detailed information is as follows.

3.3.1 Modified three-parameter hysteretic model

Typically, the experimental force-displacement data for an RC column subjected to reversed cyclic loading is composed of more than ten thousand data points, where each point is comprised of the applied displacement paired with the corresponding measured force. It is inappropriate to directly pair the column features and the corresponding experimental force-displacement data. This is because a vector constitutes an RC column's features while the experimental force-displacement data is a two-dimensional matrix with the row dimension of greater than ten thousand. Therefore, it is impractical to form a dataset where the predictors are a vector consisting of column features but the response variables are a two-dimensional matrix. As an alternative, the nonlinear relation of the experimental force-displacement data can be represented by a hysteretic model, which is governed by some critical parameters including the backbone curve and hysteretic parameters. It is practical to calibrate a hysteretic model with experimental force-displacement data by tuning the critical parameters. Once the calibrated hysteretic model can perfectly reproduce the nonlinear relation exhibited by the experimental data, the corresponding optimal critical parameters can be regarded as representative of the experimental data and thus be extracted as the response variables

as a vector. In this direction, a dataset where both predictors and response variables are vectors can be developed.

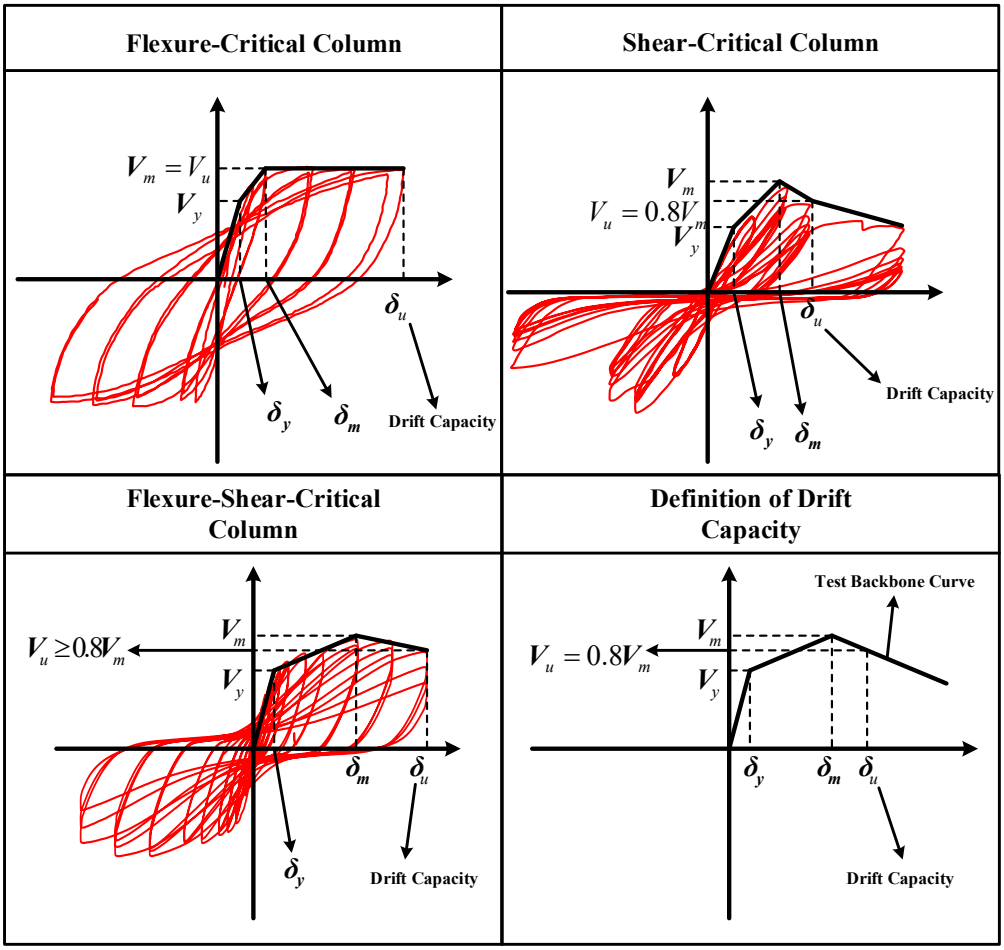


Figure 3.1 Hysteretic behavior characteristics of RC flexure-, shear-, and flexure-shear-critical columns and their definitions of cyclic backbone curves.

The selection of the hysteretic model is a very important step since it is related to the quality of the dataset. A high-quality column data can fully represent the column's nonlinear behavior observed experimentally. Therefore, the selected hysteretic model should be able to capture the various behavioral characteristics of the desired component (RC columns in this case) as observed experimentally. In this case, that includes the hardening, softening, and pinching behavior, and

stiffness and strength deterioration, as shown in Figure 3.1. Few of the existing hysteretic models are versatile enough to describe the various behaviors of RC columns, especially behaviors observed experimentally for RC non-ductile columns (e.g., shear-critical columns), such as the apparent softening behavior, pinching behavior, and significant stiffness and strength deterioration, as shown in Figure 3.1. One of the more popular hysteretic models, proposed by Ibarra et al. (2005), is a versatile tool that can capture various nonlinear characteristics of RC columns observed experimentally. Nevertheless, implementation of this model requires many hysteretic parameters (i.e., parameters governing the shape of the hysteretic loops), which increases the difficulty in calibrating the hysteretic model and thus in obtaining a high-quality dataset. The traditional three-parameter hysteretic model proposed by Park et al. (1987) is both versatile and simple, with only three hysteretic parameters. It is able to describe the hysteretic behavior of various types of RC columns by appropriately tuning these three hysteretic parameters. However, this model does not contain a softening branching in the monotonic backbone curve. This section describes the modifications to the traditional three-parameter hysteretic model, which incorporates a softening branch into the hysteretic model to describe deterioration in the backbone curve.

3.3.1.1 Backbone curve

Two types of monotonic backbone curves will be defined for the purpose of this dissertation: one is for the case with deterioration, and the other is for the case without deterioration, as shown in Figure 3.2(*a*). If no deterioration occurs, the backbone curve is defined by eight parameters including the forces and displacements at yield and maximum points in the positive and negative loading directions. After the maximum points, the gray horizontal lines are applied. If deterioration is included, a softening branch after the maximum point is initiated up to the residual point. After

the residual point, the force will remain constant as the displacement increases. The hysteretic model allows these backbone curve parameters to take on different values in the positive and negative loading directions.

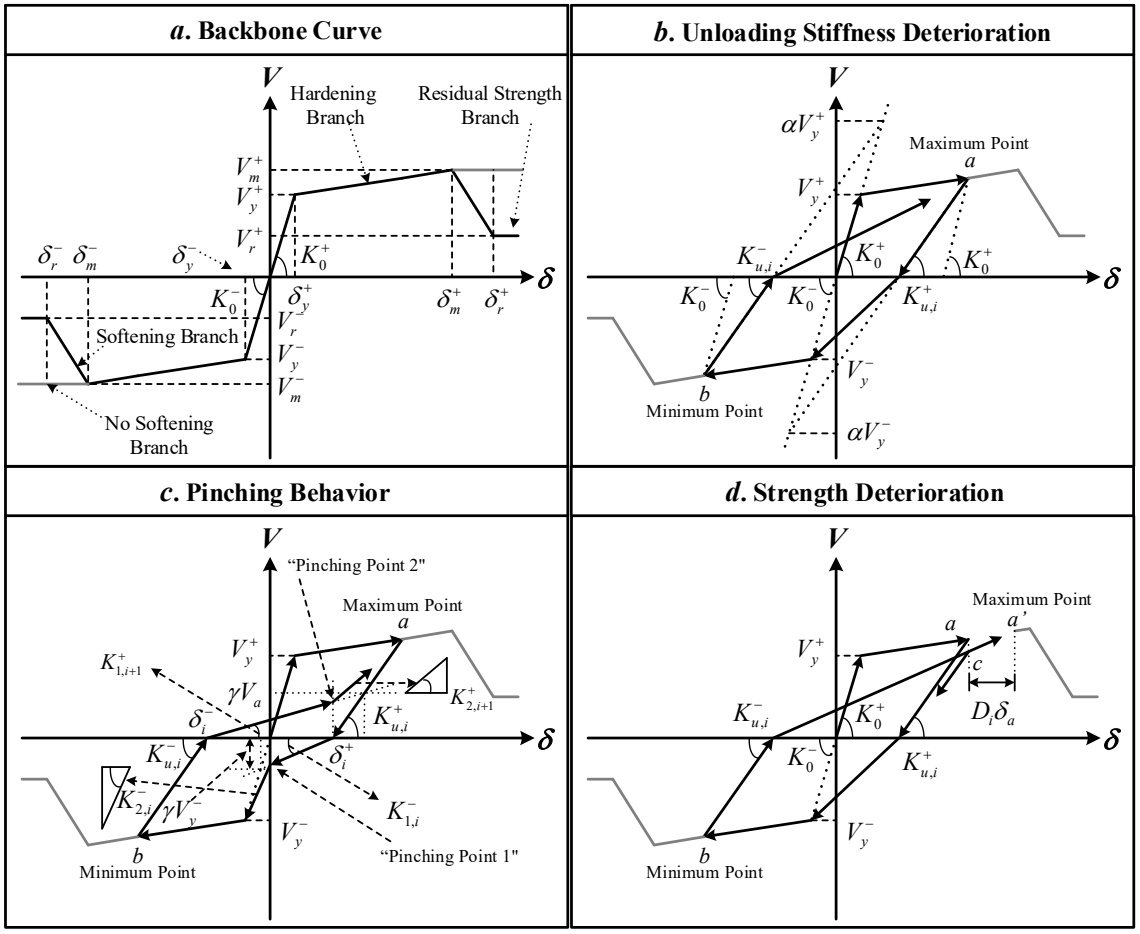


Figure 3.2 Modified three-parameter hysteretic model incorporating the deterioration in the backbone curve.

3.3.1.2 Unloading stiffness deterioration

The deterioration in unloading stiffness is incorporated into the model by introducing a point in the positive direction $\left(\alpha V_y^+ / K_0^+, \alpha V_y^+\right)$ and a point in the negative direction $\left(\alpha V_y^- / K_0^-, \alpha V_y^-\right)$, as shown in Figure 3.18(b). The parameter α has the same function as that in the traditional three-

parameter hysteretic model, which specifies the degree of stiffness deterioration. If $\alpha \to \infty$, this corresponds to stiffness deterioration which does not exist, and the smaller the value of α, the more serious the stiffness deterioration. Figure 3.2(b) shows the change in unloading stiffnesses $K_{u,i}^{+}$ and $K_{u,i}^{-}$ relative to the initial stiffnesses K_0^{+} and K_0^{-} in the positive and negative directions for a load cycle i. The equations below are used to calculate the unloading stiffnesses.

$$K_{u,i}^{+} = \frac{V_a - \alpha V_y^{-}}{\delta_a - \alpha V_y^{-}/K_0^{-}} \tag{3.1}$$

$$K_{u,i}^{-} = \frac{V_b - \alpha V_y^{+}}{\delta_b - \alpha V_y^{+}/K_0^{+}} \tag{3.2}$$

where point a (δ_a, V_a) represents the target maximum point in the positive direction at cycle i; and point b (δ_b, V_b) represents the target minimum point in the negative direction at cycle i.

Note that the maximum and minimum points are also called reversal points. Eqs. (3.1) and (3.2) are applied to all load cycles for an RC column under cyclic loading to calculate the unloading stiffness of reversal points by replacing the coordinates of a and b with the coordinates of new reversal points.

3.3.1.3 Pinching behavior

The pinching behavior depicted herein is similar to the traditional three-parameter hysteretic model except for parameter γ. When pinching behavior is included, γ will lower the force at the maximum or minimum point of the current load cycle. For example, as shown in Figure 3.2(c), the initial path of the reloading line is directed towards "pinching point 1", which can be defined by stiffness $K_{1,i}^{-}$. Then the reloading path is guided towards the minimum point of the current load cycle (i.e., yield point at the current cycle i), which can be defined by stiffness $K_{2,i}^{-}$. With continued loading in the negative direction, the target minimum point b at the current cycle i is reached and will become the new minimum point for the next load cycle. A similar procedure happens for

"pinching point 2" for the next load cycle (i.e., cycle *i*+1), and the maximum point for the next load cycle is point *a*. In this way, the value of the parameter γ can be restricted within [0,1], where 0 represents the most apparent pinching behavior, and 1 represents the opposite. The stiffnesses to describe the reloading lines in the negative and positive directions at the current and next load cycles are calculated as follows.

$$\delta_i^+ = \delta_a - \frac{V_a}{K_{u,i}^+} \tag{3.3}$$

$$\delta_i^- = \delta_b - \frac{V_b}{K_{u,i}^-} \tag{3.4}$$

$$K_{1,i}^- = \frac{\gamma V_y^-}{\delta_i^+ - \gamma V_y^- / K_0^-} \tag{3.5}$$

$$K_{2,i}^- = \frac{V_y^- - \delta_i^+ K_{1,i}^-}{\delta_y^-} \tag{3.6}$$

$$K_{1,i+1}^+ = \frac{\gamma V_a}{\delta_i^+ - \delta_i^- + \gamma V_a / K_{u,i}^+} \tag{3.7}$$

$$K_{2,i+1}^+ = \frac{V_a - (\delta_i^+ - \delta_i^-) K_{1,i}^+}{\delta_a - \delta_i^+} \tag{3.8}$$

The values in Equations 3.3-3.8 are updated for each load cycle when the maximum or minimum point is updated.

3.3.1.4 Strength deterioration

As shown in figure 3.2(*d*), cyclic strength deterioration occurs when the maximum or minimum deformation exceeds the yield deformation in the positive or negative direction for the current load cycle. Specifically, the deformation associated with point *a* in Figure 3.2(d) exceeds the yield deformation in the positive direction for cycle *i*. The reloading path in the positive direction for the next load cycle is directed towards the new maximum point *a'* rather than *a* due to deterioration. If a load reversal occurs at point *c*, the force at point *c* is less than that at point *a* due to the cyclic strength deterioration. This cyclic strength deterioration mode is similar to that in the traditional

three-parameter hysteretic model. The difference is that the deterioration parameter β defined herein is a user-defined parameter, which controls the degree of accumulated hysteretic energy dissipation affecting the cyclic strength deterioration and is calculated as follows:

$$D_i = \beta \left(\frac{\sum_{j=1}^{i} E_j}{E_{ult}} \right) \tag{3.9}$$

where $\sum_{j=1}^{i} E_j$ is the hysteretic energy dissipated in all previous cycles; and E_{ult} is the ultimate hysteretic energy dissipation capacity which is expressed as the area enclosed by the backbone curve.

3.3.2 Extraction of optimal critical parameters

By appropriately establishing the monotonic backbone curve and hysteretic parameters, the proposed hysteretic model can reproduce various behavioral characteristics of RC columns observed in the physical experiments, such as pinching behavior, stiffness and strength deterioration, hardening and softening behavior, and variability of hysteretic loop areas under repeated load reversals. More importantly, this model is very simple and only requires three hysteretic parameters to be established. This enhances the model's applicability in practice. In this dissertation, the proposed hysteretic model is calibrated via the experimental force-displacement data. The final dataset consists of the predictors that define an RC column (i.e., column features) and the response variables that define the hysteretic behavior of the column. The detailed information regarding how the dataset is developed is presented in the following sub-sections.

3.3.2.1 Extraction of monotonic backbone curve parameters

For the RC column specimens presented in **Section 3.2**, monotonic response data is not available, and only cyclic response data is available. There is no direct way to extract the optimal parameters that define the monotonic backbone curve from the cyclic response data. Haselton et al. (2009) suggested that the monotonic backbone curve can be approximated based on extrapolation from

the cyclic backbone curve. The optimal parameters associated with the cyclic backbone curve are easy to extract from the force-displacement data. Many existing methods have already been proposed to extract these parameters (Elwood and Moehle 2005; Ghannoum and Moehle 2011; Sezen and Moehle 2004). In this dissertation, the methodology suggested by Haselton et al. (2009) is utilized to relate the extracted optimal values of the cyclic backbone curve to the monotonic backbone curve. As shown in Figure 3.3, all the optimal parameters that define the cyclic backbone curve are also used to define the monotonic backbone curve except for the ultimate drift ratio (i.e., drift capacity). The ultimate drift ratio for the monotonic backbone curve is defined as two times the value δ_u for the cyclic backbone curve (i.e., $2\delta_u$). We found that this method is convenient to implement and also works very well. The monotonic backbone curve is employed rather than the cyclic backbone curve because the monotonic backbone curve will eventually shrink to the cyclic backbone curve under reversed cyclic loading due to the cyclic strength deterioration (Deierlein et al 2010; Haselton et al. 2009; Ibarra et al. 2005). Therefore, the effect of the monotonic backbone curve is to ensure that the shrunk cyclic backbone curve closely matches the one observed experimentally (Deierlein et al 2010; Haselton et al. 2009; Ibarra et al. 2005). An example to show how the monotonic backbone curve shrinks to the cyclic backbone curve due to deterioration will be provided in **Section 3.3.2.3**.

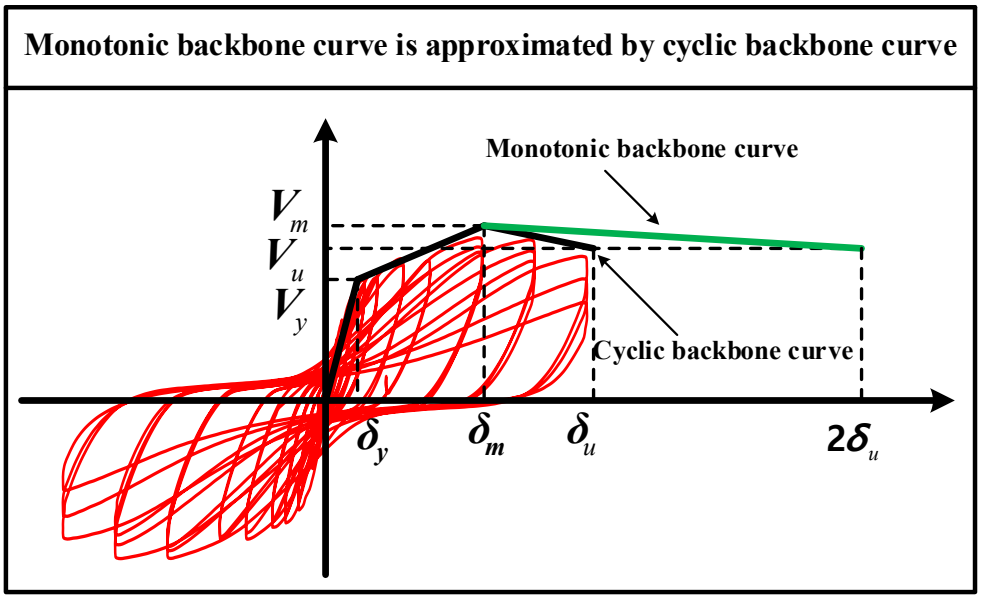


Figure 3.3 Schematic for approximating the monotonic backbone curve from the cyclic backbone curve.

For both databases, the cyclic backbone curve parameters were extracted from available hysteretic curves of base shear versus lateral displacement (i.e., experimental force-displacement data) when not specifically reported in the experimental tests. These parameters include the yield shear force (V_y), drift ratio at V_y (δ_y=Δ_y/l*100, where Δ_y is lateral drift at V_y, and l is column clear length), maximum shear force (V_m), drift ratio at V_m (δ_m=Δ_m/l*100, where Δ_m is lateral drift at V_m), ultimate shear force (V_u), and drift ratio at V_u (δ_u=Δ_u/l*100, where Δ_u is lateral drift at V_u). For V_m and δ_m, these two values can be directly extracted at the point of maximum shear.

To extract V_y and δ_y, the method proposed by Sezen and Moehle (2004) is used. The first step of this approach is to define the initial effective stiffness, which is the secant intersecting the point of the hysteretic curve at 70% of the maximum shear force. Then, δ_y is defined by the intersection of this secant with a horizontal line passing through the maximum shear force. The yield shear force V_y is defined by the force at δ_y on the hysteretic curve. As the nonlinear behavior for flexure-, shear-, and flexure-shear-critical RC columns is different (Figure 3.1), the definition

of drift capacity is also different. By observation of Figure 3.1, the behavior of a flexure-critical column typically does not contain the strain-softening branch, but for shear- and flexure-shear-critical columns, the backbone curve typically does contain a strain-softening branch. Strain softening describes the strength deterioration resulting from concrete crushing and rebar buckling and fracture.

To extract δ_u, three different cases are considered. For the first case, where there is no apparent strength deterioration phenomenon present, the drift capacity δ_u is defined as the drift at the ultimate shear force (i.e., ultimate drift ratio). For the second case, where strain-softening exists and the shear strength drops below 80% of the maximum shear force value, the method proposed by Elwood and Moehle (2005) is used. In this method, the drift capacity δ_u is defined as the displacement where the shear resistance drops to 80% of the maximum shear force. For the third case, where strain-softening exists and the shear strength does not drop below 80% of the maximum shear force value, the drift capacity δ_u is taken as the maximum drift in the backbone curve. The statistical ranges of these extracted response values for the circular and rectangular RC column databases are summarized in Tables 3.3 and 3.4, respectively. Full databases for rectangular and circular RC column databases are presented in Appendices A and B.

3.3.2.2 Optimizing the hysteretic parameters

As mentioned previously, the three hysteretic parameters α, β, and γ govern the degree of unloading stiffness, strength deterioration, and pinching behavior, respectively, which in turn, determine the shape of the hysteretic loops. Thus, they play important roles in reproducing the hysteretic behavior observed experimentally. However, extraction of the three optimal hysteretic parameters from the hysteretic curve is difficult since there is no apparent feature to indicate the optimal values. Different combinations of these three hysteretic parameters may produce

significantly different hysteretic loops, and the optimal combination should be the one from which the simulated hysteretic curve closely matches that observed experimentally. In the proposed approach, a hybrid optimization algorithm is developed to adaptively search for the optimal values of these three hysteretic parameters. Before performing the proposed hybrid optimization algorithm, the parameter space should be defined. In the proposed hysteretic model, the range of each parameter is known: $\alpha \geq 0$, $\beta \geq 0$, and $0 \leq \gamma \leq 1$. By an initial analysis, we found that there is little impact when $\alpha \geq 120$; therefore, $\alpha = 120$ can also be utilized to represent that the unloading stiffness has not deteriorated (i.e., upper bound). Further, when $\beta = 1$, the rate of strength deterioration is very high, and the monotonic backbone curve shrinks significantly; few RC columns exhibit such deterioration rates. Additionally, as introduced previously, when $\gamma = 1$, it means that there is no pinching behavior apparent in the hysteretic curve, and when $\gamma = 0$, the pinching behavior is the most pronounced. Therefore, the three hysteretic parameters are bounded as follows: $0 \leq \alpha \leq 120$, $0 \leq \beta \leq 1$, and $0 \leq \gamma \leq 1$.

The strategy of the hybrid optimization procedure is as follows. First, search for good initial values for the three hysteretic parameters using a metaheuristic global optimization algorithm, called simulated annealing (SA) (Van Laarhoven and Aarts 1987). Then, the Nelder-Mead downhill simplex optimization algorithm (NM-simplex) (Nelder and Mead 1965) is used to further search for the optimal values within a local region encompassing these good initial values. This strategy can effectively avoid the solution of local minima due to the use of the global optimization algorithm (SA). Also, it can directly find the optimal solution that is as close as possible to the exact solution by use of the NM-simplex algorithm. The specific procedure is described in the remainder of this section.

Suppose we have experimental force-displacement data $\{(\delta_i, F_i)\}_{i=1}^{n_p}$ for an RC column, where $\delta_i \in R$ represents a lateral displacement applied to the RC column and $F_i \in R$ represents a lateral force measured at this lateral displacement δ_i. There are a total n_p data points that comprise the force-displacement hysteretic curve observed in the physical experiment. We denote the lateral force estimated at the lateral displacement δ_i using the proposed hysteretic model as $f_{s,i} = f(\delta_i; \boldsymbol{B}, \alpha, \beta, \gamma)$, where $\boldsymbol{B}$ represents the monotonic backbone curve parameters and can be extracted from the experimental force-displacement data in the manner introduced in **Section 3.3.2.1**. Thus, the only unknown information in $f(\delta_i; \boldsymbol{B}, \alpha, \beta, \gamma)$ are the three hysteretic parameters. The optimal values of the three hysteretic parameters $(\alpha_o, \beta_o, \gamma_o)$ can be reached by minimizing the following objective function:

$$\text{Minimize: } J(\alpha, \beta, \gamma) = \sum_{i=1}^{n_p} \left(F_i - f(\delta_i; \boldsymbol{B}, \alpha, \beta, \gamma)\right)^2 \tag{3.10}$$

$$\text{Subject to: } 0 \leq \alpha \leq 120;\ 0 \leq \beta \leq 1;\ 0 \leq \gamma \leq 1$$

The hysteretic parameter tuning procedure using the hybrid SA-NM-simplex algorithm is formulated as follows:

1. Given the objective function $J(\alpha, \beta, \gamma)$ and lower and upper bounds of the three hysteretic parameters, select initial values for the three hysteretic parameters, denoted as $(\alpha_g, \beta_g, \gamma_g)$, where α_g, β_g, and γ_g are the selected values within the specified bounds.
2. Perform the SA procedure to tune α, β, and γ and determine good initial values $(\alpha_s, \beta_s, \gamma_s)$ (those that minimize the objective function $J(\alpha_s, \beta_s, \gamma_s)$ (Eq. 3.10)).
3. Given the good initial values $(\alpha_s, \beta_s, \gamma_s)$, set values $(\alpha_I, \beta_I, \gamma_I)$ to construct initial local regions $[\alpha_s, \alpha_s + \alpha_I]$, $[\beta_s, \beta_s + \beta_I]$, and $[\gamma_s, \gamma_s + \gamma_I]$ and set a stop criterion for the NM-simplex iterative procedure.

4. Perform the NM-simplex iterative procedure in the initial local region. For each iteration, the objective is to search for an improved combination $(\alpha_o, \beta_o, \gamma_o)$ that can continue to decrease the current minimum objective function. The initial local region and improved combination will be updated based on the NM-simplex algorithm.
5. Output the optimal combination $(\alpha_o, \beta_o, \gamma_o)$ when the stop criterion is reached.

Note that it is possible that the combination of good initial values $(\alpha_s, \beta_s, \gamma_s)$ is the final optimal combination $(\alpha_o, \beta_o, \gamma_o)$. This can happen when $J(\alpha_s, \beta_s, \gamma_s)$ is still the minimum after the NM-simplex iterations.

3.3.2.3 Extraction of optimal hysteretic parameters

The optimal hysteretic parameters can be extracted from the experimental hysteretic curve using the proposed hybrid optimization procedure described above. For implementation of this procedure, the initial values of the three hysteretic parameters $(\alpha_g, \beta_g, \gamma_g)$ for the SA tuning procedure are (50, 0, 1). The values $(\alpha_I, \beta_I, \gamma_I)$ for the NM-simplex iterative procedure are (5, 0.05, 0.1). These values are used to optimize the hysteretic parameters of all RC columns in the training set. An example is presented here to show how the proposed hybrid procedure works. Column specimen, "No. IC2", from Sritharan et al. (1996) is randomly selected from the circular RC column dataset (See appendix B). Figure 3.4(a) shows the experimental hysteretic curve enveloped by cyclic and monotonic backbone curves via the extracted optimal backbone curve parameters introduced in **Section 3.3.2.1**. Figure 3.4(b) depicts the hysteretic curve based on an initial guess combination of the three hysteretic parameters, which seems to have a large discrepancy with that observed experimentally. The SA algorithm is then used to tune these three hysteretic parameters, producing a hysteretic curve that matches better with that observed experimentally (Figure 3.4(c)). The three parameters in Figure 3.4(c) are considered as good initial

values for the NM-simplex algorithm, which then is used to tune them, and the final optimal values produce a hysteretic curve that very closely matches that observed experimentally (Figure 3.4(d)). Specifically, given the optimal values $(\alpha = 53.18,\ \beta = 0.10, \gamma = 0.98)$, the monotonic backbone curve (green line) is reduced to the cyclic backbone curve (black line) due to cyclic strength deterioration, as observed in Figure 3.4(d). Finally, the optimal hysteretic parameters $(\alpha = 53.18,\ \beta = 0.10, \gamma = 0.98)$ can be extracted.

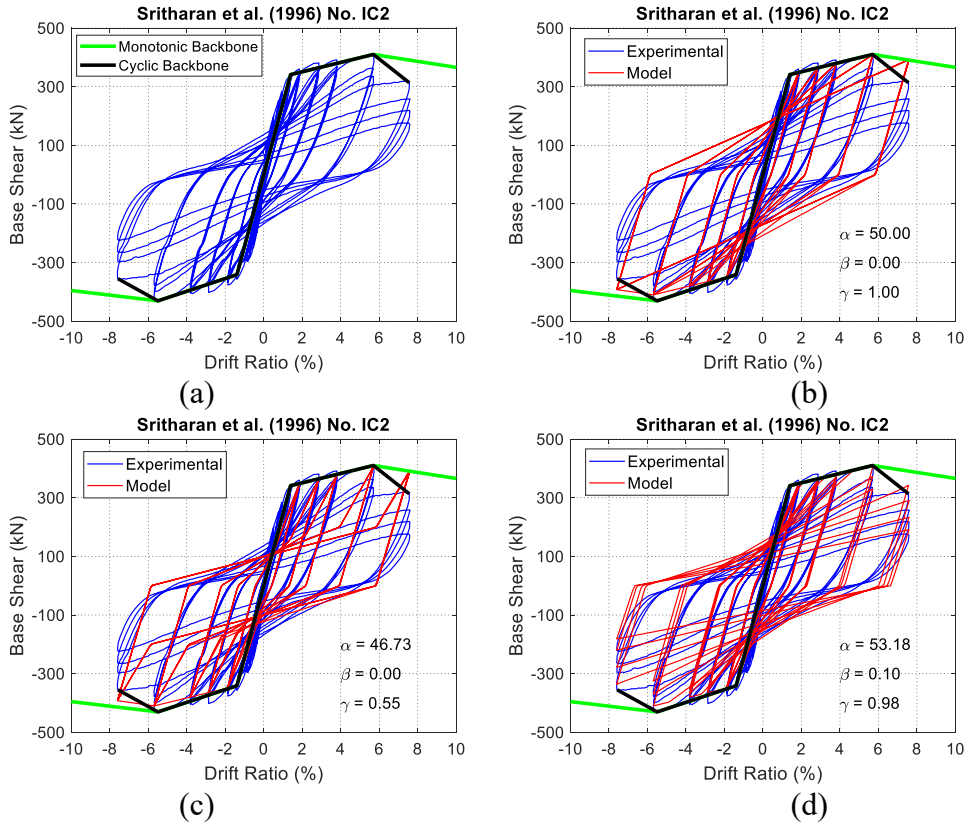


Figure 3.4 Hybrid optimization procedure for the three hysteretic parameters: (a) Experimental hysteretic curve enveloped by the monotonic and cyclic backbone curves; (b) Simulation with initial guess values of the three hysteretic parameters; (c) Result of the SA algorithm; (d) Result of the NA-simplex algorithm.

The three optimal hysteretic parameters for each of the RC columns in the rectangular and circular databases are obtained according to the aforementioned procedure. Since the forces and

displacements in the positive and negative directions for the experimental hysteretic curve are almost identical, the backbone curve parameters in the positive and negative directions are made equivalent in this work. The statistical properties of the optimal cyclic backbone curve and three hysteretic parameters for the circular and rectangular RC column databases are summarized in Tables 3.3 and 3.4. It should be noted that the original number of column specimens in the rectangular RC column database is 262. However, there are ten columns for which the full force-displacement data are not available. These ten columns are from Verderame et al. (2008) and Eom et al. (2014) and thus, are not included for the extraction of hysteretic parameters.

Finally, two datasets including the design parameters or structural features (e.g., section dimensions, material properties, and reinforcement details) and the response variables (e.g., backbone curve and hysteretic parameters) are developed and presented in Appendices A and B. For each column dataset, the structural features can define an individual RC column. The response variables represent the experimental force-displacement data, which quantifies the seismic behavior of the RC column.

Table 3.3 Statistical properties of the optimal cyclic backbone curve and hysteretic parameters for circular RC column database.

Critical Parameters	Minimum	Maximum	Mean	Std.Dev
Yield shear force, V_y (kN)	18.00	2443.90	223.37	246.42
Drift ratio at yield shear, δ_y (%)	0.18	3.23	0.97	0.55
Maximum shear force, V_m (kN)	19.00	2968.00	267.98	298.58
Drift ratio at maximum shear, δ_m (%)	0.26	14.04	2.79	2.15
Ultimate shear force, V_u (kN)	15.25	2558.40	234.14	265.50
Drift ratio at ultimate shear, δ_u (%)	0.43	14.66	4.74	2.79
Stiffness deterioration parameter, α	0.70	110.94	23.20	22.30
Strength deterioration parameter, β	0.00	0.82	0.15	0.20
Pinching parameter, γ	0.22	1.00	0.82	0.23

Table 3.4 Statistical properties of the optimal cyclic backbone curve and hysteretic parameters for rectangular RC column database.

Critical Parameters	Minimum	Maximum	Mean	Std.Dev
Yield shear force, V_y (kN)	24.00	1110.04	178.86	153.06
Drift ratio at yield shear, δ_y (%)	0.20	2.95	0.84	0.37
Maximum shear force, V_m (kN)	29.56	1338.80	212.37	181.91
Drift ratio at maximum shear, δ_m (%)	0.31	7.94	2.01	1.35
Ultimate shear force, V_u (kN)	25.65	1217.01	177.46	157.83
Drift ratio at ultimate shear, δ_u (%)	0.72	9.39	3.72	1.91
Stiffness deterioration parameter, α	0.30	119.42	19.95	22
Strength deterioration parameter, β	0.00	0.93	0.15	0.20
Pinching parameter, γ	0.31	1.00	0.89	0.19

3.4 Validation and Assessment

This section introduces the common ways to validate and assess the performance of machine learning (ML) models. Additionally, the commonly used performance metrics that are also employed in this work, are presented.

3.4.1 Validation set approach

The validation set approach is a very simple strategy for validating the performance of ML models. It involves randomly dividing the available set of observations into two parts, a learning set and a test set or hold-out set. The model is fit on the learning set with the hyper-parameters that are obtained based on a cross-validation procedure, and the fitted model is used to predict the responses for the observations in the test set. The resulting test set error provides an estimate of the test error. It is often used for the initial validation (James et al. 2013). Additionally, it should be noted that when tuning the hyper-parameters with cross-validation procedure, the learning set is split into training and validation sets. The potential hyper-parameter values are justified by the estimated error on validation set for the ML model formed by fitting the training set.

3.4.2 K-fold cross-validation approach

The validation set approach is conceptually simple and is easy to implement. However, the validation estimate of the test error can be highly variable, depending precisely on which observations are included in the learning set and which observations are included in the test set. To alleviate the randomness in selecting testing samples and therefore, enhance the robustness of the results, a K-fold cross-validation process is much more appropriate. This process is illustrated in Figure 3.5 for the case where there are five folds. The whole database is randomly and averagely divided into K data subsets or “folds” where each fold, in turn, serves as a test set. The performance of the ML models can be evaluated by averaging the results of the K data folds. Since each of the

K data folds is mutually exclusive to the others, this validation serves to assess the ML models more robustly and accurately.

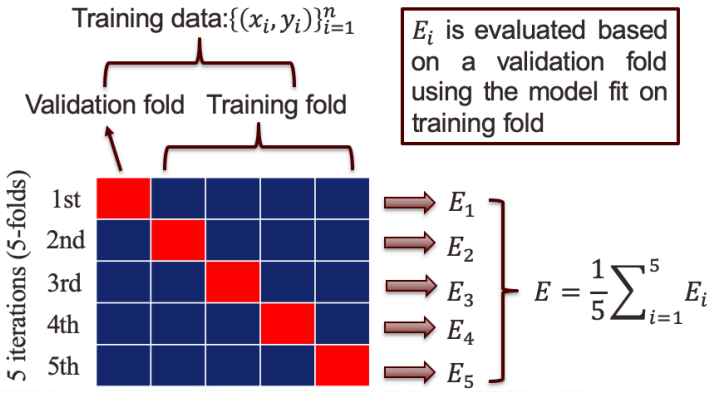


Figure 3.5 Cross-validation procedure with 5 folds for illustration.

3.4.3 Leave-One-Out (LOO) cross-validation approach

The LOO approach involves splitting the dataset into two subsets: one subset consists of a single observation $(\boldsymbol{x}_1, y_1)$ which is used for testing, while the second subset consists of all remaining observations $\{(\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_n, y_n)\}$ which then make up the learning set. In the LOO approach, this split is done *n* times, where *n* = the total number of observations in the dataset, such that each observation is predicted based on the model trained on the remainder of the observations. This approach thus results in predictions with far less bias. Additionally, in contrast to the validation set and K-fold cross-validation approaches which generate different results when applied repeatedly due to randomness in the learning and testing set splits, repeatedly performing the LOO cross validation will always yield the same results. This is because there is no randomness for the LOO cross-validation approach in the training and test set split.

3.4.4 Performance assessment metrics

The statistical metrics (coefficient of determination (R^2), robust R^2 (R_R^2), root mean squared error (RMSE), mean absolute error (MAE), and mean absolute percentage error (MAPE)) are used to quantify the prediction and generalization performance in a comprehensive manner. Given response variable $\boldsymbol{y} = \{y_i\}_{i=1}^n$ and predicted response $\hat{\boldsymbol{y}} = \{\hat{y}_i\}_{i=1}^n$, R^2, R_R^2, RMSE, MAE, and MAPE are calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2} \tag{3.11}$$

$$R_R^2 = 1 - \left(\frac{median(|\boldsymbol{y} - \hat{\boldsymbol{y}}|)}{mad(\boldsymbol{y})} \right)^2 \tag{3.12}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}} \tag{3.13}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \tag{3.14}$$

$$MAPE = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{3.15}$$

where $mad(\boldsymbol{y}) = median(|\boldsymbol{y} - median(\boldsymbol{y})|)$ is the median absolute deviation of $\boldsymbol{y}$.

Both the original and robust variant of R^2 are typically in the range of [0, 1], with 1 representing a perfect prediction. However, in some cases, R^2 could be negative and a negative R^2 value corresponds to extremely poor prediction, which means the model breaks down. Both RMSE, MAE, and MAPE values will be equal to or greater than 0, with 0 representing perfect predictions.

CHAPTER IV

COMPONENT-LEVEL DATA-DRIVEN COMPUTING FRAMEWORK*

4.1 Overview

This section presents the development of a novel data-driven framework for seismic response prediction of structural components. First, a novel multiple-output machine learning (ML) model is developed for generalized hardening behavior prediction based on the rectangular reinforced concrete (RC) column dataset presented in **Chapter III**. Second, a new locally weighted ML model is proposed for generalized softening behavior prediction based on the developed circular RC column dataset (also previously presented in **Chapter III**). Lastly, by integrating the proposed ML model with the proposed hysteretic model presented in **Chapter III**, a novel component-level data-driven framework is developed for generalized, accurate and efficient seismic response history prediction of RC structural components. The proposed framework can directly link the experimental data to nonlinear properties of target RC structural components (columns in this case) minimizing the modeling errors induced by empirical models while still employing universal laws (e.g., Newton's laws of motion). Each method is assessed and validated by comparing the numerical results with the physical experiment data.

*Section 4.2 of this chapter is reprinted with permission "Machine learning–based backbone curve model of reinforced concrete columns subjected to cyclic loading reversals" by Huan Luo and Stephanie Paal, 2018. *Journal of Computing in Civil Engineering*, 32, 04018042, Copyright [2018] by American Society of Civil Engineers.

*Section 4.3 of this chapter is reprinted with permission from "A locally weighted machine learning model for generalized prediction of drift capacity in seismic vulnerability assessments" by Huan Luo and Stephanie Paal, 2019. *Computer-Aided Civil and Infrastructure Engineering*, 34, 935-950, Copyright [2019] by John Wiley and Sons.

4.2 Hardening Behavior Prediction

The section presents a novel machine learning (ML) method for the generalized prediction of the hardening behavior in terms of the cyclic backbone curve without considering the softening behavior of RC columns covering flexure-, shear-, and flexure-shear-critical types. Figure 4.1 shows the cyclic backbone curve (drift at yield shear δ_y, yield shear V_y, drift at maximum shear δ_m, maximum shear V_m) that is often employed to quantify the hardening behavior of an RC column.

The cyclic backbone curve constructed from experimentally derived hysteresis envelope is frequently used to evaluate the seismic behavior of the RC column under cyclic loading. Strain hardening after yielding is a common behavior in the RC column subjected to cyclic loading. Traditionally, this behavior is predicted by a detailed finite element modeling, which is time-consuming and does not have a good generalization performance for flexure-shear- and shear-critical columns. Additionally, as illustrated in **Section 2.3**, the ML method called least squares support vector machines for regression (LS-SVMR) (Suykens et al. 2002) is only valid for single output. But the cyclic backbone curve is composed of four values, which is a multi-output problem. Therefore, the LS-SVMR cannot directly be used for this application. To address these shortcomings, a novel ML-based backbone curve model (ML-BCV) to rapidly predict these curves for flexure-, shear-, and flexure-shear-critical columns is developed. The novel model integrates a multi-output least squares support vector machine for regression (MLS-SVMR) to discover the mapping between input and output variables and a grid search algorithm (GSA) (Bergstra and Bengio 2012) to facilitate the training process.

This section is organized as follows. First, the formulation of MLS-SVMR, which was created to deal with the multi-output case, is described. Following this, the unique integration of the MLS-SVMR and GSA is discussed within the application to cyclic backbone curve prediction (Figure 4.1). The detailed information is presented below.

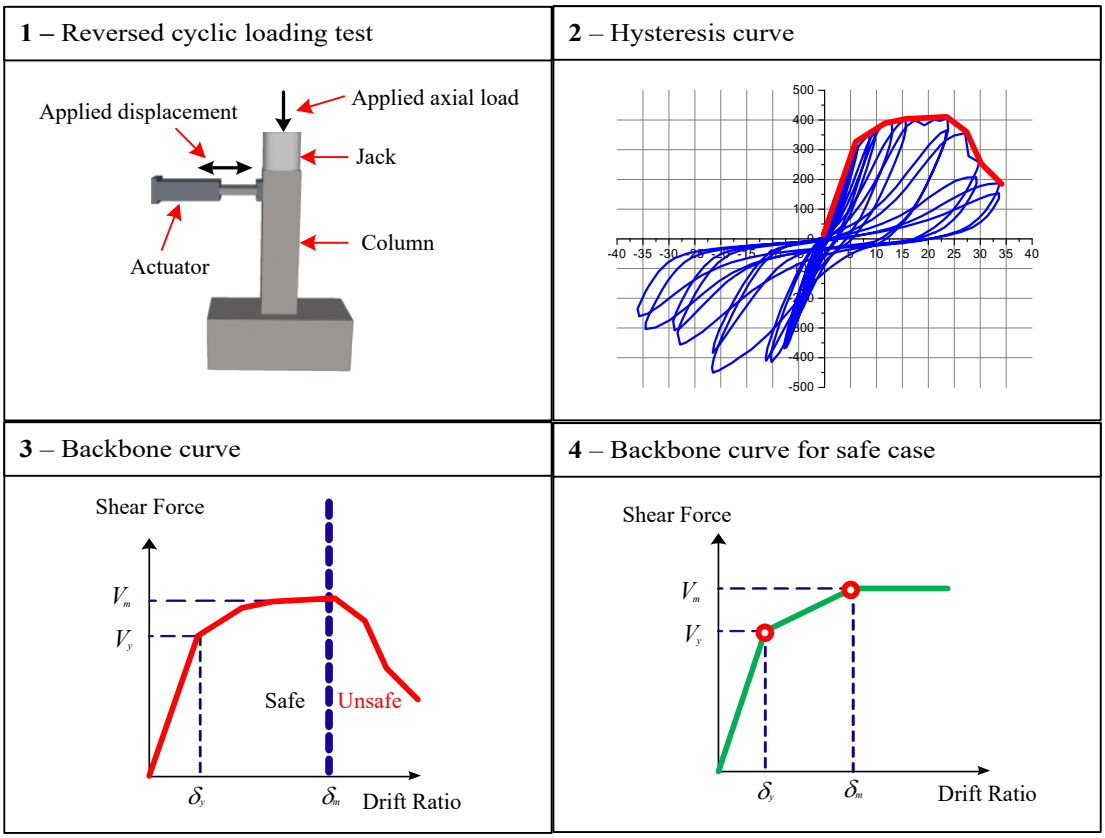


Figure 4.1 Backbone curve that quantifies the hardening behavior of the RC column subjected to cyclic loading reversals.

4.2.1 Integration of MLS-SVMR with GSA

For the remainder of this section, the following notations are used. Let R be the real numbers set. The lowercase bold letter $\boldsymbol{x} \in R^m$ indicates a column vector with m dimensions that can be represented as $\boldsymbol{x} = (x_1, x_2, \dots, x_m)^T$, $\boldsymbol{x}' \in R^n$ indicates a row vector with n dimensions that can be

represented as $\boldsymbol{x}' = (x_1, x_2, \dots, x_n)$, and the capital bold letter $\boldsymbol{X} \in R^{m\times n}$ indicates a matrix with m rows and n columns that can be represented as $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n]$, where $\boldsymbol{x}_1 = (x_{11}, x_{21}, \dots, x_{m1})^T$, $\boldsymbol{x}_2 = (x_{12}, x_{22}, \dots, x_{m2})^T$, and $\boldsymbol{x}_n = (x_{1n}, x_{2n}, \dots, x_{mn})^T$. Let training set $T = \{\boldsymbol{X}, \boldsymbol{Y}\}$, where $\boldsymbol{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n] \in R^{m\times n}$ indicates that the training set has n independent variables, and each of the independent variables has m data points, and $\boldsymbol{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \dots, \boldsymbol{y}_l] \in R^{m\times l}$ indicates that the training set has l dependent variables, and each of the dependent variables also has m data points. The learning objective of the MLS-SVMR can be transformed into the following optimization problem to find $\boldsymbol{W} = [\boldsymbol{w}_1, \boldsymbol{w}_2, \dots, \boldsymbol{w}_l] \in R^{d\times l}$ and $\boldsymbol{B} = diag(b_1, b_2, \dots, b_l) \in R^{l\times l}$:

Minimize $J(\boldsymbol{W}, \boldsymbol{E}) = \frac{1}{2} trace(\boldsymbol{W}^T \boldsymbol{W}) + \frac{1}{2} trace(\boldsymbol{E}^T \boldsymbol{E} \boldsymbol{\gamma})$ or

$$J(\boldsymbol{w}_j, e_{ij}) = \frac{1}{2}\sum_{j=1}^{l} \boldsymbol{w}_j^T \boldsymbol{w}_j + \frac{1}{2}\sum_{j=1}^{l} \gamma_j \left(\sum_{i=1}^{m} e_{ij}^2\right) \tag{4.1}$$

Subjected to $\boldsymbol{Y} = \boldsymbol{\varphi}^T \boldsymbol{W} + \boldsymbol{1}_{m\times l} \boldsymbol{B} + \boldsymbol{E}$ or

$$\mathrm{y}_{ij} = \varphi^T(\boldsymbol{x}_i') \boldsymbol{w}_j + b_j + e_{ij}, i = 1, \dots, m; j = 1, \dots, l \tag{4.2}$$

where $\boldsymbol{\gamma} = diag(\gamma_1, \gamma_2, \dots, \gamma_l) \in R^{l\times l}$ represents a diagonal matrix consisting of a positive real regularized parameter γ_j; $\boldsymbol{E} = [\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_l] \in R^{m\times l}$ represents a matrix consisting of error vectors; $\boldsymbol{1}_{m\times l}$ represents a matrix consisting of 1 elements with m rows and l columns; $\boldsymbol{\varphi} = \left[\varphi(\boldsymbol{x}_1'), \varphi(\boldsymbol{x}_2'), \dots, \varphi(\boldsymbol{x}_m')\right] \in R^{d\times m}$, $\varphi(\cdot): R^n \to R^d$ represents a mapping from n dimensions to some higher dimensional Hilbert space H with d dimensions.

The Lagrangian function for Eqs. (4.1) and (4.2) is formulated as follows.

$$L(\boldsymbol{w}_j, b_j, e_{ij}, \alpha_{ij}) = J(\boldsymbol{w}_j, e_{ij}) - \sum_{i=1}^{m}\sum_{j=1}^{l} \alpha_{ij} \left(\varphi^T(\boldsymbol{x}_i') \boldsymbol{w}_j + b_j + e_{ij} - \mathrm{y}_{ij}\right) \tag{4.3}$$

where α_{ij} represents a Lagrange multiplier.

The Karush-Kuhn-Tucker (KKT) conditions for optimality are adopted by differentiating Eq. (4.3) with the variables to yield the following set of linear equations:

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{w}_j} = 0 \rightarrow \boldsymbol{w}_j = \sum_{i=1}^{m} \alpha_{ij} \varphi(\boldsymbol{x}_i'), j = 1,...,l \\ \dfrac{\partial L}{\partial b_j} = 0 \rightarrow \sum_{i=1}^{m} \alpha_{ij} = 0, j = 1,...,l \\ \dfrac{\partial L}{\partial e_{ij}} = 0 \rightarrow e_{ij} = \dfrac{\alpha_{ij}}{\gamma_j}, i = 1,...,m; j = 1,...,l \\ \dfrac{\partial L}{\partial \alpha_{ij}} = 0 \rightarrow y_{ij} = \varphi^T(\boldsymbol{x}_i')\boldsymbol{w}_j + b_j + e_{ij}, i = 1,...,m; j = 1,...,l \end{cases} \tag{4.4}$$

Rearranging Eq. (4.4) can result in (m+1) linear equation groups, and each of the equation groups consists of (m+1) elements, which are written in matrix format as following.

$$\begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & K(\boldsymbol{x}_1', \boldsymbol{x}_1') + \dfrac{1}{\gamma_j} & K(\boldsymbol{x}_1', \boldsymbol{x}_2') & \cdots & K(\boldsymbol{x}_1', \boldsymbol{x}_m') \\ 1 & K(\boldsymbol{x}_2', \boldsymbol{x}_1') & K(\boldsymbol{x}_2', \boldsymbol{x}_2') + \dfrac{1}{\gamma_j} & \cdots & K(\boldsymbol{x}_2', \boldsymbol{x}_m') \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & K(\boldsymbol{x}_m', \boldsymbol{x}_1') & K(\boldsymbol{x}_m', \boldsymbol{x}_2') & \cdots & K(\boldsymbol{x}_m', \boldsymbol{x}_m') + \dfrac{1}{\gamma_j} \end{bmatrix} \begin{bmatrix} b_j \\ \alpha_{1j} \\ \alpha_{2j} \\ \vdots \\ \alpha_{mj} \end{bmatrix} = \begin{bmatrix} 0 \\ y_{1j} \\ y_{2j} \\ \vdots \\ y_{mj} \end{bmatrix}, j = 1,...,l \tag{4.5}$$

where $K(\boldsymbol{x}_i', \boldsymbol{x}_k') = \varphi^T(\boldsymbol{x}_i')\varphi(\boldsymbol{x}_k')$ is the kernel function that meets the Mercer rule. This means that the inner product $\varphi^T(\boldsymbol{x}_i')\varphi(\boldsymbol{x}_k')$ in the feature space has an equivalent kernel in the original input space. A kernel function is preferable rather than direct formulation of $\varphi(\boldsymbol{x}_i')$ as there are many kernel functions (i.e., linear, polynomial, and radial basis function (RBF)) that can realize the same mapping function of $\varphi(\boldsymbol{x}_i')$ in a more computationally efficient manner (because the inner product calculation of $\varphi^T(\boldsymbol{x}_i')\varphi(\boldsymbol{x}_k')$ is not necessary). As the RBF kernel works well in practice and only has one parameter, the parameter-tuning procedure is more straightforward when compared to that of the polynomial kernel function which has two parameters that need to be tuned. Also, it is well established that the RBF kernel is more powerful than the linear kernel; therefore, it is frequently used and adopted here.

$$K(\boldsymbol{x}_i', \boldsymbol{x}_k') = exp\left(-\frac{\|\boldsymbol{x}_i' - \boldsymbol{x}_k'\|^2}{2\sigma_j^2}\right) \tag{4.6}$$

where σ_j^2 represents the parameter of the RBF kernel.

The ML-BCV model is established by hybridizing MLS-SVMR and GSA as illustrated in Figure 4.2. The MLS-SVMR is adopted to learn the nonlinear mapping between the input independent and the dependent variables. In the MLS-SVMR training process, it requires eight hyper-parameters: the four regularization parameters $\gamma_1, \gamma_2, \gamma_3, \gamma_4$ that govern the penalty imposed to input data points deviating from the regression function and the four kernel parameters $\sigma_1^2, \sigma_2^2, \sigma_3^2, \sigma_4^2$ that affect the smoothness of the approximately nonlinear function. For all of these eight hyper-parameters, the lower and upper bounds are set to be 2^{-15} and 2^{15} respectively, with an interval of 2^2. This study utilizes the GSA to exhaustively and adaptively search for the most optimized set of MLS-SVMR hyper-parameters to minimize the cost function, which is the mean squared error (MSE) obtained by leave-one-out (LOO) cross-validation procedure.

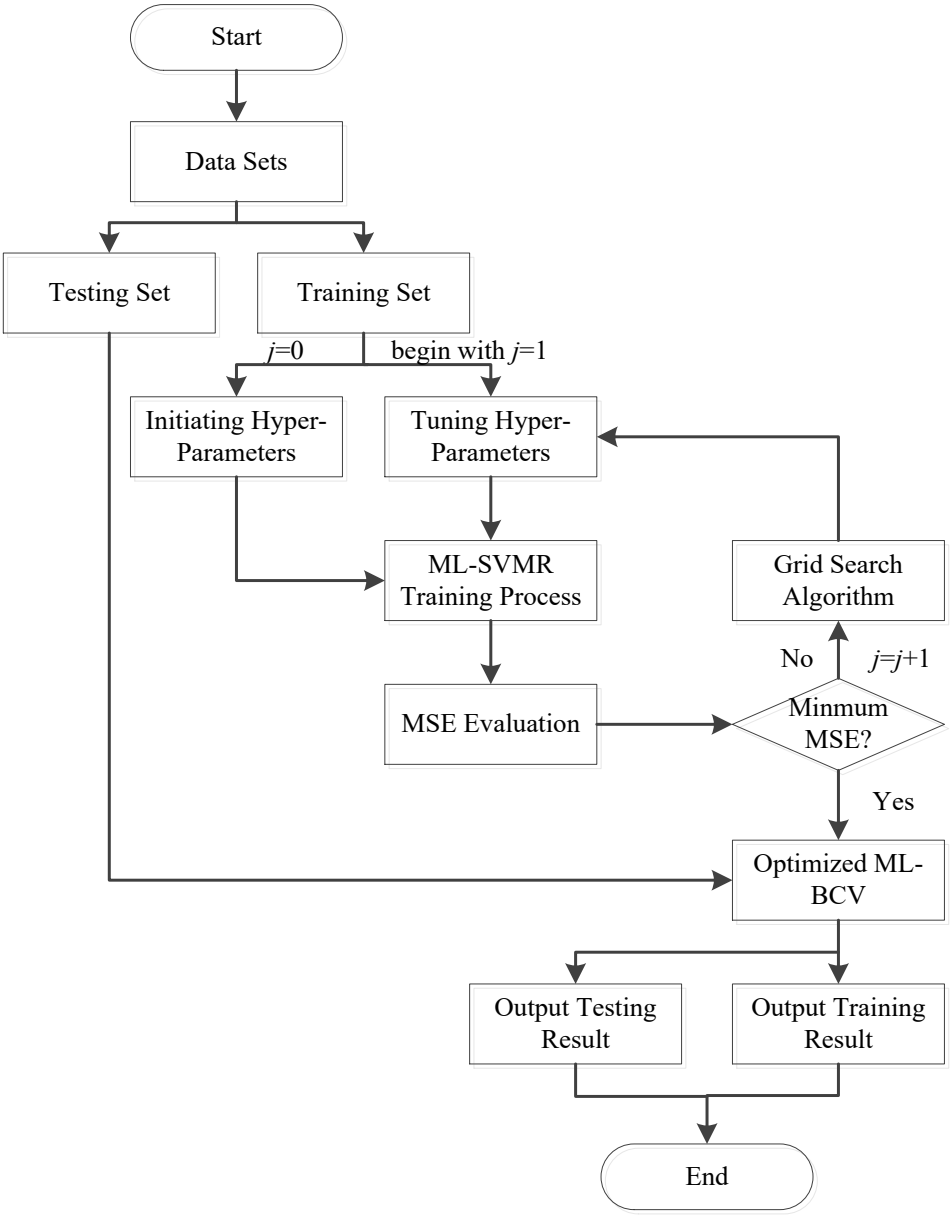


Figure 4.2 Implementation of ML-BCV.

4.2.2 Hardening behavior results

The rectangular RC column dataset presented in **Chapter III** (see Appendix A) was utilized to train, test, and validate the ML-BCV model by: (1) direct comparison with experimental results; (2) a 10-fold cross-validation procedure; and, (3) direct comparison with traditional physics-based modeling approaches for three randomly selected columns (one is flexure-critical, one is shear-

critical, and one is flexure-shear-critical). The results demonstrate the generalized performance of the proposed ML-BCV compared to traditional modeling approaches (i.e., physics-based models).

4.2.2.1 Validation set approach

This section presents the validation of the proposed ML-BCV model in predicting the hardening behavior quantified by the backbone curve of RC columns subjected to cyclic loading reversals for flexure, shear, and flexure-shear failure modes. The validation set approach presented in **Section 3.4.1** is used. First, the training and testing sets should be established. In this study, the ratio of training set to testing set is 9:1 (188 of the 208 column specimens which failed in flexure, 16 of the 18 columns which failed in shear, and 32 of the 36 specimens which failed in flexure-shear). In total, 236 of 262 columns were randomly selected for the training set, and the remaining 26 column specimens were regarded as the testing set. The training and testing results of the ML-BCV model are presented in Figure 4.3.

By observation, the training (Figure 4.3(a)) and testing (Figure 4.3(b)) results for the drift ratios at yield shear force and the training (Figure 4.3(e)) and testing (Figure 4.3(f)) results for the drift ratios at maximum shear force show that the scatter plots for drift ratios at yield and maximum shear forces in both training and testing processes closely flock together at the line of $y=x$. Further, this illustrates that the proposed approach is able to yield predicted results that agree with the actual observed values and thus, simulate both drift ratios accurately. In a similar manner, for yield and maximum shear force, the training (Figure 4.3(c) and (g)) and testing (Figure 4.3(d) and (h)) results illustrate that the proposed ML-BCV model has excellent ability to predict the yield and maximum shear forces. Both training and testing results have validated that the proposed method can reproduce experimental test data of yield and maximum shear forces of RC columns subjected to cyclic loading reversals and covering flexure, shear, and flexure-shear failure modes. The R^2

(**Section 3.4.4**) values (R^2= 0.98 for yield shear and R^2= 0.99 for maximum shear) indicate that strong correlations exist between observed and predicted yield and maximum shear forces. The statistical indicators, RMSE and R^2, for these four dependent variables in both training and testing results are summarized in Table 4.1.

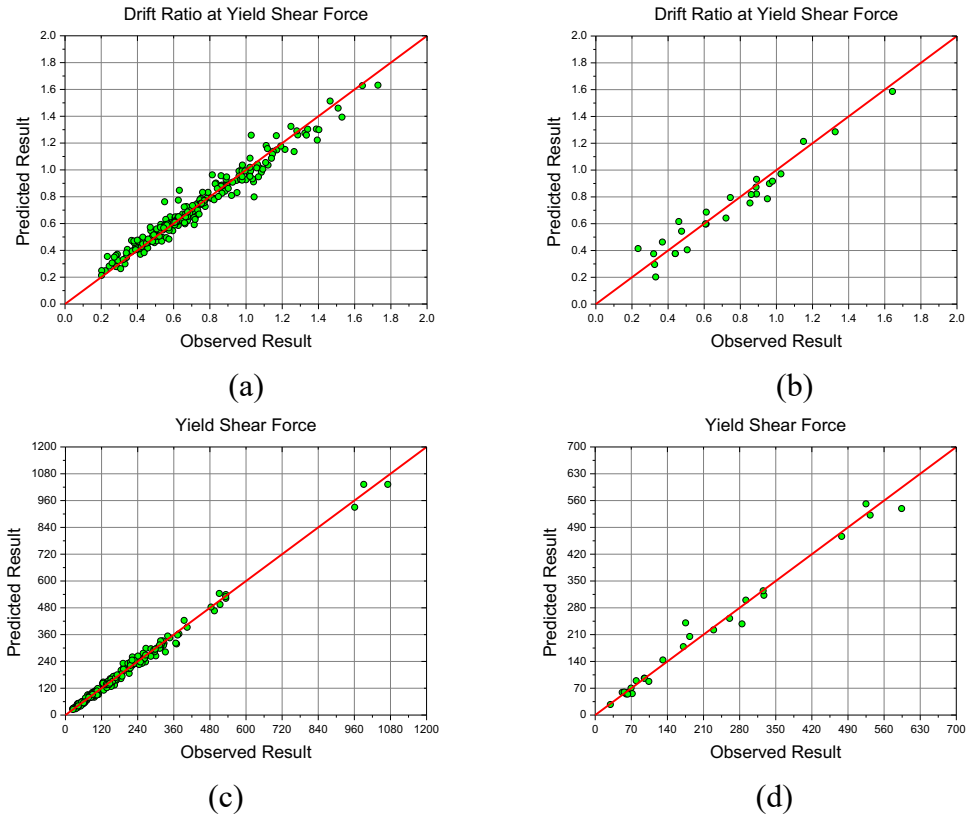


Figure 4.3 Results of training and testing the ML-BCV model: drift ratio at yield shear force for (a) training result (R^2= 0.96) and (b) testing result (R^2= 0.93); yield shear force for (c) training result (R^2= 0.99) and (d) testing result (R^2= 0.98); drift ratio at maximum shear force for (e) training result (R^2= 0.94) and (f) Testing result (R^2= 0.91); maximum shear force for (g) Training result (R^2= 1.00) and (h) Testing result (R^2= 0.99).

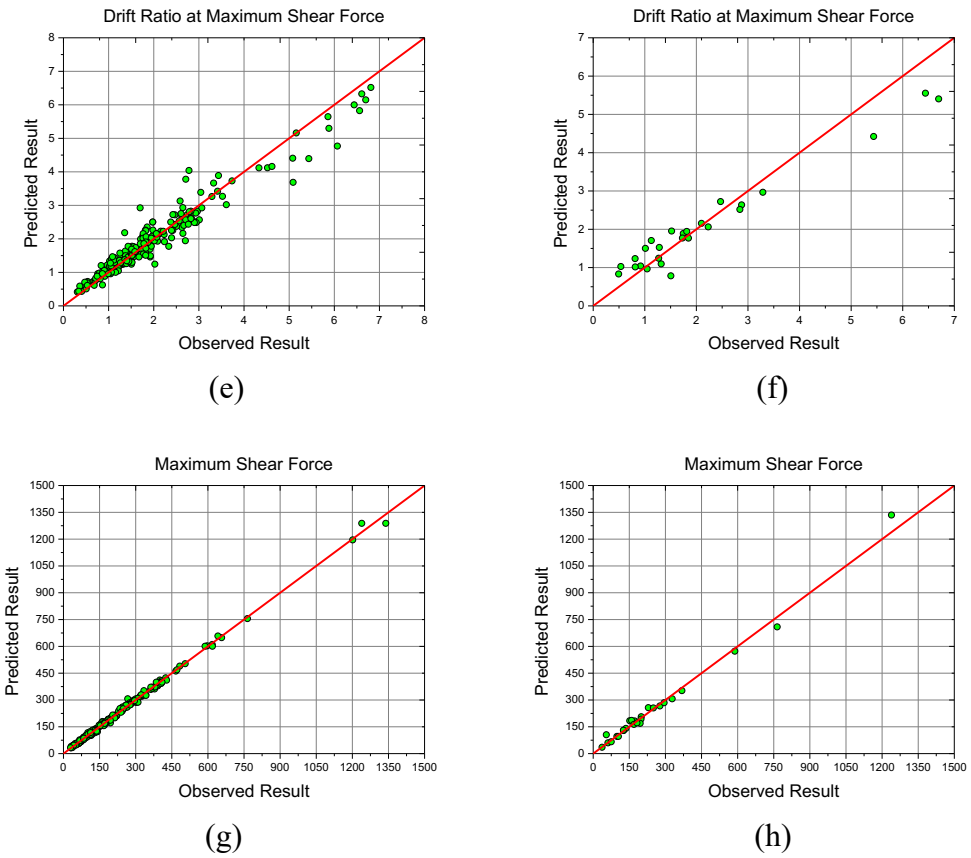


Figure 4.3 Continued.

Table 4.1 Training and testing results for the validation set approach.

Variable	Training result		Testing result	
	RMSE	R^2	RMSE	R^2
δ_y	0.06	0.96	0.08	0.93
V_y	13.25	0.99	21.74	0.98
δ_m	0.31	0.94	0.48	0.91
V_m	8.88	1	27.56	0.99

4.2.2.2 10-fold cross-validation approach

Additionally, a 10-fold cross-validation process introduced in **Section 3.4.2** is also executed to more robustly and accurately evaluate the performance of the proposed ML-BCV model. The whole database is randomly and averagely divided into 10 data subsets or "folds" where each fold, in turn, serves as a testing set. The performance of the proposed ML-BCV model can be evaluated

by averaging the results of the 10 data folds. Tables 4.2, 4.3, 4.4, and 4.5 summarize the results of the 10-fold cross-validation for drift ratio at yield shear, yield shear force, drift ratio at maximum shear, and maximum shear force, respectively. The average RMSE and R^2 of testing results for all four dependent variables are very close to those displayed in Table 4.1 for the case where the full database is employed for training and testing, illustrating that the proposed ML-BCV model is very reliable and powerful in predicting the backbone curve of RC columns subjected to cyclic loading reversals for flexure, shear, and flexure-shear failure modes.

Table 4.2 Results of the 10-fold cross-validation for drift ratio at yield shear.

Performance	Data folds										
	1	2	3	4	5	6	7	8	9	10	Mean
TrainRMSE	0.06	0.06	0.06	0.07	0.07	0.06	0.06	0.06	0.07	0.07	0.06
TrainR^2	0.96	0.96	0.96	0.95	0.95	0.96	0.95	0.96	0.95	0.95	0.96
TestRMSE	0.08	0.09	0.09	0.11	0.10	0.09	0.10	0.11	0.08	0.12	0.10
TestR^2	0.93	0.92	0.92	0.90	0.90	0.91	0.89	0.88	0.93	0.87	0.91

Table 4.3 Results of the 10-fold cross-validation for yield shear force.

Performance	Data folds										
	1	2	3	4	5	6	7	8	9	10	Mean
TrainRMSE	13.58	12.03	5.82	13.93	11.94	8.33	12.31	13.63	12.03	13.01	11.66
TrainR^2	0.99	0.99	1.00	0.99	0.99	1.00	0.99	0.99	0.99	0.99	0.99
TestRMSE	20.40	23.48	82.47	17.07	21.80	33.47	20.87	11.54	16.05	27.69	27.48
TestR^2	0.95	0.95	0.90	0.98	0.97	0.97	0.96	0.98	0.98	0.95	0.96

Table 4.4 Results of the 10-fold cross-validation for drift ratio at maximum shear.

Performance	Data folds										
	1	2	3	4	5	6	7	8	9	10	Mean
TrainRMSE	0.42	0.36	0.46	0.28	0.22	0.38	0.36	0.38	0.43	0.46	0.38
TrainR^2	0.90	0.93	0.91	0.95	0.97	0.93	0.93	0.92	0.89	0.88	0.92
TestRMSE	0.56	0.46	0.48	0.49	0.43	0.46	0.52	0.41	0.57	0.66	0.50
TestR^2	0.88	0.91	0.89	0.88	0.86	0.91	0.89	0.90	0.85	0.84	0.88

Table 4.5 Results of the 10-fold cross-validation for maximum shear force.

Performance	Data folds										
	1	2	3	4	5	6	7	8	9	10	Mean
TrainRMSE	15.07	9.25	13.41	16.37	15.74	9.66	14.09	13.93	12.65	13.56	13.37
TrainR^2	0.99	1.00	0.99	0.99	0.99	1.00	0.99	0.99	1.00	0.99	0.99
TestRMSE	21.28	39.61	32.68	27.28	21.61	37.10	27.84	19.16	26.59	15.21	26.84
TestR^2	0.97	0.92	0.98	0.98	0.98	0.98	0.99	0.98	0.90	0.98	0.97

4.2.2.3 Comparison with traditional physics-based methods

Finally, this section presents a comparison between the proposed ML-BCV model and widely-used traditional modeling approaches (i.e., distributed plasticity fiber model) to demonstrate the real-world application and full potential for this approach in practice. To validate the superiority of the proposed ML-BCV model, traditional modeling techniques were employed to simulate the hysteretic response (shear force versus lateral displacement) of RC columns with flexure-, shear-, and flexure-shear-critical modes. The classic fiber beam-column element is adopted to simulate the nonlinear cyclic response of RC columns failed in flexure. Since the classic fiber beam-column element fails to accurately reflect the nonlinear behavior of shear-critical RC columns, as illustrated in Marini and Spacone (2006), the modeling scheme proposed by Marini and Spacone (2006) is utilized to model the hysteretic force-displacement response for shear and flexure-shear critical RC columns.

Three column specimens (BG-3 from Saatcioglu and Grira (1999), 3CMD12 from Lynn (1999), and 2CLD12 from Sezen and Moehle (2002)) are randomly selected from the rectangular RC column database presented in **Chapter III** (see Appendix A). A single force-based fiber beam-column element with 5 Gauss-Lobatto integration points (i.e., monitoring sections) is employed to simulate specimen BG-3 failed in flexure. In each monitoring section, cover concrete fibers are simulated using the modified Kent and Park model (Scott et al. 1982), and the confined concrete

model proposed by Mander et al. (1988) is utilized to represent the confinement effect of the stirrups. The reinforcement fiber is modeled by the Menegotto-Pinto model (Menegotto and Pinto, 1973). For the shear- and flexure-shear critical specimens (3CMD12 and 2CLD12, respectively) the modeling strategy proposed by Marini and Spacone (2006) is used. This strategy requires an extra nonlinear V-γ constitutive law at the section level. The element, concrete, and reinforcement fibers for the two shear-critical specimens are defined in the same way as specimen BG-3. The hysteretic model proposed by Ibarra et al. (2005) is selected to represent the nonlinear shear behavior of the two shear-critical columns, and their backbone curves are defined according to the method suggested by Sezen (2008). All modeling for these three randomly selected columns has been implemented in OpenSees (Mazzoni et al. 2006).

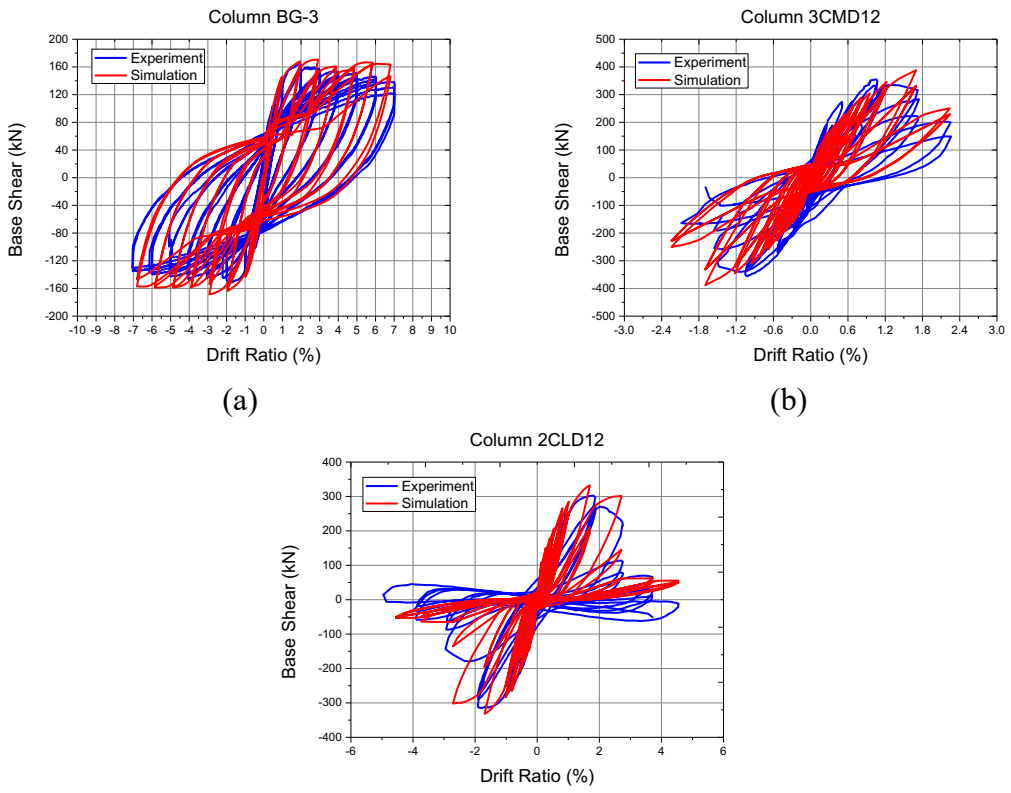


Figure 4.4 Comparison between simulated results and experimental data: (a) failure in flexure for column BG-3; (b) failure in shear for column 3CMD12; (c) failure in flexure-shear for column 2CLD12.

A comparison between the experimental data and the simulation results is presented in Figure 4.4. Figure 4.4(a) demonstrates that the simulated results, including the backbone curve and hysteretic loops, closely agree with the measured test data, illustrating that traditional modeling can accurately simulate the nonlinear response of flexure-critical columns. In contrast, Figure 4.4(b) and (c) both overestimate the column load-carrying capacity, and Figure 4.4(b) has an apparent discrepancy of initial stiffness with experimental results, while the initial stiffness in Figure 4.4(c) has good agreement with the observed data. However, both Figures 4.4(b) and 4.4(c) capture the behavior characteristics of these two shear-critical columns such as pinching of hysteretic loops, stiffness, and strength degradation.

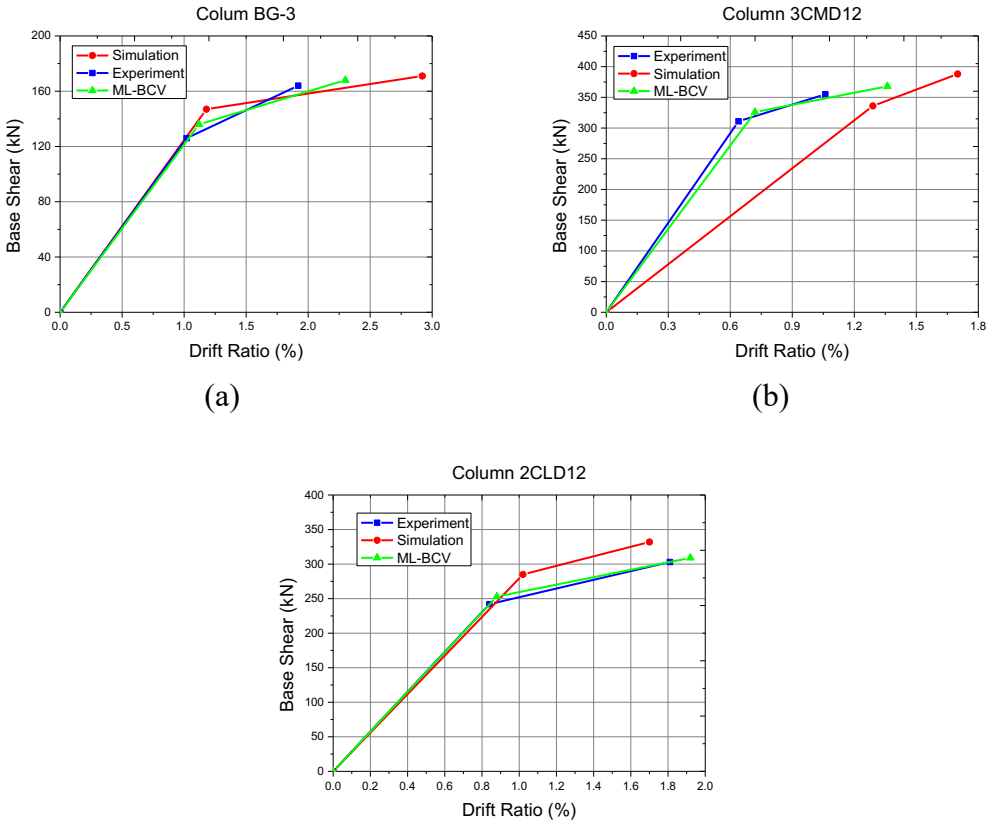


Figure 4.5 Comparison of backbone curves obtained between experiments, traditional modeling, and the proposed ML-BCV model: (a) failure in flexure for column BG-3; (b) failure in shear for column 3CMD12; (c) failure in flexure-shear for column 2CLD12.

Table 4.6 Result comparison between traditional modeling and ML-BCV.

Statistical indicators	**Traditional modeling**				**Proposed ML-BCV**			
	V_y	δ_y	V_m	δ_m	V_y	δ_y	V_m	δ_m
RMSE	31.17	0.40	25.68	0.69	12.19	0.08	8.58	0.29
MAPE (%)	14.16	46.23	7.71	39.51	5.77	9.02	2.69	18.06
R^2	0.97	0.67	0.99	0.74	1.00	0.99	1.00	0.95

The drift ratio at yield shear, yield shear force, drift ratio at maximum shear, and maximum shear force are extracted from the simulated hysteretic base shear versus drift ratio plot in order to compare the performance between the proposed ML-BCV model and traditional modeling

approaches. 259 of the 262 test columns in the database, excluding the three specimens mentioned previously, are selected as the training set, and the testing set consists of these three columns. The comparison between traditional modeling and the proposed ML-BCV model is illustrated in Figure 4.5. The results for a flexure-critical column, represented in Figure 4.5(a), show that the backbone curves obtained from both the proposed ML-BCV model and traditional modeling technique agree well with the experimental tests. However, the modeling approach overestimates the drift ratios at yield and maximum shear, whereas the ML-BCV model underestimates the drift ratio at maximum shear but overall, agrees with experimental data better than the traditional modeling approach. Figure 4.5(b) illustrates that the proposed ML-BCV model is far superior to the traditional modeling approach in predicting the backbone curve for shear-critical RC columns. Figure 4.5(c) demonstrates that the traditional modeling approach overestimates the yield and maximum shear forces of RC columns failed in flexure-shear, while the ML-BCV model yields accurate results.

To further validate the usefulness of the proposed ML-BCV model in comparison with traditional modeling approaches, the statistical indicators – mean absolute percentage error (MAPE), RMSE, and R^2 – are also adopted. The calculated metrics are provided in Table 4.6. These results show that the traditional modeling approaches are outperformed by the ML-BCV model on all accounts, where the associated RMSE, MAPE, and R^2 values are 12.19, 5.77%, and 1.00 for yield shear force, and 8.58, 2.69%, and 1.00 for maximum shear force. Thus, the ML-BCV model reduces the RMSE by roughly 61% (V_y) and 67% (V_m), reduces the MAPE by approximately 59% (V_y) and 65% (V_m), and enhances the R^2 value by approximately 3% (V_y) and 1% (V_m). Furthermore, for the predictions of drift ratio at yield and maximum shear force, the performance of the traditional modeling approach is significantly worse than that of the ML-BCV model. Notably, the proposed ML-BCV model, when compared to traditional modeling approaches, reduces the RMSE by

approximately 80% (δ_y) and 58% (δ_m), reduces the MAPE by approximately 80% (δ_y) and 55% (δ_m), and enhances the R^2 value by approximately 32% (δ_y) and 22% (δ_m). Based on these comparisons, the ML-BCV model presented in this thesis performs significantly better than that of traditional modeling approaches for both yield and maximum shear force and drift ratios and agrees well with experimental tests**.** Therefore, the novel ML-BCV model is deemed the most appropriate means for predicting the backbone curves of RC columns subjected to reversed cyclic loading across all failure modes.

4.3 Softening Behavior Prediction

In this section, a novel machine learning (ML) method is proposed to predict the softening behavior of RC columns in a generalized and accurate way. To do this, the peak response of the softening behavior represented via the drift capacity was extracted from the circular RC column database developed in **Chapter III** (see Appendix B) and is regarded as the response variable. The drift capacity of RC columns is an important indicator to quantify the seismic vulnerability of RC frame buildings; however, it is challenging to accurately predict this value as the nonlinear behavior can vary greatly by column type. Further, the variation of nonlinear behavior for flexure-, shear-, and flexure-shear-critical columns can make the circular RC column dataset presented in **Chapter III** (see Appendix B) have different data patterns in different local regions. The global ML methods, which are required to fit the full training set, may not be able to fully capture this variation under this circumstance. Instead, a local model may be more appropriate, and the performance of global models can often be improved by localizing their learning capabilities via use of locally-weighted training criteria (Bottou and Vapnik, 1992; Vapnik, 1992; Vapnik and Bottou, 1993). This means that for different regions of the input space, there will be individual models that attempt to fit only the nearby training data (points which are relevant within the specific regions), conceivably avoiding the negative influences from irrelevant training data far away from that location, which the global model cannot avoid (Atkeson et al., 1997b).

This section presents a novel, local ML model, called locally-weighted LS-SVMR (LWLS-SVMR), which integrates LS-SVMR and locally-weighted training criteria to enhance and generalize the prediction of the drift capacity of RC columns, regardless of the column type. The details of the proposed LWLS-SVMR model are presented in the remainder of this section. Additionally, in the development of the LWLS-SVMR, several hyper-parameter values need to be

tuned. To tune these values, a hybrid coupled simulated annealing (CSA) (Xavier-de-Souza et al., 2010) and grid search algorithm (GSA) (Bergstra and Bengio, 2012) is developed. The mathematical formulation of the proposed LWLS-SVMR model and the hyper-parameter tuning procedure are presented in the following.

4.3.1 Development of LWLS-SVMR

This section describes the mathematical basis for the novel LWLS-SVMR model. As previously mentioned, the LWLS-SVMR model combines the excellent nonlinear mapping capabilities of the LS-SVMR algorithm and the properties of locally-weighted learning. For the remainder of this section, the following notations are utilized. Let R be the real numbers set; $\boldsymbol{x}_i \in R^p$ is a row vector with p dimensions (i.e., p variables) which can be written as $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ip})$, and $\boldsymbol{x}'_{\mathrm{i}} \in R^p$ is a column vector with p dimensions which can be written as $\boldsymbol{x}'_i = (x_{i1}, \dots, x_{ip})^T$; $y_i \in R$ is a real number; $\boldsymbol{X} \in R^{n \times p}$ is an n × p matrix which can be written as $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^T$; the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is an $n \times (p+1)$ matrix which includes n data points and each data point contains p predictors (i.e., $\boldsymbol{x}_i \in R^p$) and one response (i.e., $y_i \in R$).

The basic procedure of the LWLS-SVMR model is as follows:

(1) Define the query point $(\boldsymbol{x}_q), q = 1, \dots, m$ (m is the total number of points to be predicted) as a point which is not included in the training set, where the corresponding response value $\hat{y}_q$ is still unknown and not considered in the query process;

(2) For each query point $(\boldsymbol{x}_q), q = 1, \dots, m$: define a subset $\{(\boldsymbol{x}_{(s)}, y_{(s)})\}_{s=1}^r$ from the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ by a parameter f_q;

where f_q can take any value in the range $(0, 1]$, the number of data points in the subset r is equivalent to $Ceil(f_q * n)$, and the points $(\boldsymbol{x}_{(s)}, y_{(s)})$, $s = 1, \dots, r$ in the subset

$\{(\boldsymbol{x}_{(s)}, y_{(s)})\}_{s=1}^{r}$, are determined and sorted by the Euclidean distance metric.

(2a) Calculate the Euclidean distance from each data point in the training set to each query point:

$$d_{qi} = \|\boldsymbol{x}_i - \boldsymbol{x}_q\|, i = 1, \dots, n; q = 1, \dots, m \tag{4.7}$$

Then, for each query point, there is a distance-vector $\boldsymbol{d}_q = (d_{q1}, \dots, d_{qn}), q = 1, \dots m$;

(2b) Sort the entries in each distance vector increasingly such that a new sorted distance vector $\boldsymbol{d}_{(q)} = (d_{(q1)}, \dots, d_{(qn)}), q = 1, \dots m$ is obtained;

(2c) Select the data points in the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, corresponding to the first r entries in the sorted distance vector $\boldsymbol{d}_{(q)}$ (i.e., $d_{(q1)}, \dots, d_{(qr)}$), as the subset $\{(\boldsymbol{x}_{(s)}, y_{(s)})\}_{s=1}^{r}$.

(3) After the subset is determined, the LS-SVMR fitting procedure (Suykens et al., 2002) is performed to calculate $\boldsymbol{w}' = (w_1, w_2, \dots, w_h)^T \in R^h$ and $b \in R$ given the subset and weights which minimize the following objective function:

$$J(\boldsymbol{w}', e_s) = \frac{1}{2}(\boldsymbol{w}')^T\boldsymbol{w}' + \frac{1}{2}\gamma_q \sum_{s=1}^{r} \beta_q(\boldsymbol{x}_{(s)}) e_s^2, q = 1, \dots, m \tag{4.8}$$

Subject to: $y_{(s)} = (\boldsymbol{w}')^T \varphi(\boldsymbol{x}'_{(s)}) + b + e_s, s = 1, \dots, r$ (4.9)

where $e_s \in R, s = 1, \dots, r$ is the error variable; $\gamma_q \in R, q = 1, \dots, m$ is a regularization parameter; $\beta_q(\boldsymbol{x}_{(s)}) \in R, s = 1, \dots, r, q = 1, \dots, m$ is a weight that can take any value in the range $[\varepsilon, 1]$ used to determine the level of contribution from data points in a subset around the query point; $\varepsilon \in R$ is a real number approaching 0; $\varphi(\boldsymbol{x}'_{(s)})$ is a feature vector; and $\varphi(\cdot)$: $R^p \rightarrow R^h$ is a mapping function from p dimensions to a higher h-dimensional feature space. Note: $\boldsymbol{x}'_{(s)}$ is a column vector; thus $\varphi(\boldsymbol{x}'_{(s)})$ is also a column vector.

If $\beta_q(\boldsymbol{x}_{(s)})$ takes a value approaching ε, the point $(\boldsymbol{x}_{(s)}, y_{(s)})$ is far away from the query point $(\boldsymbol{x}_q, \hat{y}_q)$ (relatively large Euclidean distance) and plays a lesser role in the determination of $\hat{y}_q$; while, if $\beta_q(\boldsymbol{x}_{(s)})$ takes a value approaching one, the point is close to the query point (relatively small Euclidean distance) and plays a large role in the determination of $\hat{y}_q$.

(3a) The Lagrangian function is established to solve Eq. (4.8) and Eq. (4.9):

$$L(\boldsymbol{w}', b, e_s; \alpha_s) = J(\boldsymbol{w}', e_s) - \sum_{s=1}^{r} \alpha_s \left((\boldsymbol{w}')^T \varphi(\boldsymbol{x}'_{(s)}) + b + e_s - y_{(s)} \right) \tag{4.10}$$

where $\alpha_s \in R, s = 1, \ldots, r$ is a Lagrange multiplier.

(3b) The Karush-Kuhn-Tucker (KKT) conditions for optimality are used by differentiating the variables in (4.10), which results in the following:

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{w}'} = 0 \rightarrow \boldsymbol{w}' = \displaystyle\sum_{s=1}^{r} \alpha_s \varphi(\boldsymbol{x}'_{(s)}) \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow 0 = \displaystyle\sum_{s=1}^{r} \alpha_s \\ \dfrac{\partial L}{\partial e_s} = 0 \rightarrow e_s = \dfrac{\alpha_s}{\gamma_q \beta_q(\boldsymbol{x}_{(s)})}, \; s = 1, \ldots, r; \; q = 1, \ldots, m \\ \dfrac{\partial L}{\partial \alpha_s} = 0 \rightarrow y_{(s)} = (\boldsymbol{w}')^T \varphi(\boldsymbol{x}'_{(s)}) + b + e_s, \; s = 1, \ldots, r \end{cases} \tag{4.11}$$

(3c) Rearranging (4.11) and eliminating $\boldsymbol{w}'$ and e_s, the following matrix equation can be obtained:

$$\begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & K(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(1)}) + \dfrac{1}{\gamma_q \beta_q(\boldsymbol{x}_{(1)})} & K(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(2)}) & \cdots & K(\boldsymbol{x}_{(1)}, \boldsymbol{x}_{(r)}) \\ 1 & K(\boldsymbol{x}_{(2)}, \boldsymbol{x}_{(1)}) & K(\boldsymbol{x}_{(2)}, \boldsymbol{x}_{(2)}) + \dfrac{1}{\gamma_q \beta_q(\boldsymbol{x}_{(2)})} & \cdots & K(\boldsymbol{x}_{(2)}, \boldsymbol{x}_{(r)}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & K(\boldsymbol{x}_{(r)}, \boldsymbol{x}_{(1)}) & K(\boldsymbol{x}_{(r)}, \boldsymbol{x}_{(2)}) & \cdots & K(\boldsymbol{x}_{(r)}, \boldsymbol{x}_{(r)}) + \dfrac{1}{\gamma_q \beta_q(\boldsymbol{x}_{(r)})} \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_r \end{bmatrix} = \begin{bmatrix} 0 \\ y_{(1)} \\ y_{(2)} \\ \vdots \\ y_{(r)} \end{bmatrix} \tag{4.12}$$

where $q = 1, \dots, m$ and the kernel function is: $K(\boldsymbol{x}_{(s)}, \boldsymbol{x}_{(t)}) = \varphi^T(\boldsymbol{x}'_{(s)})\varphi(\boldsymbol{x}'_{(t)}), s = 1, \dots, r; t = 1, \dots, r$

(3d) For the determination of $\beta_q(\boldsymbol{x}_{(s)}) \in R, s = 1, \dots, r; q = 1, \dots, m$, for each query point $\boldsymbol{x}_q$, let $d_{(qr)}$ be the distance from $\boldsymbol{x}_q$ to the r^{th} nearest neighbor $\boldsymbol{x}_{(r)}$ (i.e., $d_{(qr)}$ is the maximum distance compared to $d_{(q1)}, \dots, d_{(q(r-1))}$), and let: $\beta_q(\boldsymbol{x}_{(s)}) = T(d_{(qr)}^{-1}\|\boldsymbol{x}_{(s)} - \boldsymbol{x}_q\|)$, where $T(\cdot)$ is a tricube weight function, which is defined as the following:

$$T(g) = f(x) = \begin{cases} (1 - |g|^3)^3, & |g| < 1 \\ \varepsilon, & |g| \geq 1 \end{cases} \tag{4.13}$$

where ε can take any value close to zero, which in this case is equal to $10e - 4$ to avoid a zero in the denominator in (4.12).

(4) Solving (4.12), the Lagrange multiplier $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_r)$ and b are obtained which can then be utilized to predict the query point $\boldsymbol{x}_q$ using the following:

$$y(\boldsymbol{x}_q) = \sum_{s=1}^{r} \alpha_s K(\boldsymbol{x}_q, \boldsymbol{x}_{(s)}) + b \tag{4.14}$$

The final general form of the drift capacity model presented in this work is given by (4.14) for an individual query point. The approach developed in this section implements the RBF kernel function, which is given in Eq. (4.6).

4.3.2 Hybrid optimization algorithm

As mentioned previously, there are three hyper-parameters, f_q, γ_q, and σ_q^2 that need to be accurately defined in the training process as they can significantly affect the accuracy level of the predicted results. Proper tuning of these parameters is necessary to optimize the proposed LWLS-SVMR. The parameter f_q is employed to establish the optimum subset size and the weights for the data points in the subset that yield the best prediction for each query point $\boldsymbol{x}_q$. The regularization parameter γ_q controls the penalty inflicted on the data points diverging from the regression function in the subset, and the kernel parameter σ_q^2 governs the smoothness of the approximately nonlinear function. If all three hyper-parameters are suitably selected, it can guarantee high predictive capabilities and good generalization performance. Grid search algorithm (GSA) (Bergstra and Bengio, 2012) is an effective optimization technique that has been very widely used to determine hyper-parameter values. However, GSA requires a range of space for the hyper-parameters to be defined manually. In the proposed model, the range of each hyper-parameter is known: f_q is within $(0, 1]$, γ_q is within $(0, \infty)$, and σ_q^2 is within $(0, \infty)$.

Theoretically, broader ranges for γ_q and σ_q^2, and narrower intervals in the ranges of all three hyper-parameters will result in the most optimum parameter pairs; however, the computational cost will also increase with an increase in range and number of intervals. A much more efficient alternative is to first determine starting values for f_q, γ_q, and σ_q^2 that are close to the optimum pairs $\left(f_{qo}, \gamma_{qo}, \sigma_{qo}^2\right)$. Then, a refined search for the optimum pairs $\left(f_{qo}, \gamma_{qo}, \sigma_{qo}^2\right)$ can be performed using GS by setting an appropriate region encompassing the starting values and defining a stop criterion for the iteration. As the range of f_q is already determined and does not contain an infinite number, it is more straightforward to separate f_q from γ_q and σ_q^2 during the optimization of these three hyper-parameters.

A global optimization algorithm called coupled simulated annealing (CSA) (Xavier-de-Souza et al., 2010) is employed to determine the starting values. CSA was originally proposed to reduce the sensitivity to initialization parameters. This is done by means of an acceptance temperature which regulates the variance of the associated probabilities while guiding the optimization process to quasi-optimal runs. Ultimately, this results in considerably higher optimization efficiency than alternate global optimization algorithms (Xavier-de-Souza et al., 2010).

The hyperparameter tuning procedure using the hybrid CSA-GSA algorithm, and as implemented within the LWLS-SVMR algorithm, is formulated as the following:

(1) For each query point $\boldsymbol{x}_q, q = 1, \dots, m$, discretize the parameter space with interval a_q for f_q (i.e., f_q takes the values $a_q, 2a_q, \dots, 1$);

(2) For each $f_{qi}, i = 1, \dots, length(f_q)$, given the query point $\boldsymbol{x}_q$, determine a subset and weights for the corresponding subset data points;

(3) Perform the CSA procedure in the subset to tune γ_q and σ_q^2. Using leave-one-out cross-validation, determine the starting values $(\gamma_{qs}, \sigma_{qs}^2)$ as those with minimum mean squared errors (MSE);

(4) Given the starting values $(\gamma_{qs}, \sigma_{qs}^2)$, set an appropriate region encompassing $(\gamma_{qs}, \sigma_{qs}^2)$ and the stop criterion, and use GS to exhaustively search for the optimum pair $(\gamma_{qo}, \sigma_{qo}^2)$ that can continue to decrease the minimum MSE obtained in (3) (it is possible that the starting values are the optimum pair);

(5) Store $\{f_{qi}\}_{i=1}^{length(f_q)}$, $\{(MSE)_i\}_{i=1}^{length(f_q)}$, and optimum pairs $\{(\gamma_{qo}, \sigma_{qo}^2)_i\}_{i=1}^{length(f_q)}$, and extract the optimum pair $(f_{qo}, \gamma_{qo}, \sigma_{qo}^2)$ that has the minimum MSE among them.

The detailed implementation procedure is presented in Figure 4.6.

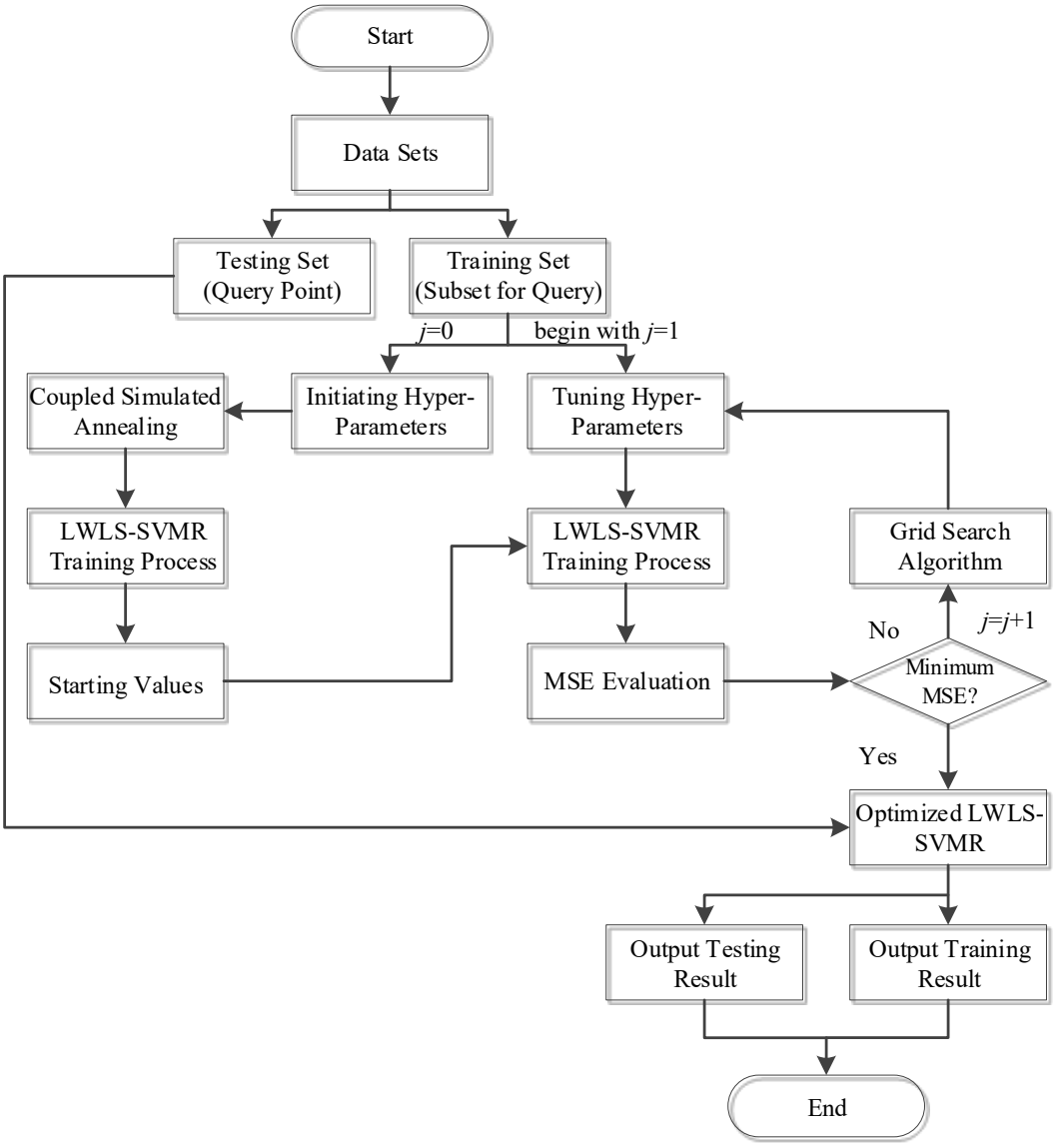


Figure 4.6 Implementation of the proposed LWLS-SVMR

4.3.3 Softening behavior results

The circular RC column database presented in **Chapter III** (see Appendix B) covering flexure-, shear-, and flexure-shear-critical specimens was used to train and test the LWLS-SVMR. The proposed LWLS-SVMR was validated by comparison with LS-SVMR, a popular local learning

approach (locally weighted quadratic regression (LWQR)), and a suitable, traditional empirical equation (Elwood and Moehle 2005).

4.3.3.1 Validation set approach

To use the validation set approach as introduced in **Section 3.4.1**, the exclusive training and testing sets need to first be defined. 112 specimens (70% of the total dataset) are randomly selected from the circular RC column database presented in **Chapter III** (see Appendix B) to form the training set, and the remaining 48 specimens (30%) comprise the testing set. The LS-SVMR, LWQR, and the proposed LWLS-SVMR models are created by fitting the training set based on the LOO cross-validation procedure to avoid overfitting. However, the LWQR and proposed LWLS-SVMR just need to fit subsets of the training set to answer query points in both training and testing sets, while the LS-SVMR needs to fit the full training set to predict the same query points in both the training and testing sets. The training and testing results of the LS-SVMR, LWQR, and proposed LWLS-SVMR models are presented in Figure 4.7 (a, b, and c, respectively). The detailed statistical indicators (R^2, RMSE, and MAPE) for LS-SVMR, LWQR, and the proposed LWLS-SVMR in both training and testing results are summarized in Table 4.7. By inspection, the training and testing results predicted by the proposed LWLS-SVMR show the best agreement with the experimentally observed results. In comparison to the testing results obtained by the LS-SVMR and LWQR, the proposed LWLS-SVMR increases the R^2 by roughly 20% and 33%, respectively, decreases RMSE values by approximately 32% and 53%, respectively, and decreases and MAPE values by approximately 39% and 54%, respectively. From this initial direct comparison of these three ML models, the proposed LWLS-SVMR outperforms both the popular global and local models.

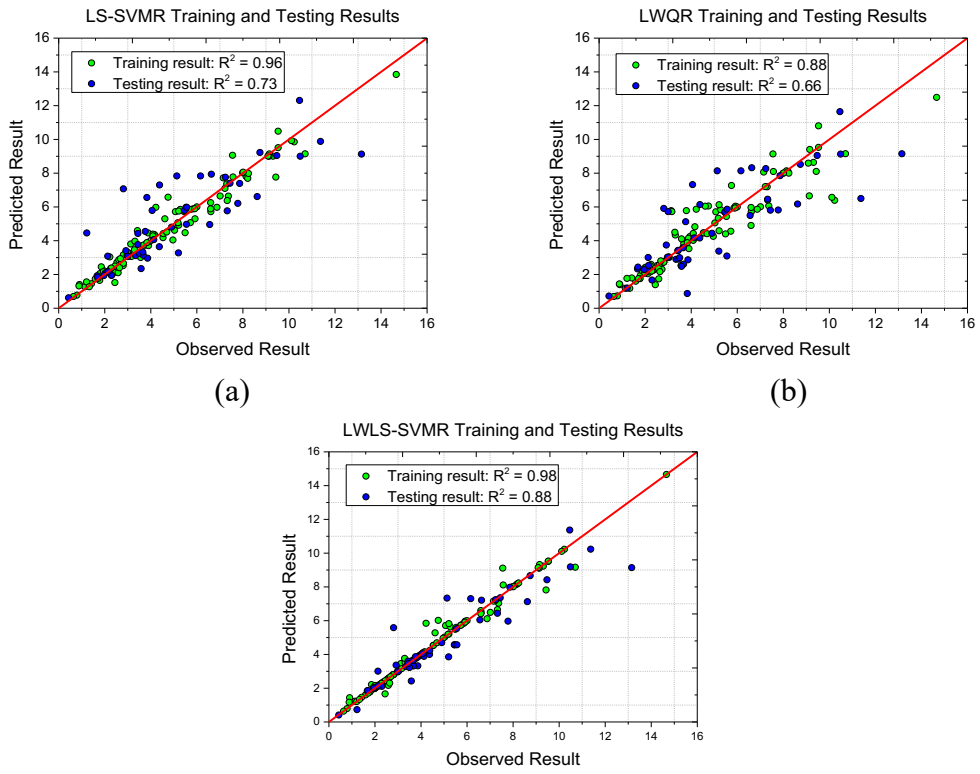


Figure 4.7 Training and testing results of LS-SVMR, LWQR, and proposed LWLS-SVMR.

Table 4.7 Comparison of training and testing results for LS-SVMR, LWQR, and LWLS-SVMR.

Models	Training Result			Testing Result		
	R^2	RMSE (%)	MAPE (%)	R^2	RMSE (%)	MAPE (%)
LS-SVMR	0.96	0.54	9.59	0.73	1.49	26.60
LWQR	0.88	0.93	13.26	0.66	1.66	27.20
LWLS-SVMR	0.98	0.39	4.19	0.88	1.01	12.42

4.3.3.2 10-fold cross-validation approach

A 10-fold cross-validation process presented in **Section 3.4.2** is also employed to alleviate the inherent randomness in selecting training and testing samples when using the validation set approach. In turn, the 10-fold cross-validation procedure will result in increasingly robust results. The fitting procedure for the three models is the same as that introduced in the previous section

for the validation set approach; however, as the division of training and testing sets is different, the model fitting procedure will be performed on each of the 10 different training sets individually. Then, each of these models is used to predict for the data in the corresponding testing set, such that all of the points in the dataset are tested. The performance of the LS-SVMR, LWQR, and proposed LWLS-SVMR models can be evaluated via averaging the results of the 10 data folds. Figure 4.8 present the testing results of the 10-fold cross-validation procedure in terms of R^2, RMSE, and MAPE for LS-SVMR, LWQR, and the proposed LWLS-SVMR model. The average R^2, RMSE and MAPE metrics for the testing results associated with all the ML models are summarized in Figure 4.8. As shown in Figure 4.8, the average of the 10-fold cross-validation results for all three ML models are comparable to the testing results obtained previously via the validation set approach, maintaining proof that the proposed LWLS-SVMR can reliably perform better than LS-SVMR and LWQR models.

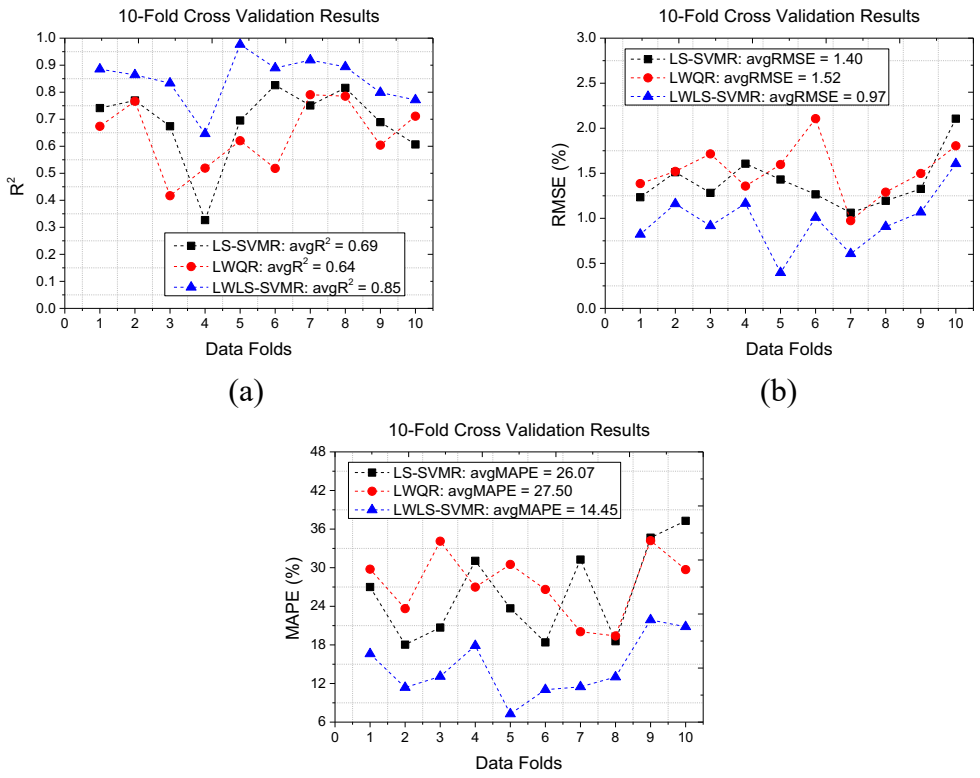


Figure 4.8 Results of 10-fold cross-validation using LS-SVMR, LWQR, and proposed LWLS-SVMR in terms of (a) R^2, (b) RMSE, and (c) MAPE.

4.3.3.3 Comparison with physics-based methods

Finally, this section presents a comparison between LS-SVMR, LWQR, LWLS-SVMR and the widely-used empirical model (Elwood and Moehle 2005) for predicting the drift capacity of RC columns using a leave-one-out (LOO) cross-validation procedure presented in **Section 3.4.3**. This comparison demonstrates the real-world application and full potential for the novel LWLS-SVMR model within the civil engineering realm and with respect to drift capacity prediction practices in general. The predicted results using the LOO procedure are partitioned according to the column failure mode. In addition to the quantification metrics used above (i.e., R^2, RMSE, and MAPE), the mean and coefficient of variation (CV) of predicted to observed results are also employed in

this comparison as these two qualification metrics are commonly used to measure the predictive performance of an empirical equation in civil engineering.

A comparison between the predicted and observed results for the LS-SVMR, LWQR, LWLS-SVMR, and empirical model developed by Elwood and Moehle (2005) is presented in Figure 4.9. The detailed statistical metrics are summarized in Table 4.8. From Table 4.8, it is evident that the Elwood and Moehle (2005) model both over- and under-estimates the drift capacity of RC flexure-critical columns. This leads to the highest CV of predicted to observed results, negative R^2 values, the highest RMSE and MAPE values, and the highest mean of predicted to observed results (Table 4.8). In comparison to the poor prediction capability of the empirical model, Table 4.8 illustrates that the three ML models all show significant improvement in accurately predicting the drift capacity of RC flexure-critical columns. The proposed LWLS-SVMR performs best among these three ML models with the highest R^2 and lowest RMSE, MAPE, mean, and CV of predicted to observed results (0.81, 1.16%, 14.61%, 1.04, and 0.24, respectively). The global LS-SVMR model performs better than the local LWQR model in terms of R^2, RMSE, MAPE, and CV of predicted to observed results. The proposed LWLS-SVMR model exhibits an increase in the R^2 value by approximately 33% when compared to the global LS-SVMR model and 62% when compared to LWQR. Comparably, RMSE and MAPE values are decreased by roughly 30% and 45% (as compared to the global LS-SVMR) and 38% and 48% (as compared to the local LWQR), respectively.

The empirical equation performs even worse for flexure-shear-critical columns and overestimates the drift capacity for nearly every flexure-shear-critical specimen. In contrast, the performance is improved for all three AI models. Still, the proposed LWLS-SVMR performs best among all models and the global LS-SVMR exhibits better performance than the local LWQR

approach. In tune with the results for flexure-critical columns, the proposed LWLS-SVMR experiences an increase in the R^2 value by roughly 10% in comparison to the global LS-SVMR and 23% for the local LWQR, and demonstrates a decrease in RMSE and MAPE values by roughly 19% and 34% for the global LS-SVMR and 30% and 35% for the local LWQR, respectively.

Finally, for RC shear-critical columns, all of the models perform worse when compared accordingly for both RC flexure- and flexure-shear-critical columns. Consistent with the previously presented results, the empirical equation performs the worst, overestimating the drift capacity for all the shear-critical columns in the dataset. This may be attributed to the fact that the good predictors for shear-critical columns may not be the same as those for the other two types of RC columns. Moreover, the shear-critical columns may negatively influence the global LS-SVMR model in the training process. The local quadratic function may not be able to reasonably represent the complex nonlinearity of shear-critical columns. This leads to the failure to capture the complex nonlinear relationship between predictors and response. The global LS-SVMR model performs better than the local LWQR model, but both of these two AI models perform poorly in predicting the drift capacity of RC shear-critical columns. The proposed LWLS-SVMR still performs best among all approaches as it can avoid the negative interference from the other two types columns in the training process and can reasonably represent the complex nonlinear relationship for shear-critical columns. This is exhibited via an increase in the R^2 values by roughly 89% in comparison to the global LS-SVMR model and 313% in comparison to the local LWQR model. Also, an ultimate decrease in the RMSE and MAPE values by roughly 28% and 37%, respectively for the global LS-SVMR and 36% and 46%, respectively for the local LWQR is achieved.

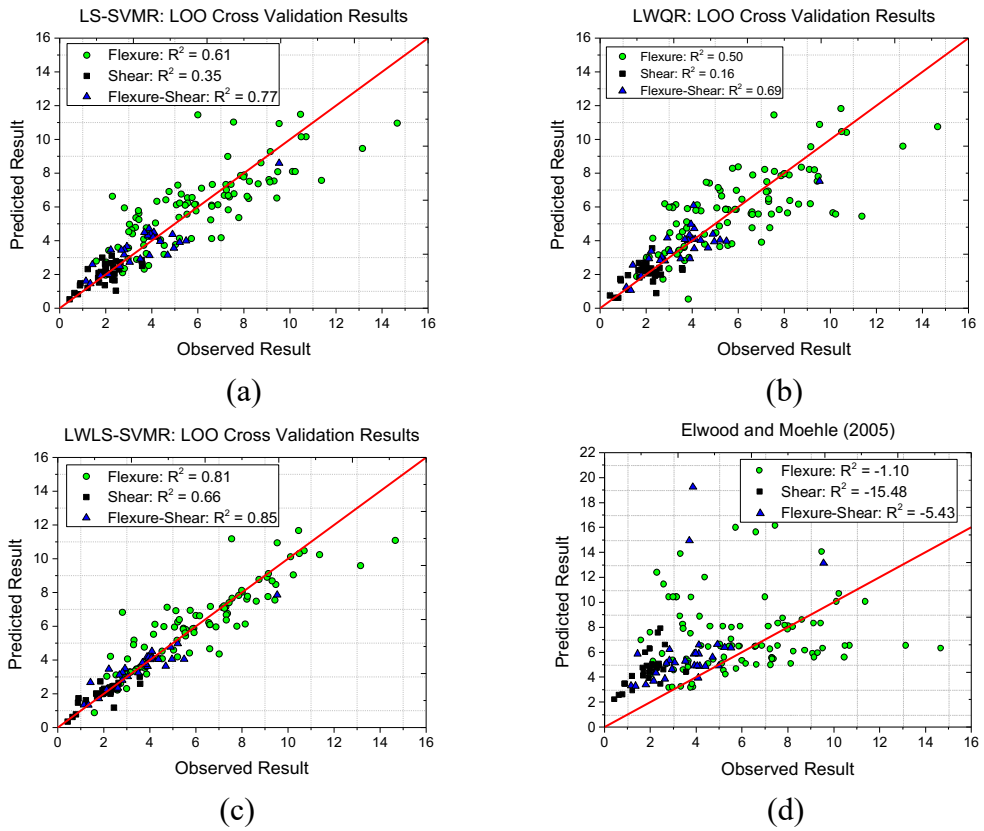


Figure 4.9 Results comparison between LS-SVMR, LWQR, LWLS-SVMR, and empirical model using LOO cross-validation procedure.

The generalization performance of the proposed AI model is measured by combining the predictions for all three types of RC columns. Notably, the proposed LWLS-SVMR model performs the best among the three AI models and the empirical equation. This is because it produces the highest R^2 value and lowest RMSE and MAPE values (0.88, 0.96%, and 14.58%, respectively). Further, it yields a value for the mean of predicted to observed results which is closer to 1 than all other models and has the smallest CV (1.04 and 0.23) (Table 4.8). In addition, the global LS-SVMR model performs better than the local LWQR model, but the performance of all the AI models developed in this work is noticeably better than the empirical equation in terms of all represented qualification metrics.

Table 4.8 Results comparison between all models using the LOO cross-validation procedure.

Models	Failure modes	RMSE (%)	MAPE (%)	R^2	Mean	CV
LS-SVMR	Flexure	1.66	26.78	0.61	1.07	0.36
	Shear	0.58	25.52	0.35	1.09	0.29
	Flexure-shear	0.77	19.53	0.77	1.06	0.23
	Combined	1.37	25.17	0.75	1.07	0.33
LWQR	Flexure	1.87	28.23	0.50	1.06	0.37
	Shear	0.66	30.01	0.16	1.11	0.33
	Flexure-shear	0.88	19.78	0.69	1.06	0.24
	Combined	1.54	27.00	0.69	1.07	0.34
LWLS-SVMR	Flexure	1.16	14.61	0.81	1.04	0.24
	Shear	0.42	16.12	0.66	1.07	0.24
	Flexure-shear	0.62	12.81	0.85	1.04	0.21
	Combined	0.96	14.58	0.88	1.04	0.23
Elwood and Moehle (2005)	Flexure	3.85	65.62	-1.10	1.49	0.66
	Shear	2.92	164.78	-15.48	2.65	0.32
	Flexure-shear	4.02	87.19	-5.43	1.87	0.52
	Combined	3.72	89.50	-0.82	1.79	0.59

4.4 Seismic Response History Prediction

In this section, a novel hybrid ML-physics based data-driven framework – where an ML method is used to directly link the experimental data to nonlinear properties of a target component, and a physical model that meets universal laws is used to perform the seismic analysis – is proposed to predict the seismic response history of RC structural components in an efficient, generalized, and accurate way. For the physics-based model, the distributed plasticity approach is widely used to predict the seismic response of RC structural components and can accurately reflect the physical behavior of RC ductile components. However, it is computationally expensive and cannot reasonably capture the physical behavior of RC non-ductile components. This section introduces a novel data-driven framework that constructs a hybrid-ML-physics based computational procedure for generalized seismic response history prediction of both RC components in a more accurate and efficient way. The proposed framework is applied for RC flexure-critical, shear-critical, and flexure-shear-critical columns under cyclic loads as well as of a full-scale RC bridge column subjected to six consecutive ground motions. The generalized prediction capability and computational efficiency of the proposed framework is validated by comparison with physics-based modeling approaches based on the experimental data.

4.4.1 Component-level data-driven framework

This section presents the novel data-driven framework to predict the hysteretic behavior and time-history response quantities of target structural components subjected to both quasi-static cyclic loading and ground motions in a more generalized, accurate, and efficient way. Based on the results of response quantities, the engineering demand parameters (EDPs) of interest can be extracted. The approach herein utilizes the proposed, novel hybrid ML-physics-based methodology, where ML is used to directly link the experimental data with the nonlinear properties of target structural

components and the physical model is used to perform the seismic analysis, making full use of the advantages of both techniques.

The framework consists of six components, as shown in Figure 4.10.

1. The first step is to collect the physical experimental data of the target structural component subjected to quasi-static cyclic loading. The experimental data should include the structural features and the corresponding observed force-displacement data. The structural features are those associated with design details that can define the target structural component, such as the structural geometry and material properties. The good predictors $\boldsymbol{x}$, where $\boldsymbol{x} \in R^p$, can then be identified from the structural features (specific details regarding the selection procedure will be introduced later). The force-displacement data denoted as $(\boldsymbol{\delta}, \boldsymbol{F})$, where $\boldsymbol{\delta} \in R^{n_p}$ is the displacement vector containing n_p applied displacements and $\boldsymbol{F} \in R^{n_p}$ is the vector of force values measured experimentally at the corresponding displacements $\boldsymbol{\delta}$, is observed or measured via the quasi-static cyclic loading test to construct the hysteretic behavior (i.e., force-displacement hysteretic relationship) of the target structural component.
2. The second step is to choose an appropriate physical model (i.e., hysteretic model) that can represent the hysteretic behavior of the target structural component subjected to cyclic load reversals (hardening and softening behavior, stiffness and strength deterioration, and pinching behavior). Here, it is important to identify the critical parameters denoted as $\boldsymbol{\Theta}$, where $\boldsymbol{\Theta} \in R^{n_\theta}$ is the parameter vector containing n_θ critical parameters, which define the selected hysteretic model. The function of each critical parameter should be clearly identified (e.g., which critical parameters control the shape of the backbone curve and the hysteretic loop). Then, the estimated forces $\widehat{\boldsymbol{F}}$ from the

selected hysteretic model can be expressed as a function of critical parameters $\boldsymbol{\Theta}$ and applied displacements $\boldsymbol{\delta}$, which is denoted as $\widehat{\boldsymbol{F}} = f(\boldsymbol{\delta}; \boldsymbol{\Theta})$ where $f(\cdot)$ represents the selected hysteretic model.

3. The third step is to calibrate the selected hysteretic model with collected force-displacement data $(\boldsymbol{\delta}, \boldsymbol{F})$ by tuning the critical parameters $\boldsymbol{\Theta}$ using a selected optimization algorithm. The parameter tuning procedure stops once the calibrated hysteretic model can perfectly reproduce the hysteretic behavior exhibited by the experimental data $(\boldsymbol{\delta}, \boldsymbol{F})$. The optimal critical parameters $\boldsymbol{\Theta}$, which are denoted as the response variables $\boldsymbol{y}$, will then be recorded.

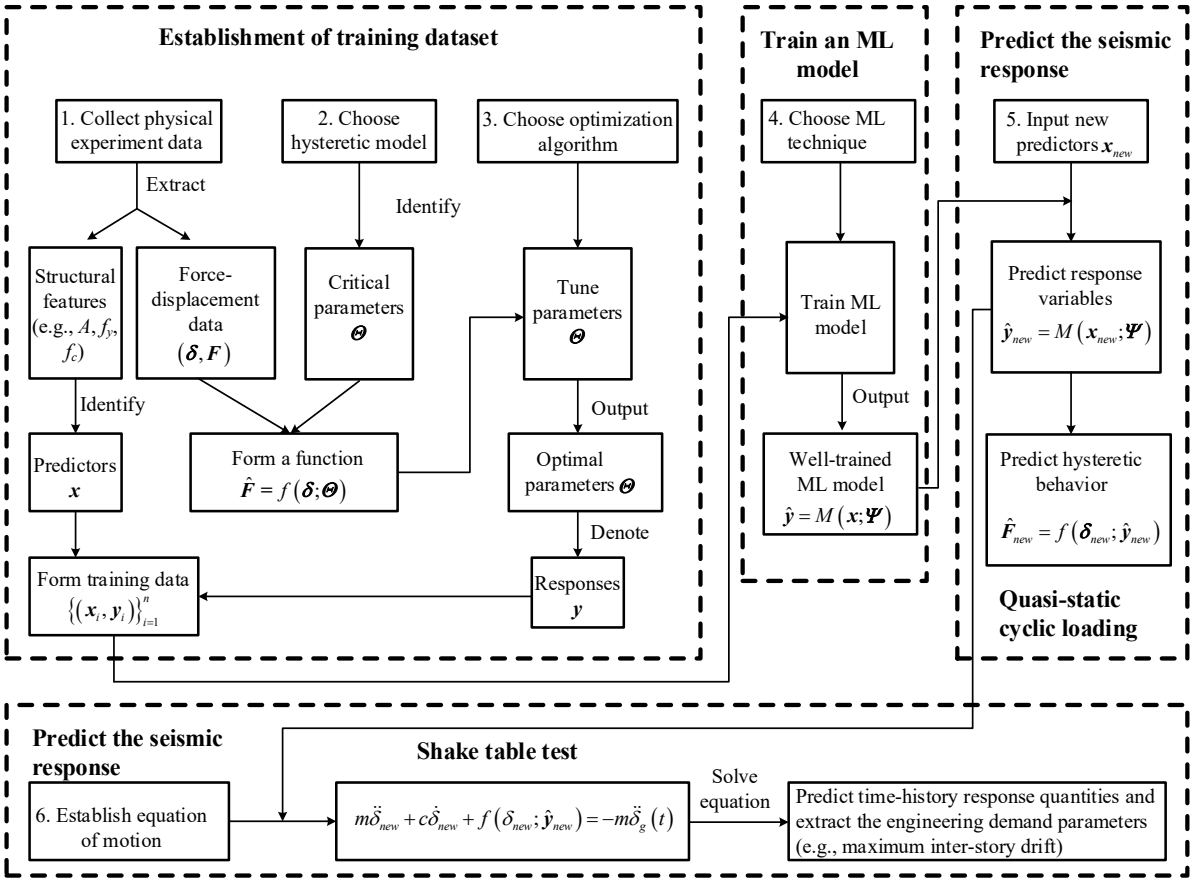


Figure 4.10 Data-driven framework for predicting seismic response of structural components under quasi-static cyclic loading and shake table tests.

Steps 1 through 3 will be performed *n* times if there are *n* target structural component specimens collected, such that a training set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i \in R^p$ and $\boldsymbol{y}_i \in R^{n_\theta}$, can be formed. In the training set, $\boldsymbol{x}_i$ represents the predictors and $\boldsymbol{y}_i$ represents the optimal critical parameters $\boldsymbol{\theta}$ that signify the response variables governing the hysteretic behavior of the target RC structural component.

4. the fourth step involves utilizing the training set to construct an ML model based on a k-fold cross-validation procedure. The well-trained ML model is denoted as $\widehat{\boldsymbol{y}} = M(\boldsymbol{x}; \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} \in R^{n_\varphi}$ is the optimal ML model parameter vector containing n_φ parameters, and $M(\cdot)$ represents the selected ML technique.
5. Steps 5 and 6 are to use the well-trained ML model to predict the hysteretic behavior and time-history response quantities of a new target structural component that is not covered in the training set under cyclic loading and ground motions, respectively. Specifically, in Step 5, the new target structural component is featured by predictors $\boldsymbol{x}_{new}$ (note that the predictors are known information). The component's hysteretic behavior can be predicted by first inputting $\boldsymbol{x}_{new}$ into the well-trained ML model to obtain the corresponding responses $\widehat{\boldsymbol{y}}_{new} = M(\boldsymbol{x}_{new}; \boldsymbol{\Psi})$. Then, the predicted responses $\widehat{\boldsymbol{y}}_{new}$ can be input into the selected hysteretic model to obtain the estimated force $\widehat{\boldsymbol{F}}_{new} = f(\boldsymbol{\delta}_{new}; \widehat{\boldsymbol{y}}_{new})$ at the applied displacements $\boldsymbol{\delta}_{new}$ under the displacement-controlled quasi-static cyclic loading (note that the applied displacements $\boldsymbol{\delta}_{new}$ in the displacement-controlled quasi-static cyclic loading test are known information). Finally, the hysteretic behavior of the new target structural component is predicted by the force-displacement relationship exhibited by the obtained data $(\boldsymbol{\delta}_{new}, \widehat{\boldsymbol{F}}_{new})$.

6. In Step 6, the time-history response quantities of the new target structural component subjected to a ground motion $\ddot{\delta}_g(t)$ can be predicted by establishing and solving the equation of motion, as shown in Figure 4.10. The most important part in Step 6 is to incorporate the predicted hysteretic behavior (e.g., tangent stiffness or resisting force) from Step 5 into the equation of motion. After all the time histories are completed, the EDPs of interest (e.g., maximum inter-story drift) can be extracted from the time-history analysis results. The proposed approach is a hybrid ML -physics-based approach, which makes full use of the advantages of ML techniques in mapping the highly nonlinear relations exhibited by the physical experimental data and the advantages of physical models satisfying mechanics and physical laws.

4.4.2 An illustrative example: RC columns

This section presents the detailed procedure for the establishment of the proposed data-driven framework specifically for RC columns. The proposed framework can be implemented for any structural component as long as the experimental dataset is available. The circular RC column dataset presented in **Chapter III** (see Appendix B) is used to validate the novel data-driven framework in generalized seismic response history prediction of these columns under both quasi-static cyclic loading and shake table tests. The hysteretic model developed in **Chapter III** is utilized. Since the column dataset has been presented in **Chapter III**, there is no need to specify an optimization algorithm at this time.

4.4.3 Data-driven seismic response solvers

At this point, the proposed framework is employed to predict the seismic response of RC columns subjected to both displacement-controlled quasi-static cyclic loading and dynamic ground motions. Two data-driven seismic response solvers, one for displacement-controlled quasi-static

cyclic loading and another for dynamic ground motions, are developed to implement the proposed data-driven computing framework. For a structural component (e.g., column) subjected to lateral earthquake loads, it can be equivalent to a single degree of freedom (SDOF) model. **Algorithm 4.1** presented below is used to implement the proposed approach for an SDOF model subjected to displacement-controlled quasi-static cyclic loading.

Algorithm 4.1: Implementation of proposed SDOF model under quasi-static cyclic loading

1. Development of hysteretic modeler:
   Given an RC column training set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$, AI model $M(\cdot)$, hysteretic model $f(\cdot)$, and a new RC column;
   (a) translate the new column into predictors, denoted as a query point $\boldsymbol{x}_{new}$;
   (b) train an AI model $M(\boldsymbol{x}; \boldsymbol{\Psi})$ based on the RC column training set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$;
   (c) predict the response variables for the new column, denoted as $\widehat{\boldsymbol{y}}_{new} = M(\boldsymbol{x}_{\text{new}}; \boldsymbol{\Psi})$;
   (d) form a hysteretic modeler for the new column, denoted as $[f_s, k] = f(\delta; \widehat{\boldsymbol{y}}_{new})$;
2. Predict hysteretic response using the developed hysteretic modeler:
   Given the displacement history $\boldsymbol{\delta} = (\delta^1, \ldots, \delta^D)^T$, hysteretic modeler $[f_s, k] = f(\delta; \widehat{\boldsymbol{y}}_{new})$;
   for d = 1 to D **do**
      (a) calculate the lateral force at the lateral displacement δ^d, denoted as $f_s(\delta^d) = f(\delta^d; \widehat{\boldsymbol{y}}_{new})$;
   end for

Given any RC column (flexure-, shear-, or flexure-shear-critical), **Algorithm 4.1** can be used to predict the seismic response history in terms of the predicted hysteretic force-displacement curve $\left\{\left(\delta^d, f_s(\delta^d)\right)\right\}_{d=1}^{D}$. It should be noted that the hysteretic modeler can adaptively produce the force $f_s(\delta^d)$ and tangent stiffness $k(\delta^d)$ at the lateral displacement δ^d, which are critical components for the nonlinear time-history analysis. **Algorithm 4.2** is developed to implement the proposed approach for the SDOF model subjected to dynamic ground motions. **Algorithm 4.2** is a hybrid algorithm coupling the Newmark average acceleration method, modified Newton-Raphson iteration, and the hysteretic modeler developed in this work. After implementing **Algorithm 4.2**, the time-history response quantity of interest can be obtained. This can be the time-displacement response $\{(t_t, \delta^t)\}_{t=1}^{T}$ or force-displacement response $\left\{\left(\delta^t, f_s(\delta^t)\right)\right\}_{t=1}^{T}$.

Algorithm 4.2: Implementation of proposed SDOF model under dynamic ground motions

1. Initialization:

   Given the ground motion $\left\{\left(\ddot{\delta}_g(t_t)\right)\right\}_{t=1}^{T}$, hysteretic modeler $f(\delta;\hat{\mathbf{y}}_{new})$, mass m and damping constant c for the new column;

   (a) calculate the initial tangent stiffness for the new column from the hysteretic modeler $k^0 = f(\delta;\hat{\mathbf{y}}_{new})$;

   (b) select an appropriate time interval Δt and calculate the earthquake force: $p^t = -m\ddot{\delta}_g(t_t)$;

   (c) calculate the Newmark coefficients: $A = 4m/\Delta t + 2c$; $B = 2m$;

2. Solving the equation of motion for an SDOF system by the hybrid algorithm:

   Given the initial condition of the new RC column, i.e., p^0, δ^0, and $\dot{\delta}^0$, $f_s(\delta^0)$, and known information from step 1;

   (a) calculate the $\ddot{\delta}^0 = \left(p^0 - c\dot{\delta}^0 - f_s(\delta^0)\right)/m$;

   for t = 1 to T **do**

   (a) $\Delta\hat{p}^{t-1} = p^t - p^{t-1} + A\dot{\delta}^{t-1} + B\ddot{\delta}^{t-1}$;

   (b) $\hat{k}^{t-1} = k^{t-1} + 2c/\Delta t + 4m/(\Delta t)^2$;

   (c) calculate the $\Delta\delta^{t-1}$, k^t, $f_s(\delta^t)$ using modified Newton-Raphson and hysteretic modeler $f(\delta;\hat{\mathbf{y}}_{new})$

   Given $f_s(\delta^{t-1})$, δ^{t-1}, $\Delta\hat{p}^{t-1}$, $\hat{k}^{t-1}$, k^{t-1}, maximum number of iteration N, and tolerance *tol*

   (a) initial assignment: $f_s(\delta_0^t) = f_s(\delta^{t-1})$, $\delta_0^t = \delta^{t-1}$, $\Delta R_1 = \Delta\hat{p}^{t-1}$, $\hat{k} = \hat{k}^{t-1}$, $k = k^{t-1}$;

   for $j_n = 1$ to N **do**

   (a) $\Delta\delta_{j_n} = \Delta R_{j_n}/\hat{k}$;

   (b) $\delta_{j_n}^t = \delta_{j_n-1}^t + \Delta\delta_{j_n}$;

   (c) calculate the $k_{j_n}^t$ and $f_s\left(\delta_{j_n}^t\right)$ using the hysteretic modeler: $\left[f_s\left(\delta_{j_n}^t\right), k_{j_n}^t\right] = f\left(\delta_{j_n}^t;\hat{\mathbf{y}}_{new}\right)$;

   (d) $\Delta f_{j_n} = f_s\left(\delta_{j_n}^t\right) - f_s\left(\delta_{j_n-1}^t\right) + \left(\hat{k} - k\right)\Delta\delta_{j_n}$;

   (e) $\Delta R_{j_n+1} = \Delta R_{j_n} - \Delta f_{j_n}$;

   (f) calculate the convergence criterion: $\Delta\delta = \sum_{i_n=1}^{j_n}\Delta\delta_{i_n}$, $eps = \left\|\Delta\delta_{j_n}\right\|/\|\Delta\delta\|$;

   (g) $\Delta\delta^{t-1} = \Delta\delta$, $k^t = k_{j_n}^t$, and $f_s(\delta^t) = f_s\left(\delta_{j_n}^t\right)$;

   if $eps \le tol$ **do**

   (a) break the loop;

   end if

   end for j_n

   (d) $\Delta\dot{\delta}^{t-1} = 2\Delta\delta^{t-1}/\Delta t - 2\dot{\delta}^{t-1}$;

   (e) $\Delta\ddot{\delta}^{t-1} = 4\Delta\delta^{t-1}/(\Delta t)^2 - 4\dot{\delta}^{t-1}/\Delta t - 2\ddot{\delta}^{t-1}$;

   (f) $\delta^t = \delta^{t-1} + \Delta\delta^{t-1}$, $\dot{\delta}^t = \dot{\delta}^{t-1} + \Delta\dot{\delta}^{t-1}$, and $\ddot{\delta}^t = \ddot{\delta}^{t-1} + \Delta\ddot{\delta}^{t-1}$;

   end for t

4.4.4 Numerical results

This section presents the numerical experiments carried out to validate the proposed data-driven framework in generalized seismic response history prediction of RC columns under displacement-controlled quasi-static cyclic loading and dynamic shake table tests. For the displacement-controlled quasi-static cyclic loading test, six RC columns are randomly selected from the circular RC column dataset presented in **Chapter III** (see Appendix B) to serve as the test specimens. The remaining 154 columns serve as the training set. For the dynamic shake table test, a full-scale circular RC bridge column subjected to six earthquake (EQ) ground motions is selected as the test

specimen. In each case, the proposed approach is compared with widely-used traditional modeling techniques based on experimental data. All the numerical experiments are performed using a Desktop PC with the Processor: Intel(R) Xeon(R) CPU E3-1270 v6 @ 3.80 GHz.

4.4.4.1 Displacement-controlled quasi-static cyclic loading test

To validate the capabilities of the novel framework, widely-used distributed plasticity approaches were employed to predict the hysteretic response of the flexure-, shear-, and flexure-shear-critical RC columns. The classic fiber beam-column element (Spacone et al. 1996a; 1996b) is utilized to model the nonlinear cyclic response of flexure-critical RC columns. However, the classic fiber beam-column element fails to reasonably capture the nonlinear hysteretic behavior of shear- and flexure-shear-critical RC columns, as illustrated in Deierlein et al. (2010) and Marini and Spacone (2006). There are many existing methods proposed to address this shortcoming (Elwood 2004; LeBorgne and Ghannoum 2013; Marini and Spacone 2006; Sasani 2007), and the modeling strategy proposed by Marini and Spacone (2006) is adopted to predict their hysteretic force-displacement responses. For the proposed framework, the locally-weighted least-squares support vector machine for regression (LWLS-SVMR) presented in **Section 4.3** is selected as the ML technique to train the model on the training set consisting of 154 circular column specimens introduced in **Section 3.2**. The predictors used in this section include the gross column cross-sectional area A_g, concrete compressive strength f_c', longitudinal reinforcement yield stress f_{yl}, longitudinal reinforcement area A_{sl}, column effective depth d, concrete cover c, transverse reinforcement yield stress f_{yt}, transverse reinforcement area A_{st}, stirrup spacing s, shear span a, and applied axial load P. The response variables are the optimal backbone curve and hysteretic parameters introduced in **Chapter III**. It should be noted that other ML techniques can be

employed. However, regardless of the ML technique employed, the objective is to accurately predict these response variables.

Six column specimens (N5 from Cheok and Stone 1986, NH3 from Vu et al. 1999, No. 13 and No. 19 from Ghee et al. 1989, S1 from McDaniel 1997, and UCI5 from Hamilton et al. 2002) are randomly selected from the circular RC column dataset presented in **Chapter III** (see Appendix B) to serve as the test column specimens. For the widely-used traditional modeling techniques, a single force-based fiber beam-column element with five Gauss-Lobatto integration points (i.e., monitoring sections) is employed to model specimens N5 and NH3 (flexure-critical columns). In each monitoring section, the cover concrete fiber is simulated using the modified Kent and Park model (Scott et al. 1982), and the core concrete fiber is simulated by the confined concrete model proposed by Mander et al. (1988) to represent the confinement effect of the stirrups. The reinforcement fiber is modeled by the Menegotto-Pinto model (Menegotto and Pinto 1973). For the remaining specimens, where specimens No.19 and S1 are shear-critical columns and specimens No.13 and UCI5 are flexure-shear-critical columns, the modeling strategy proposed by Marini and Spacone (2006) is used. This strategy requires an extra nonlinear shear constitutive law at the section level. The element, concrete, and reinforcement fibers for the four shear- and flexure-shear-critical specimens are defined in the same way as the flexure-critical specimens. The 'hysteretic' material in OpenSees (Mazzoni et al. 2006) is used to model the nonlinear shear behavior, and the backbone curves are defined according to the method suggested by Sezen (2008). Since there is still no effective method to define the parameters regarding the pinching behavior and strength and stiffness deterioration for the 'hysteretic' material, they are calibrated with the corresponding experimental data. All numerical models for these six randomly selected columns are implemented in OpenSees.

For the proposed data-driven framework, **Algorithm 4.1** is used. Specifically, the selected six columns are first featured by predictors (i.e., six columns are expressed as six query points where the response variables need to be predicted). Then, for each query point, the LWLS-SVMR is used to predict the response variables based on the aforementioned 154 training data. The detailed information for how the LWLS-SVMR works can be found in **Section 4.3**. Matlab 2018a is used to implement the proposed approach.

Figures 4.11, 4.12, and 4.13 present a comparison of the results between the proposed AI-enhanced framework and traditional modeling techniques; ground truth is defined as the experimental test results. Figure 4.11 demonstrates that the proposed approach effectively captures the unloading stiffness and cyclic strength deteriorations observed experimentally for the two flexure-critical columns. However, the traditional method cannot accurately reflect these behavioral characteristics, which is evident based on an apparent discrepancy with the experimental data in unloading stiffness for specimen NH3 and the inaccurate predictions of the cyclic strength deteriorations for flexure-critical specimen N5. Figure 4.12 illustrates that the proposed approach can reflect the hysteretic behavior of the two shear-critical columns, with accurate prediction of the lateral strength, cyclic strength deterioration, and pinching behavior when compared with those observed experimentally. The traditional model also reflects the pinching behavior and cyclic strength deterioration, but it fails to accurately describe the significant strength deterioration for column No. 19 and under-estimates the lateral strength of column S1. Figure 4.13 shows that the traditional method overestimates the initial stiffness for the two flexure-shear-critical columns, while the proposed approach accurately predicts their initial stiffnesses. For all six of the selected column specimens, the hysteretic curves predicted by the

proposed approach show closer agreement with the experimental data in comparison with those predicted by the widely-used traditional modeling techniques.

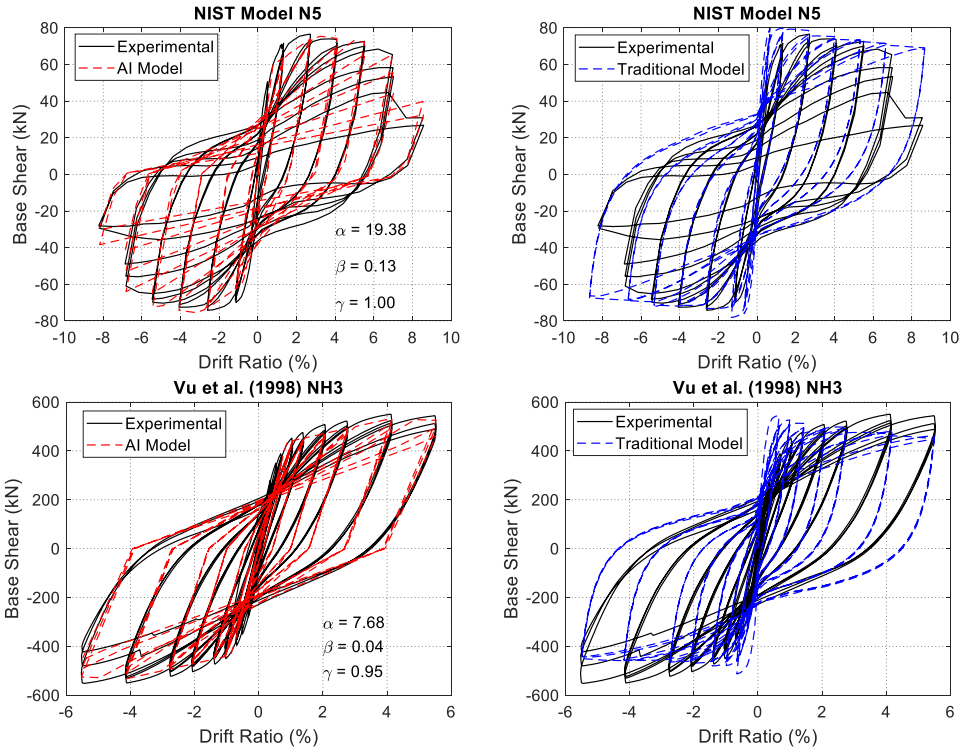


Figure 4.11 Comparison of results between proposed AI-based framework, experimental data, and widely-used traditional model for two flexure-critical columns.

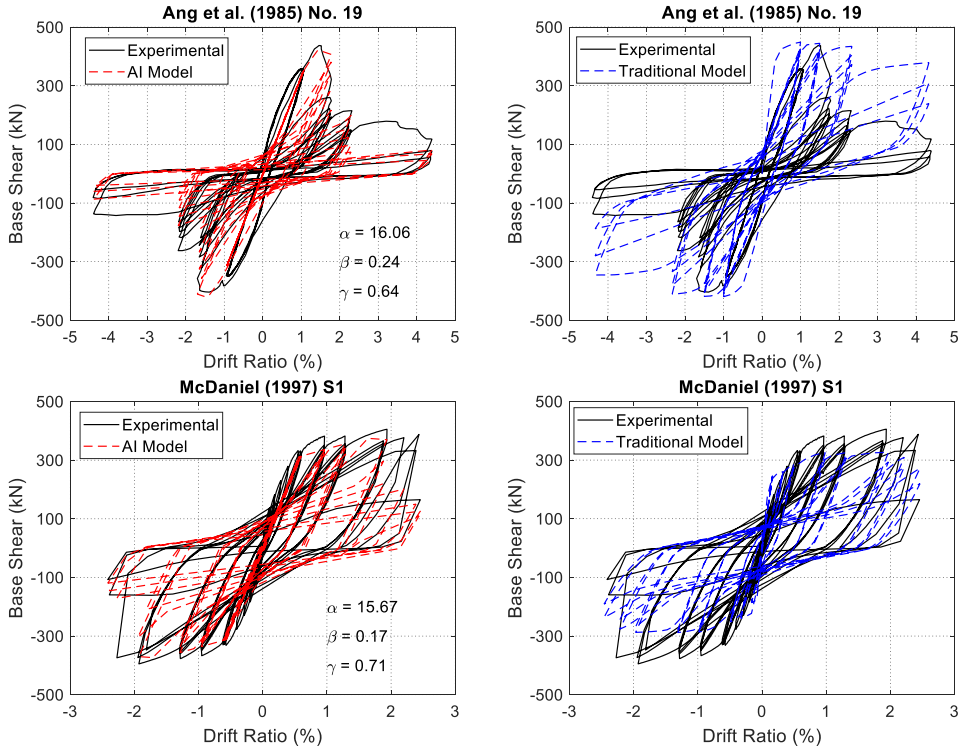


Figure 4.12 Comparison of results between proposed AI-based framework, experimental data, and widely-used traditional model for two shear-critical columns.

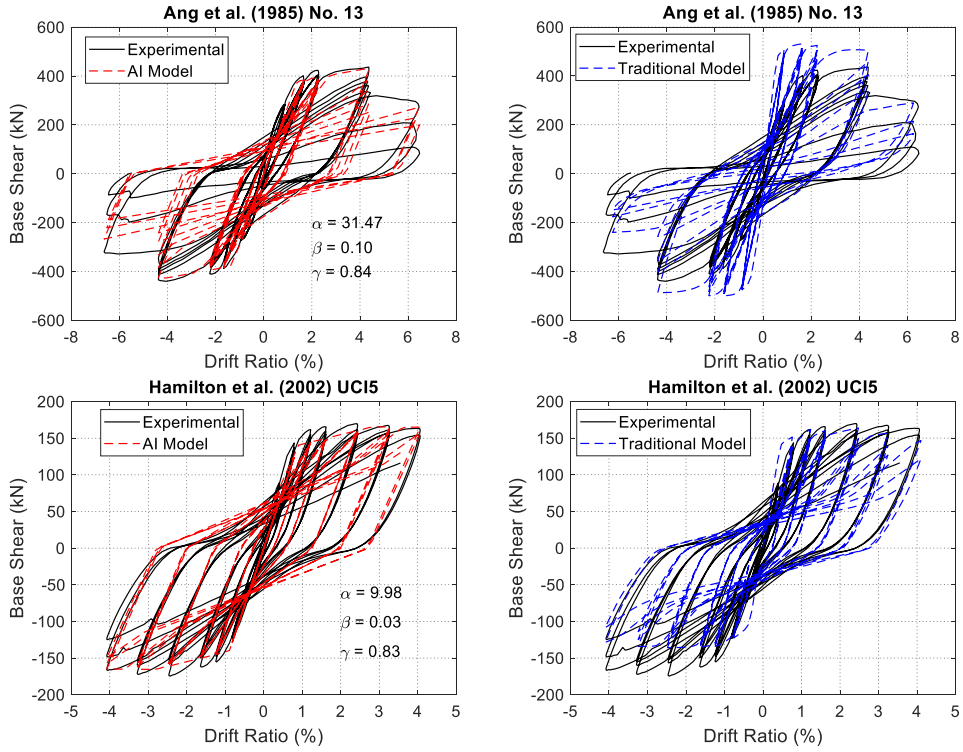


Figure 4.13 Comparison of results between proposed AI-based framework, experimental data, and widely-used traditional model for two flexure-shear-critical columns.

Additionally, the cyclic backbone curve parameters (i.e., V_y, δ_y, V_m, δ_m, V_u, δ_u), the accumulated hysteretic energy dissipation, and the computational time for all six selected columns are utilized to further compare these two approaches. The statistical indicators (mean absolute error (MAE), root mean squared error (RMSE), and coefficient of determination (R^2)), are used to quantify the performance. Note that a negative R^2 value corresponds to a very poor performance. The calculated metrics are provided in Table 4.9.

The calculated results show that the proposed data-driven framework presented in this work significantly outperforms the traditional approaches on all accounts. Specifically, the proposed approach enhances the R^2 values by approximately 47% (V_y), 9% (V_m), and 18% (V_u), reduces the RMSE by roughly 87% (V_y), 68% (V_m), and 56% (V_u), and reduces the MAE by

approximately 91% (V_y), 61% (V_m), and 61% (V_u). Furthermore, for the predictions of drift ratio at yield and maximum shear force, the performance of traditional approaches is significantly worse than that of the proposed method. Notably, the proposed method, when compared to traditional approaches, reduces the RMSE by approximately 85% (δ_y) and 87% (δ_m), reduces the MAE by approximately 89% (δ_y) and 85% (δ_m), and enhances the R^2 from negative values to 0.9463 (δ_y) and 0.9441 (δ_m). Moreover, for the prediction of the accumulated hysteretic energy dissipation, the proposed method enhances the R^2 value by approximately 37% and reduces the RMSE and the MAE by roughly 78% and 77%, respectively. Perhaps most importantly, the computational time for predicting the hysteretic curves of all six selected columns using the proposed method only requires 4 seconds in total, while using the traditional approaches takes 1,016 seconds. The proposed approach significantly reduces the computational cost. Based on these comparisons, the proposed approach performs significantly better than traditional methods; therefore, it is deemed that the proposed approach is the most appropriate means for generalized seismic response prediction of RC columns subjected to reversed cyclic loading, especially in near-real-time scenarios.

Table 4.9 Performance metrics for the proposed approach and widely-used traditional methods in predicting seismic response of the six selected column specimens.

Quantification Indicators	**AI Model**			**Traditional Model**		
	R^2	RMSE	MAE	R^2	RMSE	MAE
V_y (kN)	**0.9953**	**9.34**	**5.99**	0.6784	77.40	63.68
δ_y (%)	**0.9463**	**0.06**	**0.04**	-1.5945	0.40	0.35
V_m (kN)	**0.9905**	**16.13**	**12.84**	0.9073	50.51	33.00
δ_m (%)	**0.9441**	**0.25**	**0.21**	-2.1386	1.90	1.36
V_u (kN)	**0.9649**	**29.38**	**20.74**	0.8188	66.71	52.94
δ_u (%)	**0.8712**	**0.66**	**0.40**	0.5418	1.25	0.73
Dissipated Energy (kJ)	**0.9872**	**9.49**	**7.13**	0.7231	44.12	30.62
Computational Time (s)		**4**			1016	

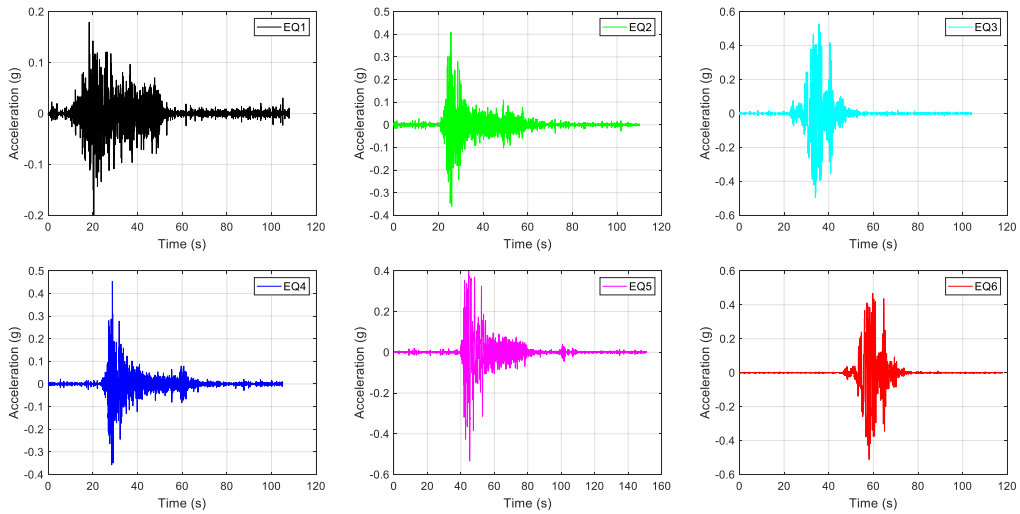


Figure 4.14 Six earthquake (EQ) levels that serve as sequential input for the full-scale RC bridge column.

4.4.4.2 Dynamic shake table tests

To validate the performance of the proposed data-driven framework in predicting the seismic response of the RC column subjected to ground motions, a full-scale RC bridge column specimen subjected to six consecutive ground motions is used as an example. These shake table tests were organized by the Pacific Earthquake Engineering Research (PEER) Center, and the detailed information regarding the physical experimental set up, structural features, ground motions, and results can be found in Terzic et al. (2015) and Schoettler et al. (2012). The six earthquake (EQ) levels (or ground motions) are presented in Figure 4.14. For traditional modeling approaches, since this RC bridge column is designed as a flexure-critical column (Terzic et al. 2015), the fiber beam-column element is also used to model the seismic response history. The element type, integration method, number of integration points, and material constitutive models described in **Section 4.4.4.1** for the two flexure-critical columns are also used here to establish the numerical model of the full-scale RC bridge column. For the proposed approach, **Algorithm 4.2** is used. Specifically, the bridge column is first featured as a query point by the predictors introduced in **Section 4.4.4.1**.

Then, the hysteretic modeler for this RC bridge column is formed using the LWLS-SVMR based on the 154 training data specimens introduced previously. Finally, the established hysteretic modeler is incorporated into **Algorithm 4.2** for dynamic response prediction. For both approaches, the damping ratio is set to 0.03. The time step (or time interval) is set to 0.0042s. Since the bridge column is not repaired after each ground motion (Terzic et al. 2015), the six ground motions are grouped sequentially and applied as a single ground motion that serves as the input to the two numerical models, which have been implemented in OpenSees and Matlab 2018a. Time-history results are presented in Figures 4.15 and 4.16.

Figure 4.15 presents the comparison of the predicted time-displacement results between the traditional approach and the proposed method, with the experimental data serving as the ground truth. The proposed method achieves better agreement with the experimental data for EQ1, EQ3, EQ5, and EQ6 over the full-time histories when compared with the traditional approach. Although both the proposed and the traditional approaches have apparent discrepancies with the experimental data for EQ2 and EQ4 over the full-time histories, the proposed method captures the maximum drift ratio (i.e., peak drift ratio) more accurately than the traditional modeling approach. The maximum drift ratio is an important engineering demand parameter (EDP) which is typically used to quantify the seismic performance of an RC structure (Bracci et al. 1997; Deierlein et al. 2010; Moehle 2014). Therefore, in this sense, the proposed AI-enhanced framework still performs better than the traditional model for EQ2 and EQ4, and thus, for all ground motions.

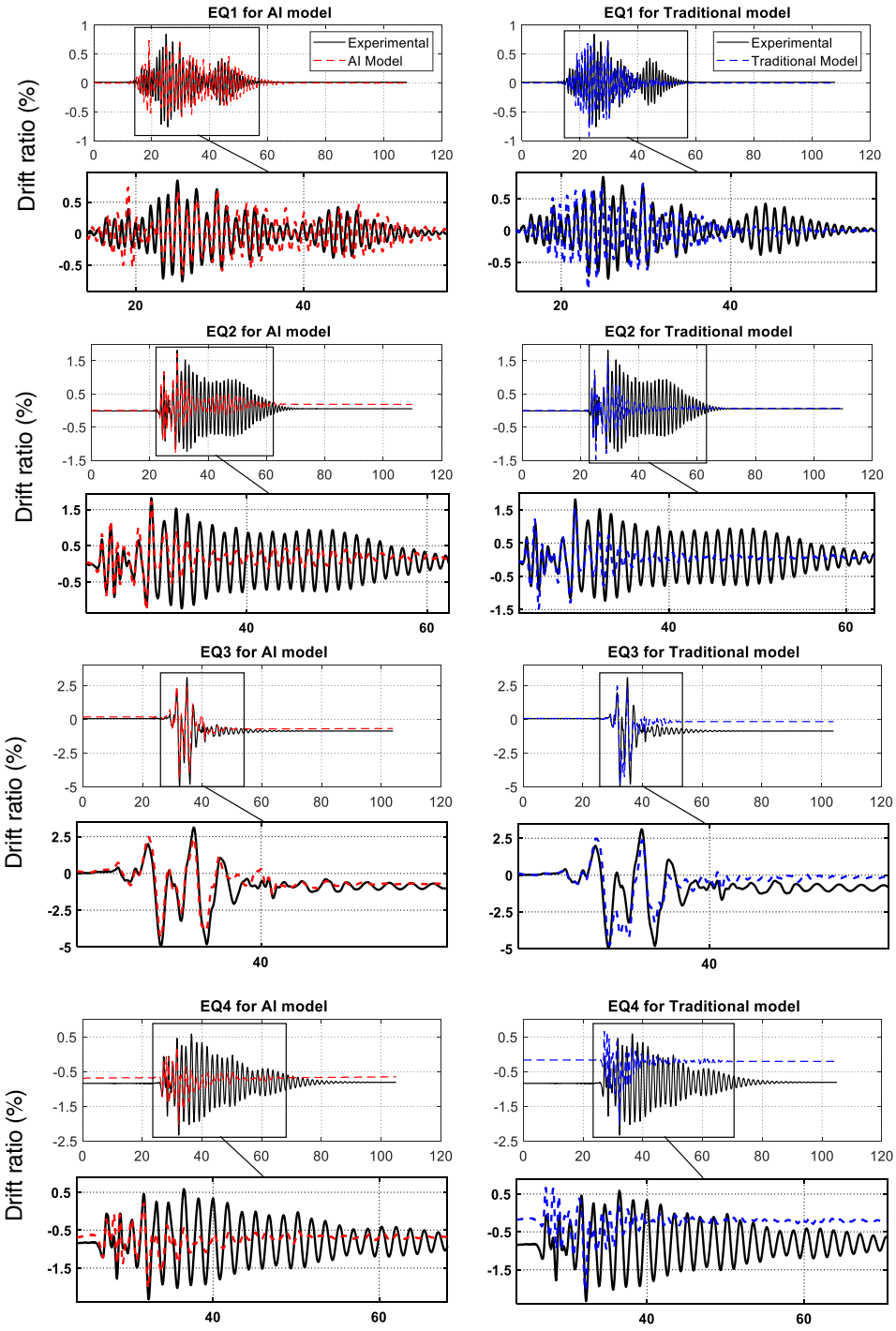


Figure 4.15 Comparison of time (s) vs. displacement (%) between the traditional approach and the proposed AI model, with the experimental data serving as the ground truth.

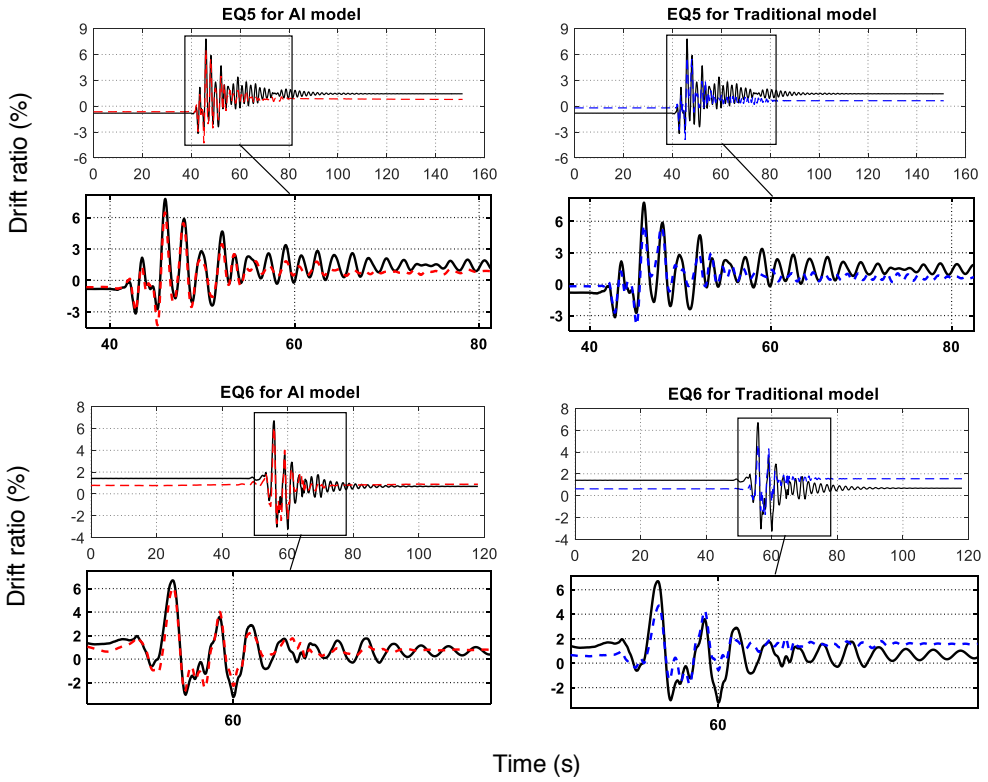


Figure 4.15 Continued.

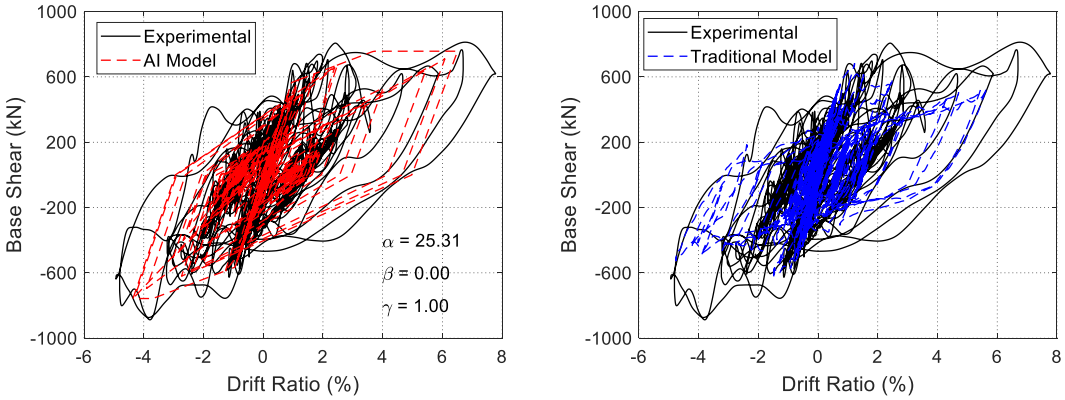


Figure 4.16 Hysteretic curves for the proposed AI-based framework, the traditional modeling approach, and the experimental data for all six ground motions.

Figure 4.16 shows the comparison of the results regarding the predicted hysteretic curves under all six ground motions. It is observed that the proposed data-driven framework accurately reflects the capacity in terms of the hysteretic energy dissipation (i.e., the area encompassed by the

hysteretic loops), while the traditional modeling approach underestimates the capacity of the column in this manner. Moreover, the maximum drift ratio, residual drift ratio, maximum shear, and hysteretic energy dissipation are considered as the engineering demand parameters (EDPs) of interest in this work to better compare the prediction performance of the RC bridge column subjected to six ground motions. Figure 4.17 presents the predicted EDPs for each ground motion from the proposed and traditional approaches. It is clearly evident from this figure that in most cases the proposed data-driven framework achieves a closer agreement with the experimental data than the traditional modeling approach when predicting these EDPs.

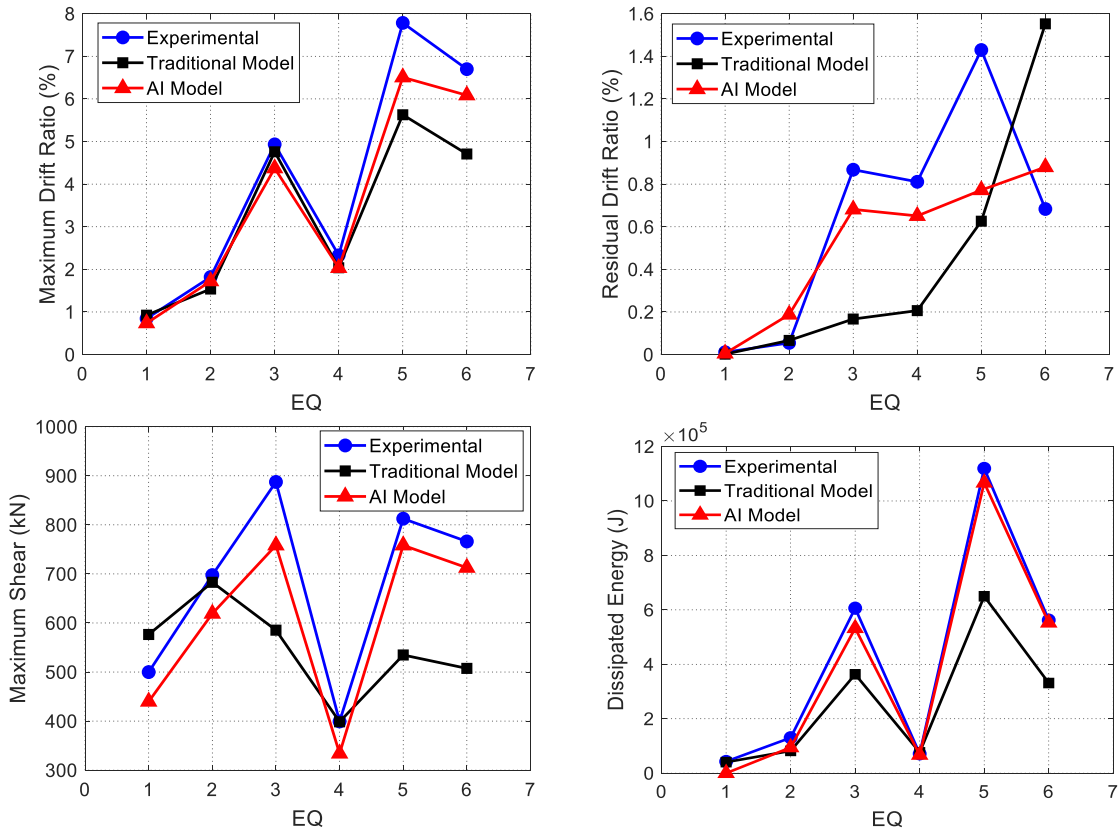


Figure 4.17 Predicted maximum drift ratio, residual drift ratio, maximum shear, and accumulated hysteretic energy dissipation for each of the six earthquake (EQ) levels.

Additionally, the prediction performance for these EDPs in Figure 4.17 are quantified by the MAE, RMSE, and R^2 metrics. The computational time for the RC bridge column analysis under all six ground motions is also utilized to compare the computational cost of the existing and proposed approaches. The calculated metrics are provided in Table 4.10. The calculated results show that the traditional approach is significantly outperformed by the proposed method for all metrics and all response quantities. The proposed approach enhances the R^2 value by approximately 21% for maximum drift ratio and 58% for hysteretic energy dissipation and increases the R^2 value from a negative value (-0.5633) to 0.6198 for residual drift ratio and from -0.3449 to 0.7964 for maximum shear force. Further, the proposed approach reduces the RMSE and MAE by roughly 47% and 41%, respectively for maximum drift ratio, 51% and 56%, respectively for residual drift ratio, 61% and 53%, respectively for maximum shear force, and 82% and 78%, respectively for hysteretic energy dissipation. More importantly, the computational time for predicting the seismic response history of the full-scale RC bridge column under all six ground motions using the proposed approach only requires 137 seconds, while that using the traditional approach is 10,991 seconds. The proposed approach significantly enhances the computational efficiency and shows great potential for regional seismic risk quantification, which requires hundreds of thousands of non-linear time-history analyses. Thus, the proposed approach performs significantly better than the traditional method for all seismic response quantities and agrees better with the experimental data.

Table 4.10 Performance metrics for the proposed AI-enhanced framework and the widely-used traditional modeling technique in predicting the seismic response of a full-scale RC bridge column subjected to six ground motions.

Quantification Indicators	AI Model			Traditional Model		
	R^2	RMSE	MAE	R^2	RMSE	MAE
Maximum Drift Ratio (%)	**0.9391**	**0.64**	**0.49**	0.7793	1.21	0.83
Residual Drift Ratio (%)	**0.6198**	**0.30**	**0.22**	-0.5633	0.61	0.50
Maximum Shear (kN)	**0.7964**	**78.04**	**73.49**	-0.3449	200.61	155.11
Dissipated Energy (kJ)	**0.9872**	**43.64**	**36.46**	0.6241	236.40	166.64
Computational Time (s)		**137**			10991	

4.4.4.3 Discussion of results

From the comparison of the results of both displacement-controlled quasi-static cyclic loading and shake table tests, it can be concluded that the proposed data-driven framework significantly outperforms the widely-used traditional modeling approaches in predicting the seismic response history of RC columns. The hysteretic curves predicted via the proposed approach will not perfectly match with the experimental data when the experimental hysteretic curves are smoother. This is because the polygonal hysteretic model is utilized to construct the hysteretic modeler where every branch of the force-displacement diagram follows a linear relationship. However, for the RC columns, the polygonal hysteretic model is sufficient to model the hysteretic behavior, and the predicted results presented in this section also demonstrate this fact. For other components (e.g., components constructed by steel material) having smoother experimental hysteretic curves, a smooth hysteretic model (e.g., Sivaselvan and Reinhorn 2000) can be used.

Additionally, when predicting the hysteretic curves of the test shear- and flexure-shear-critical columns using the widely-used traditional approach, accurate definition of the shear constitutive laws is required. However, there is still no unified method available to accurately define the shear behavior parameters (e.g., parameters regarding pinching and strength and

stiffness deterioration). Although these parameters can be calibrated with the experimental data, it is impractical when experimental data is not available. Nevertheless, the proposed approach does not suffer from this drawback and is applicable to flexure-, shear-, and flexure-shear-critical columns. More importantly, the proposed approach is extremely computationally efficient, exhibiting a significant reduction in computational time in comparison with the widely-used traditional approaches. These characteristics of the proposed approach demonstrate its' great potential in quantifying regional seismic risk and for other near-real-time scenarios.

Finally, although this work utilizes RC columns as an example to illustrate the proposed data-driven framework, it is a generalized approach and can be applied to any structural component of interest. Further, the establishment of the training dataset is an important factor, which is closely related to the collected physical experimental data, the selected hysteretic model, and the optimization algorithm.

4.5 Summary

This chapter has presented the development and validation of a novel component-level data-driven framework for generalized, accurate and efficient seismic response prediction of structural components. First, a novel machine learning-based backbone curve model (ML-BCV) is proposed for hardening behavior prediction in terms of an RC column's cyclic backbone curve without consideration of the softening branch. The proposed model consists of a modified LS-SVM to address the multi-output case (MLS-SVMR) and a GSA to more effectively facilitate the training process and more accurately predict the hardening behavior of RC columns subjected to reversed cyclic loading for flexure, shear, and flexure-shear failure modes. Using the MLS-SVMR, the nonlinear function that maps a multi-dependent variable output space from a multi-independent variable input space is ascertained. Then, a GSA optimization algorithm assisted training process is adopted to exhaustively and adaptively search for the most proper hyper-parameters for the MLS-SVMR. This proposed ML-BCV model can accurately predict the bi-linear backbone curve and thus, existing capacity and structural performance of RC columns solely based on the material and geometric properties, applied loads, and failure modes without human intervention, intelligence, or any assumptions. This makes the proposed ML-BCV a more robust approach than traditional modeling techniques. Additionally, a 10-fold-cross validation procedure is embedded in the objective GSA optimization function to establish a desirable prediction model that prevents overfitting and is robust with highly generalized performance. The predicted performance results prove that this strategy is capable of overcoming the problem of overfitting and reaches high accuracy in both training and testing results. The proposed ML-BCV was also compared with traditional modeling approaches, and it was found that the performance of the newly proposed ML-BCV model yields more accurate results than traditional modeling approaches.

Furthermore, a novel ML model, LWLS-SVMR, which integrates LS-SVMR with locally-weighted training criterion is proposed for softening behavior prediction in terms of an RC column's drift capacity. The proposed LWLS-SVMR can overcome the possible negative interference from irrelevant data points, and thus more accurately discover highly complex nonlinear relationships between influential factors and response values. An efficient strategy for hyper-parameter tuning of the proposed LWLS-SVMR was also developed using a hybrid global optimization algorithm, CSA, and an exhaustive searching algorithm, GSA, to facilitate the training process. CSA was used to determine appropriate starting values for the hyper-parameters, and GSA was then utilized to further exhaustively search for the optimum pairs within a small region encompassing the initial values. In order to demonstrate the superiority of the proposed LWLS-SVMR, results obtained from validation set, 10-fold cross-validation, and leave-one-out (LOO) cross-validation approaches were compared with those obtained by three popularly used models: a global ML model (LS-SVMR), an existing, local ML model (LWQR), and an empirical model (Elwood and Moehle 2005). The results proved that the proposed LWLS-SVMR outperformed all models in predicting the drift capacity across RC flexure-, shear-, and flexure-shear-critical columns.

Finally, a novel data-driven framework is proposed for predicting the seismic response history of structural components under displacement-controlled quasi-static cyclic loading and shake-table tests. The proposed data-driven framework is a hybrid approach, coupling an ML technique (e.g., the proposed LWLS-SVMR) and a physical model (i.e., hysteretic model). In this way, ML is used to directly link the experimental data with the nonlinear properties of target structural components, and the physical model is used to perform the seismic analysis, efficiently leveraging the advantages of both approaches. To validate the performance of the proposed

approach, RC columns are selected as an illustrative example. Two data-driven seismic response solvers are developed to implement the proposed method. The numerical results validate that the proposed approach significantly outperforms the widely-used distributed plasticity approaches in predicting the seismic response history of RC columns under both quasi-static cyclic loading and shake table tests. Moreover, the proposed method significantly enhances the computational efficiency for both cases in comparison with the widely-used traditional approaches, yielding great potential for regional seismic risk quantification and other near-real-time needs in a more accurate and efficient way.

Thus, with all three models, we now arrive at full ML-based prediction of the seismic response of reinforced concrete columns. It should be noted that the proposed three models can also be applied to other structural components if the corresponding component dataset is available.

CHAPTER V

SYSTEM-LEVEL DATA-DRIVEN COMPUTING FRAMEWORK

5.1 Overview

Existing physics-based modeling approaches do not have a good compromise between performance and computational efficiency in predicting the seismic response of reinforced concrete (RC) structural systems. The high-fidelity models have reasonable predictive performance but are computationally demanding, while more simplified models may be computationally efficient, but do not have as good of performance. This chapter presents a novel data-driven computational framework for the seismic response history prediction of RC structural systems to remedy this problem.

The proposed system-level data-driven framework integrates the component-level data-driven framework presented in **Section 4.4** with a simplified shear building model. The component-level data-driven framework can directly link the experimental data to nonlinear properties of any structural component, while the shear building model (that meets the universal laws such as Newton's law of motion) can perform a seismic analysis at the system level.

Two data-driven seismic response solvers are developed to implement the proposed approach. The proposed system-level computational framework is utilized for seismic response prediction of a large-scale 3-bay, 3-story RC frame under cyclic loads as well as of two small-scale 3-bay, 9-story RC frames subjected to four and six consecutive ground motions respectively. Compared to the experimental data, the results demonstrate that the proposed system-level data-driven framework outperforms the widely used distributed plasticity fiber model in both prediction capabilities and computational efficiency. Therefore, the framework is deemed a promising tool to achieve a good compromise between computational cost and performance. The detailed information is presented below.

5.2. Methodology

This section presents the novel system-level data-driven framework to predict the hysteretic behavior and time-history response quantities of target RC structures (e.g., RC frames) subjected to both quasi-static cyclic loading and ground motions in a generalized, accurate, and efficient way. The framework includes three steps. First, component-level hysteretic modelers are formed for all the main load-bearing elements in the target structural system based on the hybrid model presented in Chapter IV. The second step is to formulate a multi-degree of freedom (MDOF) numerical model for the target structural system by incorporating the hysteretic modelers developed in Step 1 into the well-established structural model. Finally, two data-driven seismic response solvers are developed to solve the MDOF model subjected to earthquake loads (i.e., displacement-controlled quasi-static cyclic loading or ground motions). In such a way, the hysteretic behavior or time-history response quantities of the target structure subjected to earthquake loads can be obtained. Each of these three steps will be introduced in detail in the following sub-sections.

5.2.1 Component-level hysteretic modelers

The component-level hysteretic modeler is denoted as $[f_s, k] = f(\delta; \boldsymbol{y})$, where $\boldsymbol{y} \in R^{n_\theta}$ is the optimal critical parameter vector containing n_θ critical parameters that define a hysteretic model, and $f(\cdot)$ represents the hysteretic model. This modeler is employed to produce the force f_s and tangent stiffness k for the component in a structure at a deformation δ at each load step or time instant. The component could be a beam, column, wall, etc. The modeler is a hybrid-ML-physics-based model, as shown in Figure 5.1. For different types of components, the modeler also varies. Specifically, given the collected physical experimental data (i.e., structural features and force-deformation data) of n components (e.g., either beam, column, or wall in a frame-wall structure),

a training dataset $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ can be developed using the method presented in Chapter III. The training dataset consists of the necessary structural features (e.g., specimen geometry and material properties) denoted as $\boldsymbol{x}_i \in R^p$ that serve as predictors and an optimal critical parameter vector $\boldsymbol{y}_i \in R^{n_\theta}$ that serves as the response variables.

In the example presented in Figure 5.1, three types of training sets – one for beams, one for columns, and one for walls – can be developed. Given these three training sets, three well-trained ML models – one for beams, one for columns, and one for walls – can be formed by learning the nonlinear relations exhibited by the training sets using ML techniques. The well-trained ML model is denoted as $\widehat{\boldsymbol{y}} = M(\boldsymbol{x}; \boldsymbol{\Psi})$, where $\boldsymbol{\Psi} \in R^{n_\psi}$ is the optimal ML model parameter vector containing n_ψ parameters and $M(\cdot)$ represents the ML technique. Note that the use of ML techniques is flexible. The ML techniques used for these three components could be the same or different from each other, as shown in Figure 5.1. Regardless of the types of ML techniques employed, it is necessary that the techniques have high generalization performance. Then, each component in the target structure needs to be featured by corresponding predictors, as shown in Figure 5.1. In such a way, each component is expressed as a query point denoted as $\boldsymbol{x}_{new} \in R^p$. Note that the p value may be different for each type of component, but it must be the same as those in the training sets (i.e., the type and number of predictors for beam, column, and wall may be different, but within the space of each component, the predictors must be the same). These predictors for all the components in the target structure are input to the corresponding well-trained ML models to obtain the optimal critical parameter vector $\widehat{\boldsymbol{y}}_{new} = M(\boldsymbol{x}_{new}; \boldsymbol{\Psi})$, as shown in Figure 5.1. The optimal critical parameter vector $\widehat{\boldsymbol{y}}_{new}$ is then applied to the hysteretic model to form the component-level hysteretic modeler $[f_s, k] = f(\delta; \widehat{\boldsymbol{y}}_{new})$. Note that the use of hysteretic models is also flexible. The hysteretic models used for these three components could be the same or different from each other.

Regardless of the type of hysteretic models used, the requirements are that the hysteretic models must be the same as those used for the training sets and must reflect various hysteretic behaviors experienced by these components experimentally (e.g., pinching behavior and stiffness and strength deterioration). The detailed information regarding how the component-level hysteretic modelers are developed can be found in Chapter III.

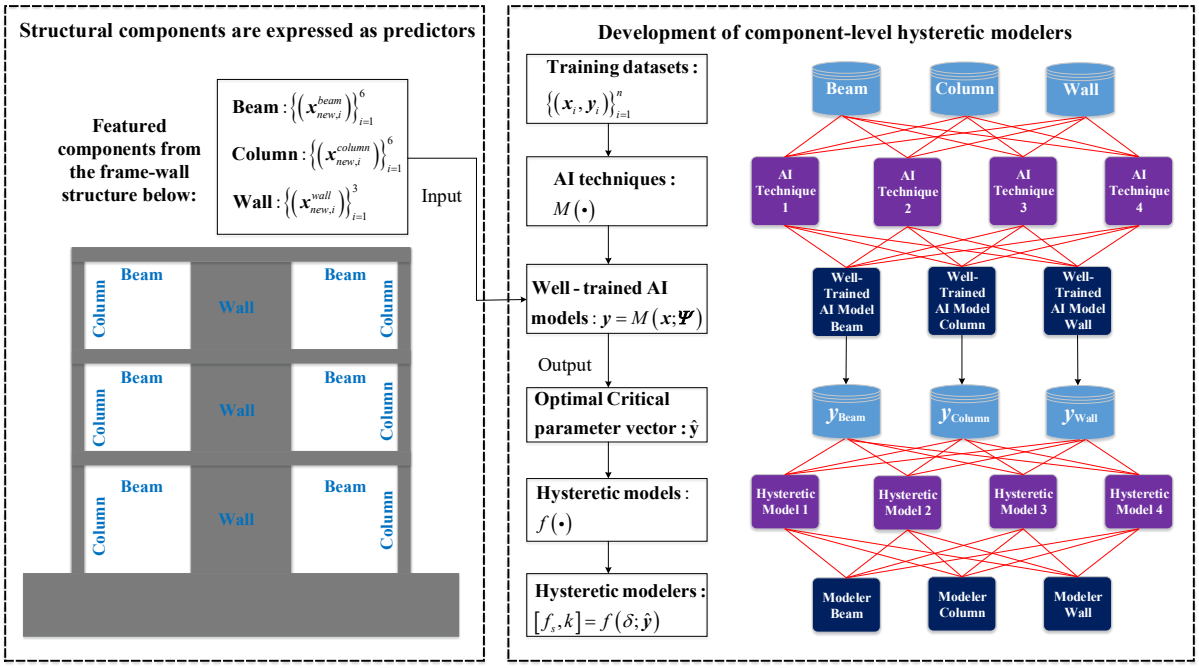


Figure 5.1 Procedure for establishing component-level hysteretic modelers for structural components in a structural system.

5.2.2 Formulation of an MDOF Model

For a structure subjected to earthquake loads, the component-level hysteretic modelers developed in **Section 5.2.1** for the components in the structure can generate the tangent stiffness and resisting force in their deflected DOFs given the displacement information. Then, these tangent stiffnesses and resisting forces can be assembled to form a structure stiffness matrix and resisting force vector for further calculation at the system level. In this dissertation, an MDOF model is formulated in

terms of the simplified shear building model to predict the seismic response of RC frames. It should be noted that a high-fidelity model for any structural system can be developed and then seamlessly integrated into this framework but will not be discussed here since the aim of this dissertation is to reduce the computational cost while maintaining good prediction performance. The following assumptions are made to formulate the system-level MDOF model: 1) axial deformations are ignored in all structural components; 2) masses for each story are idealized as lumped at the nodes of the discretized structure; and, 3) all beams are axially and flexurally rigid. A schematic sketch of the proposed MDOF model is presented in Figure 5.2. Assume a planar RC frame structure has *n*-stories, with each story having *l*-bays, as shown in Figure 5.2. Based on the aforementioned assumptions, the nodal mass at each story, denoted as m_{ij}, where $i = 1, \dots, n$ represents the story and $j = 1, \dots, l+1$ represents the column along the bay direction, has the same lateral translational DOF. Thus, the mass matrix for this structure is a diagonal matrix in the lateral DOF direction, which is written as follows:

$$\boldsymbol{M} = \begin{bmatrix} \sum_{j=1}^{l+1} m_{1j} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \sum_{j=1}^{l+1} m_{nj} \end{bmatrix} \tag{5.1}$$

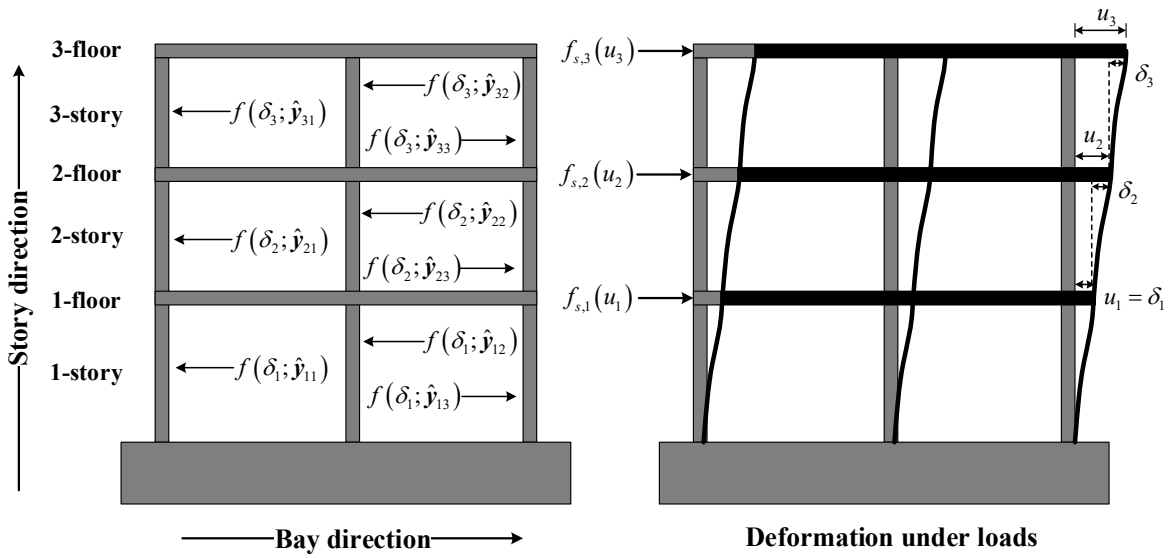


Figure 5.2 Schematic sketch of the proposed MDOF model.

The mass matrix will remain constant throughout the response history. As shown in Figure 5.2, the hysteretic property (e.g., tangent stiffness or shear force) for each column in each story is obtained by the hysteretic modeler presented in **Section 5.2.1**. The hysteretic property is denoted as $[f_{s,ij}, k_{ij}] = f(\delta_i; \hat{\mathbf{y}}_{ij})$ where $f_{s,ij}$ and k_{ij} are the lateral shear force and tangent stiffness of column j located at story i and obtained by the modeler $f(\delta_i; \hat{\mathbf{y}}_{ij})$ given the i^{th}-story relative displacement (or story drift) δ_i, respectively. The calculation of δ_i is $\delta_i = u_i - u_{i-1}$, $i \geq 2$, and when $i = 1$, $\delta_1 = u_1$, which means the relative story displacement δ_1 is equal to the lateral displacement u_1 at the first floor. The u_i is the lateral displacement relative to the ground at floor i. Note that the hysteretic property for each column in each story could be the same or they could vary from one another, depending on the selected hysteretic model and obtained optimal critical parameter vector $\hat{\mathbf{y}}_{ij}$. Due to the above assumptions, the structure stiffness matrix is a symmetric tri-diagonal matrix, which is written below:

$$\boldsymbol{K}=\begin{bmatrix} \sum_{j=1}^{l} k_{1j}+k_{2j} & -\sum_{j=1}^{l} k_{2j} & & 0 \\ -\sum_{j=1}^{l} k_{2j} & \sum_{j=1}^{l} k_{2j}+k_{3j} & \ddots & \\ & \ddots & \ddots & -\sum_{j=1}^{l} k_{nj} \\ 0 & & -\sum_{j=1}^{l} k_{nj} & \sum_{j=1}^{l} k_{nj} \end{bmatrix} \tag{5.2}$$

The structure stiffness matrix $\boldsymbol{K}$ will be updated when the column tangent stiffness k_{ij} changes due to nonlinear behavior throughout the response history. For the damping component, Rayleigh damping is used, which is a combination of mass-proportional and stiffness-proportional damping. The Rayleigh damping matrix is given by:

$$\boldsymbol{C}=a_0\boldsymbol{M}+a_1\boldsymbol{K} \tag{5.3}$$

The coefficients a_0 and a_1 can be determined from specified damping ratios ζ_{i_m} and ζ_{j_m} for the i_mth and j_mth modes, respectively. The detailed information regarding the calculation of a_0 and a_1 can be found in Chopra (2007). Given the mass, stiffness, and damping components, an MDOF model for an RC frame structure subjected to ground motions can be formulated.

$$\boldsymbol{M}\ddot{\boldsymbol{u}}+\boldsymbol{C}\dot{\boldsymbol{u}}+\boldsymbol{f}_S(\boldsymbol{u})=-\boldsymbol{M}\boldsymbol{1}\ddot{u}_g(t) \tag{5.4}$$

where $\boldsymbol{u}=(u_1,u_2,\dots,u_n)^T$ is a displacement vector along the structures' height, and each element represents the lateral floor displacement relative to the ground; $\boldsymbol{f}_S(\boldsymbol{u})$ is a lateral resisting force vector along the structures' height determined by the structure stiffness matrix $\boldsymbol{K}$ and corresponding displacement vector $\boldsymbol{u}$, or directly assembled by the story shear force $\sum_{j=1}^{l+1} f_{s,ij}$, $i=1,\dots,n$; $\boldsymbol{1}=(1,\dots,1)\in R^n$ is a column vector; and, $\ddot{u}_g(t)$ is the ground motion.

Note that Eq. (5.4) can be applied to both linear and nonlinear systems. This is because when solving Eq. (5.4), the structure stiffness matrix $\boldsymbol{K}$ is not constant and will be updated to determine the resisting force vector $\boldsymbol{f}_S(\boldsymbol{u})$ from the column tangent stiffness corresponding to the

deformation and state of each column. The solvers developed to obtain the hysteretic behavior and time-history response quantities will be introduced in the next sub-section.

5.2.3 Data-driven seismic response solvers

For the linear analysis, the initial structural stiffness matrix is used throughout the entire time history. Therefore, the $\boldsymbol{f}_S(\boldsymbol{u})$ term in Eq. (5.4) can be changed to $\boldsymbol{K}\boldsymbol{u}$ where $\boldsymbol{K}$ represents the initial structure stiffness matrix and will remain constant. For the nonlinear analysis, the structural stiffness matrix $\boldsymbol{K}$ is not constant and will be updated to determine $\boldsymbol{f}_S(\boldsymbol{u})$ from the column tangent stiffness corresponding to the deformation and state of each column in each story. Specifically, given the relative story displacement δ_i and state (e.g., loading or unloading) of each column in each story, the column hysteretic modelers can adaptively produce the column shear force and tangent stiffness $\left[f_{s,ij}, k_{ij}\right] = f\left(\delta_i; \widehat{\boldsymbol{y}}_{ij}\right)$ and record the current state. The recorded current state can trail if the deformation is in the loading branch, unloading branch, or at the reversal point where a transition happens between loading and unloading. Thus, this can inform the hysteretic modelers to determine the column shear force and tangent stiffness for the next load step or time instant. The produced shear force $f_{s,ij}$ and tangent stiffness k_{ij} for each column can be respectively assembled to a resisting force vector $\boldsymbol{f}_S(\boldsymbol{u})$ and structure stiffness matrix $\boldsymbol{K}$ for further calculation. Eq. (5.2) can be used to assemble a structure stiffness matrix $\boldsymbol{K}$ from the column tangent stiffness k_{ij}. Since the force-displacement relation is nonlinear, the direct calculation of the resisting force vector by $\boldsymbol{f}_S(\boldsymbol{u}) = \boldsymbol{K}\boldsymbol{u}$ is no longer valid. The static equilibrium constraint is used to directly assemble the resisting force vector $\boldsymbol{f}_S(\boldsymbol{u})$ from the column shear force $f_{s,ij}(\delta_i)$, $i = 1, \dots, n; j = 1, \dots, l+1$, produced by the hysteretic modelers. Figure 5.3 is an example to illustrate how the resisting force $f_{s,2}(u_2)$ at the 2nd floor is formed using the static equilibrium constraint.

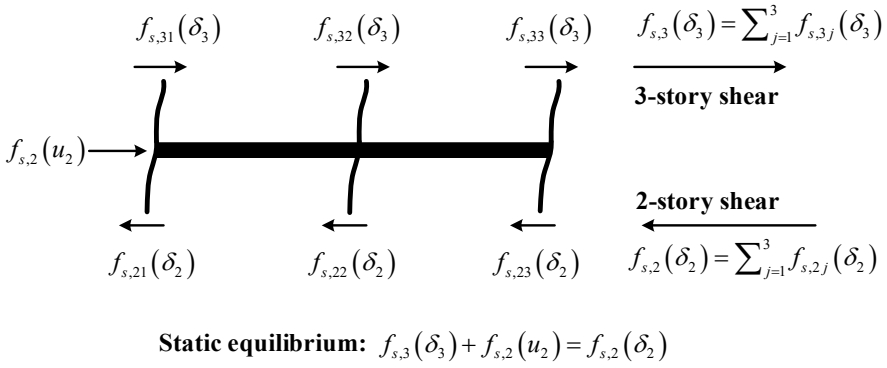


Figure 5.3 Determination of the resisting force from story shear by static equilibrium.

Specifically, given the shear force $f_{s,ij}(\delta_i)$ for each column at story *i*, the i^{th}-story story shear force can be calculated as $f_{s,i}(\delta_i) = \sum_{j=1}^{l+1} f_{s,ij}(\delta_i)$. The resisting force vector $\boldsymbol{f}_S(\boldsymbol{u})$ consists of the resisting force $f_{s,i}(u_i)$ at each floor, which is denoted as $\boldsymbol{f}_S(\boldsymbol{u}) = \left(f_{s,1}(u_1), f_{s,2}(u_2) \dots, f_{s,n}(u_n)\right)^T$. The resisting force $f_{s,i}(u_i)$ at floor *i* is made up of two contributions: $f_{s,i}(\delta_i)$ from the story of floor *i* below, and $f_{s,i+1}(\delta_{i+1})$ from the story of floor *i* above, as shown in Figure 5.3. To maintain static equilibrium, the following equation can be established:

$$f_{s,i+1}(\delta_{i+1}) + f_{s,i}(u_i) = f_{s,i}(\delta_i),\ 1 \le i \le n-1 \tag{5.5}$$

where, when $i = n$, the resisting force $f_{s,n}(u_n)$ equals $f_{s,n}(\delta_n)$.

This is because there is no story above floor *n*. So, the resisting force vector $\boldsymbol{f}_S(\boldsymbol{u})$ can be re-written as follows:

$$\boldsymbol{f}_S(\boldsymbol{u}) = \left(f_{s,1}(\delta_i) - f_{s,2}(\delta_2), \dots, f_{s,n-1}(\delta_{n-1}) - f_{s,n}(\delta_n), f_{s,n}(\delta_n)\right)^T \tag{5.6}$$

Thus, Eq. (5.6) can be used to assemble a resisting force vector $\boldsymbol{f}_S(\boldsymbol{u})$ with each column shear force in each story, updated for each time instant. For the displacement-controlled quasi-static cyclic loading, the floor displacement information $\boldsymbol{u}$ is known, and the quantity of interest is

regarding the hysteretic relationship between base shear and roof displacement or story shear and story drift (i.e., relative story displacement). The prediction of these quantities using the proposed data-driven MDOF model is straightforward. The following solver (**Algorithm 5.1**) is developed to implement the proposed approach to predict the hysteretic response of an RC frame subjected to quasi-static cyclic loading.

Algorithm 5.1: Implementation of proposed MDOF model under quasi-static cyclic loading

---

1. Development of hysteretic modelers:
   Given an RC column training set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$ and a target RC frame with *n* stories and *l* bays
   (a) translate the columns in each story in the target RC frame into predictors, denoted as query points $\left\{\left(\boldsymbol{x}_{\text{new},ij}\right)\right\}_{i=1}^{n\times(l+1)}$;
   (b) train an AI model $M(\boldsymbol{x};\boldsymbol{\Psi})$ based on the RC column training set $\{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^{n}$;
   (c) predict the response for each column in the target RC frame, denoted as $\hat{\boldsymbol{y}}_{ij} = M\left(\boldsymbol{x}_{\text{new},ij};\boldsymbol{\Psi}\right)$;
   (d) form a hysteretic modeler for each column, denoted as $\left[f_{s,ij}, k_{ij}\right] = f\left(\delta_i;\hat{\boldsymbol{y}}_{ij}\right)$, i = 1,…,n; j = 1,…,l+1;
2. Predict hysteretic response using proposed AI-enhanced MDOF model:
   Given the displacement history $\boldsymbol{U} = (\boldsymbol{u}^1, \ldots, \boldsymbol{u}^D)^T$, hysteretic modeler $\left[f_{s,ij}, k_{ij}\right] = f\left(\delta_i;\hat{\boldsymbol{y}}_{ij}\right)$, i = 1,…,n; j = 1,…,l+1
   for d = 1 to D **do**
      for i = 1 to n **do**
         (a) when i = 1, calculate the relative story displacement or story drift $\delta_1^d = u_1^d$
         (b) when $i \neq 1$, calculate the relative story displacement or story drift $\delta_i^d = u_i^d - u_{i-1}^d$
         for j = 1 to l+1 **do**
            (a) calculate the shear and tangent stiffness $\left[f_{s,ij}\left(\delta_i^d\right), k_{ij}\left(\delta_i^d\right)\right] = f\left(\delta_i^d;\hat{\boldsymbol{y}}_{ij}\right)$ for each column;
         end for j
         (a) calculate and record the story shear $f_{s,i}\left(\delta_i^d\right) = \sum_{j=1}^{l+1} f_{s,ij}\left(\delta_i^d\right)$;
         (b) calculate and record the story stiffness $k_i\left(\delta_i^d\right) = \sum_{j=1}^{l+1} k_{ij}\left(\delta_i^d\right)$;
      end for i
      (a) assemble the structure stiffness matrix $\boldsymbol{K}^d$ according to $\left\{\left(k_i\left(\delta_i^d\right)\right)\right\}_{i=1}^{n}$ using Eq. (5.2);
      (b) assemble the resisting force vector $\boldsymbol{f}_S(\boldsymbol{u}^d)$ according to $\left\{\left(f_{s,i}\left(\delta_i^d\right)\right)\right\}_{i=1}^{n}$ using Eq. (5.6)
      (c) output $\left\{\left(f_{s,i}\left(\delta_i^d\right)\right)\right\}_{i=1}^{n}$, $\boldsymbol{f}_S(\boldsymbol{u}^d)$, and $\boldsymbol{K}^d$.
   end for d

By implementing **Algorithm 5.1**, one can obtain the hysteretic response of both roof displacement ($\{(u_n^d)\}_{d=1}^{D}$) versus base shear ($\left\{\left(f_{s,1}(\delta_1^t)\right)\right\}_{d=1}^{D}$) and story i drift ($\left\{\left(\delta_i^d\right)\right\}_{d=1}^{D}$) versus story i shear ($\left\{\left(f_{s,i}\left(\delta_i^d\right)\right)\right\}_{d=1}^{D}$) for a target RC frame structure. Further, **Algorithm 5.1** can also output the structure stiffness matrix $\{(\boldsymbol{K}^d)\}_{d=1}^{D}$ and resisting force vector $\left\{\left(\boldsymbol{f}_S(\boldsymbol{u}^d)\right)\right\}_{d=1}^{D}$ given the entire

displacement history $\boldsymbol{U} = (\boldsymbol{u}^1, \ldots, \boldsymbol{u}^D)^T$, which are important components for the nonlinear time-history analysis. Thus, **Algorithm 5.1** will be used in **Algorithm 5.2** below to calculate the structure stiffness matrix $\boldsymbol{K}$ and resisting force vector $\boldsymbol{f}_s(\boldsymbol{u})$ given the displacement information $\boldsymbol{u}$, which is denoted as $[\boldsymbol{f}_s(\boldsymbol{u}), \boldsymbol{K}] = Algorithm5.1(\boldsymbol{u})$. The nonlinear dynamic analysis involves using a numerical method to solve the equations of motion presented in Eq. (5.4). In this section, a hybrid algorithm coupling the Newmark average acceleration (NAA) method, modified Newton-Raphson (MNR) iteration, and **Algorithm 5.1** is developed to solve Eq. (5.4). The detailed procedure is presented below.

Algorithm 5.2: Implementation of proposed MDOF model under dynamic ground motions

---

1. Initialization:

   Given the ground motion $\left\{\left(\ddot{u}_g(t_t)\right)\right\}_{t=1}^{T}$, hysteretic modeler $\left[f_{s,ij}, k_{ij}\right] = f\left(\delta_i; \hat{\boldsymbol{y}}_{ij}\right)$, i = 1,…,n; j = 1,…,l+1;

   (a) calculate the nodal mass m_{ij} in each story for the target RC frame;

   (b) calculate the initial tangent stiffness for each column from the hysteretic modeler: $\left[f_{s,ij}, k_{ij}\right] = f\left(\delta_i; \hat{\boldsymbol{y}}_{ij}\right)$;

   (c) calculate the mass, initial stiffness, and damping matrix $\boldsymbol{M}$, $\boldsymbol{K}^0$, and $\boldsymbol{C}$ using Eqs.(5.1-5.3), respectively;

   (d) select an appropriate time interval Δt and calculate the earthquake forces: $\boldsymbol{p}^t = -\boldsymbol{M}\boldsymbol{1}\ddot{u}_g(t_t)$;

   (e) calculate the Newmark coefficients: $\boldsymbol{A} = 4\boldsymbol{M}/\Delta t + 2\boldsymbol{C}$; $\boldsymbol{B} = 2\boldsymbol{M}$;

2. Solving Eq. (5.4) by the hybrid algorithm:

   Given the initial condition of the target RC frame, i.e., $\boldsymbol{p}^0$, $\boldsymbol{u}^0$, and $\dot{\boldsymbol{u}}^0$, $\boldsymbol{f}_S(\boldsymbol{u}^0)$, and known information from step 1;

   (a) calculate the $\ddot{\boldsymbol{u}}^0 = \boldsymbol{M}^{-1}\left(\boldsymbol{p}^0 - \boldsymbol{C}\dot{\boldsymbol{u}}^0 - \boldsymbol{f}_S(\boldsymbol{u}^0)\right)$;

   for t = 1 to T **do**

   (a) $\Delta\hat{\boldsymbol{p}}^{t-1} = \boldsymbol{p}^t - \boldsymbol{p}^{t-1} + \boldsymbol{A}\dot{\boldsymbol{u}}^{t-1} + \boldsymbol{B}\ddot{\boldsymbol{u}}^{t-1}$;

   (b) $\hat{\boldsymbol{K}}^{t-1} = \boldsymbol{K}^{t-1} + 2\boldsymbol{C}/\Delta t + 4\boldsymbol{M}/(\Delta t)^2$;

   (c) calculate the $\Delta\boldsymbol{u}^{t-1}$, $\boldsymbol{K}^t$, $\boldsymbol{f}_s(\boldsymbol{u}^t)$ using modified Newton-Raphson and *algorithm 1*

   Given $\boldsymbol{f}_s(\boldsymbol{u}^{t-1})$, $\boldsymbol{u}^{t-1}$; $\Delta\hat{\boldsymbol{p}}^{t-1}$, $\hat{\boldsymbol{K}}^{t-1}$, $\boldsymbol{K}^{t-1}$, maximum number of iteration N, and tolerance *tol*

   (a) initial assignment: $\boldsymbol{f}_s(\boldsymbol{u}_0^t) = \boldsymbol{f}_s(\boldsymbol{u}^{t-1})$, $\boldsymbol{u}_0^t = \boldsymbol{u}^{t-1}$, $\Delta\boldsymbol{R}_1 = \Delta\hat{\boldsymbol{p}}^{t-1}$, $\hat{\boldsymbol{K}} = \hat{\boldsymbol{K}}^{t-1}$, $\boldsymbol{K} = \boldsymbol{K}^{t-1}$;

   for j_n = 1 to N **do**

   (a) $\Delta\boldsymbol{u}_{j_n} = \hat{\boldsymbol{K}}^{-1}\Delta\boldsymbol{R}_{j_n}$;

   (b) $\boldsymbol{u}_{j_n}^t = \boldsymbol{u}_{j_n-1}^t + \Delta\boldsymbol{u}_{j_n}$;

   (c) calculate the $\boldsymbol{K}_{j_n}^t$ and $\boldsymbol{f}_s\left(\boldsymbol{u}_{j_n}^t\right)$ using the algorithm 1: $\left[\boldsymbol{f}_s\left(\boldsymbol{u}_{j_n}^t\right), \boldsymbol{K}_{j_n}^t\right] = Algorithm5.1\left(\boldsymbol{u}_{j_n}^t\right)$;

   (d) $\Delta\boldsymbol{f}_{j_n} = \boldsymbol{f}_s\left(\boldsymbol{u}_{j_n}^t\right) - \boldsymbol{f}_s\left(\boldsymbol{u}_{j_n-1}^t\right) + \left(\hat{\boldsymbol{K}} - \boldsymbol{K}\right)\Delta\boldsymbol{u}_{j_n}$;

   (e) $\Delta\boldsymbol{R}_{j_n+1} = \Delta\boldsymbol{R}_{j_n} - \Delta\boldsymbol{f}_{j_n}$;

   (f) calculate the displacement convergence criterion: $\Delta\boldsymbol{u} = \sum_{i_n=1}^{j_n}\Delta\boldsymbol{u}_{i_n}$, $eps = \left\|\Delta\boldsymbol{u}_{j_n}\right\|/\|\Delta\boldsymbol{u}\|$

   (h) $\Delta\boldsymbol{u}^{t-1} = \Delta\boldsymbol{u}$, $\boldsymbol{K}^t = \boldsymbol{K}_{j_n}^t$, and $\boldsymbol{f}_s(\boldsymbol{u}^t) = \boldsymbol{f}_s\left(\boldsymbol{u}_{j_n}^t\right)$;

   if $eps \le tol$ **do**

   (a) break the loop;

   end if

   end for j_n

   (d) $\Delta\dot{\boldsymbol{u}}^{t-1} = 2\Delta\boldsymbol{u}^{t-1}/\Delta t - 2\dot{\boldsymbol{u}}^{t-1}$;

   (e) $\Delta\ddot{\boldsymbol{u}}^{t-1} = 4\Delta\boldsymbol{u}^{t-1}/(\Delta t)^2 - 4\dot{\boldsymbol{u}}^{t-1}/\Delta t - 2\ddot{\boldsymbol{u}}^{t-1}$;

   (f) $\boldsymbol{u}^t = \boldsymbol{u}^{t-1} + \Delta\boldsymbol{u}^{t-1}$, $\dot{\boldsymbol{u}}^t = \dot{\boldsymbol{u}}^{t-1} + \Delta\dot{\boldsymbol{u}}^{t-1}$, and $\ddot{\boldsymbol{u}}^t = \ddot{\boldsymbol{u}}^{t-1} + \Delta\ddot{\boldsymbol{u}}^{t-1}$;

   end for t

By implementing **Algorithm 5.2**, the time-history response quantities of interest, such as time versus roof displacement and the distribution of peak story drift ratio along the floors can be obtained. It should be noted that the displacement convergence criterion in **Algorithm 5.2** for the proposed MDOF model is satisfactory. This is because the numerical values in the displacement vector have the same units and do not suffer the problems associated with inconsistent units that bring in significant errors (Chopra 2007).

5.3 Development of a Training Set

For the purpose of this study, the rectangular RC column dataset presented in Chapter III (see Appendix A) is used to evaluate the performance of the novel system-level, data-driven framework in predicting the seismic response of RC frames under both displacement-controlled quasi-static cyclic loading and dynamic ground motions. Additionally, as shake table tests for large RC frames with several stories and bays (e.g., RC frame with more than 6 stories and 2 bays) are not available, shake table tests for smaller RC frames will be used. Since small RC frames have column features outside the range of the dataset, the dataset presented in Chapter III (see Appendix A) is supplemented with 20 small-scale RC column specimens to reduce potential sample bias. These column specimens are taken form Cecen (1979). Thus, the final number of column specimens in the dataset is 272.

The nine optimal critical parameters employed to define a hysteretic modeler for each of the 20 columns in the dataset are obtained according to the method presented in Chapter III. The statistical properties of the optimal cyclic backbone curve and three hysteretic parameters for the 272 column specimens are summarized in Table 5.1.

Table 5.1 Statistical properties of the optimal cyclic backbone curve and hysteretic parameters.

Critical Parameters	Minimum	Maximum	Median	Mean	Std.Dev
Yield shear force, V_y (kN)	1.60	1071.01	130.50	163.72	149.05
Drift ratio at yield shear, δ_y (%)	0.20	1.73	0.79	0.85	0.37
Maximum shear force, V_m (kN)	1.84	1338.80	155.09	194.63	178.50
Drift ratio at maximum shear, δ_m (%)	0.31	7.94	1.69	1.99	1.33
Ultimate shear force, V_u (kN)	1.64	1217.01	126.89	163.03	155.51
Drift ratio at ultimate shear, δ_u (%)	0.72	9.39	3.15	3.60	1.88
Stiffness deterioration parameter, α	0.30	119.42	9.37	21.09	21.98
Strength deterioration parameter, β	0.00	0.93	0.06	0.14	0.20
Pinching parameter, γ	0.31	1.00	0.98	0.87	0.19

5.4 Numerical Results

This section presents the numerical experiments carried out to validate the proposed data-driven framework in generalized seismic response prediction of RC frame structures under displacement-controlled quasi-static cyclic loading and dynamic shake table tests. For the displacement-controlled quasi-static cyclic loading test, a large-scale (1:2) physical experimental model of a 3-bay, 3-story RC frame structure is selected from Xie et al. (2015) to serve as the test specimen. For the dynamic shake table test, two small-scale (1:15) physical experimental models of 3-bay, 9-story RC frame structures are selected from Schultz (1986) to serve as the test specimens. One is subjected to four earthquake (EQ) ground motions and another is subjected to six EQ ground motions. In each case, the proposed approach is compared with the widely used distributed plasticity fiber model based on experimental data. All the numerical experiments are performed using a Desktop PC with the Processor: Intel(R) Xeon(R) CPU E3-1270 v6 @ 3.80 GHz.

5.4.1 Displacement-controlled quasi-static cyclic loading tests

This section presents a comparison between the proposed system-level data-driven framework and the widely used fiber model to demonstrate the real-world application and full potential of the proposed approach. To validate the superiority of the novel framework, the classic fiber beam-column element is utilized to model the nonlinear cyclic response of the RC frame. For the proposed framework, the locally weighted least-squares support vector machine for regression (LWLS-SVMR) presented in **Section 4.3** is selected as the ML technique. The predictors used in **Section 4.4** are also utilized here. The response variables are those presented in Table 5.1.

A large-scale (1:2) physical experiment of a 3-bay, 3-story RC frame subjected to displacement-controlled quasi-static cyclic loading is selected from Xie et al. (2015) for this comparison. The lateral load distribution for this experimental test is an inverse triangle, and the

entire loading process is controlled by the displacement of the top floor (i.e., roof displacement). The detailed information regarding the structural geometry, material properties, reinforcement details, and load pattern can be found in Xie et al. (2015). For the widely used distributed plasticity fiber model, a single force-based fiber beam-column element (Spacone et al. 1996a; 1996b) with five Gauss-Lobatto integration points (i.e., monitoring sections) is employed to model each of the columns and beams in the selected RC frame. In each monitoring section, cover concrete fiber is simulated using the modified Kent and Park model (Scott et al. 1982), and the core concrete fiber is simulated by the confined concrete model proposed by Mander et al. (1988) to represent the confinement effect of the stirrups. The reinforcement fiber is modeled by the Menegotto-Pinto model (Menegotto and Pinto 1973). OpenSees (Mazzoni et al. 2006) is used to implement the RC frame numerical model. For the proposed data-driven framework, all the beams are assumed rigid in axial and flexure as introduced in **Section 5.2.2**. All the columns in the frame are first expressed by predictors as query points where the response variables need to be predicted. Then, for each query point, the LWLS-SVMR is used to predict the response variables based on the 272 training data points presented in **Section 5.3** (see Appendix A). **Algorithm 5.1** is used to implement the proposed approach using Matlab 2018a. By implementing **Algorithm 5.1**, the hysteretic responses (roof displacement versus base shear and story drift ratio versus story shear) are produced.

Figure 5.4 presents a comparison of the results between the proposed framework and traditional physics-based modeling techniques where ground truth is defined as the experimental tests. Figure 5.4(a-b) demonstrates that both methods reasonably capture the global nonlinear response of the RC frame in terms of the hysteretic relation of roof displacement versus base shear. The proposed approach effectively reflects the cyclic strength deterioration and softening behavior observed experimentally, while the fiber model fails to reasonably capture these types of behavior.

Although both methods reasonably predict the overall hysteretic response, the proposed approach achieves better prediction than the fiber model. The hysteretic curve predicted by the proposed approach has better agreement with the experimental results than that simulated by the fiber model. The story drift ratio versus story shear is extracted and presented in Figure 5.4(c-h). Both methods reasonably predict the lateral capacity of the RC frame, where the lateral strength (i.e., maximum shear force) predicted by both methods are close to those observed experimentally. However, the fiber model still does not reasonably capture the softening behavior induced by cyclic strength deterioration, while the proposed approach can effectively reflect these types of behavior characteristics observed experimentally. In total, the proposed approach can reasonably reflect the hysteretic behavior of the RC frame. The hysteretic curves for each story predicted by the proposed approach have a reasonable agreement with experimental results as shown in Figure 5.4(c,e,g). The story behaviors predicted by the fiber model show some discrepancy with the experimental results, as shown in Figure 5.4(d,f,h).

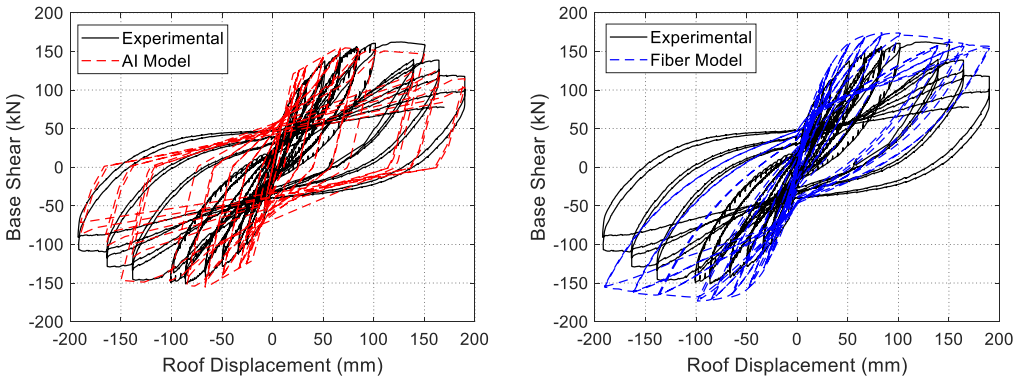


Figure 5.4 Comparison of results between the proposed AI-enhanced framework, experimental data, and widely-used traditional model (i.e., Fiber Model) for the selected RC frame

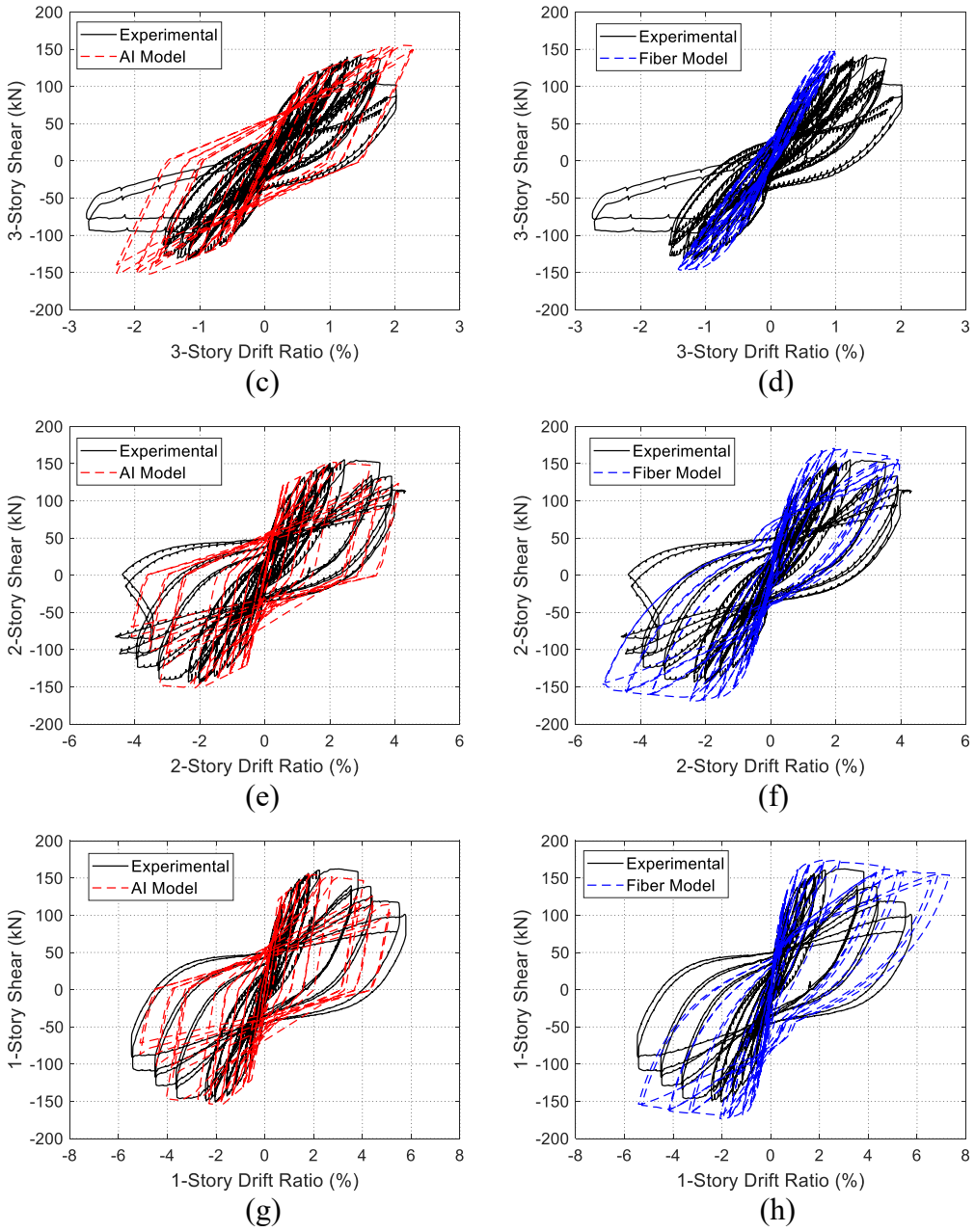


Figure 5.4 Continued.

Perhaps most importantly, the computational time for predicting the hysteretic curve of the selected RC frame using the proposed method only requires *10* seconds, while using the traditional fiber model takes *1,672* seconds (or roughly 30 minutes). Therefore, the proposed approach significantly reduces the computational cost. Based on these comparisons, the proposed approach

presented in this section performs better than the traditional physics-based method. Thus, it is deemed that the proposed approach is the most appropriate means for seismic response prediction of RC frames subjected to reversed cyclic loading, especially for application in near-real-time scenarios.

5.4.2 Dynamic shake table tests

To validate the performance of the proposed framework in predicting the seismic response of RC frames subjected to ground motions, two small-scale (1:15) 3-bay, 9-story RC frame specimens – structure SS1 subjected to four consecutive unidirectional ground motions and structure SS2 subjected to six consecutive unidirectional ground motions – are used as illustrative examples. These shake table tests were organized by Schultz (1986). The difference between these two test specimens is that the columns in frame SS2 have a higher longitudinal reinforcement ratio than those in frame SS1. The detailed information regarding the physical experimental set-up, structural features, ground motions, and shake table test results can be found in Schultz (1986).

For traditional physics-based modeling approaches, the fiber beam-column element is also used to model the seismic response of the two small-scale RC frames. The element type, integration method, number of integration points, and material constitutive models described in **Section 5.4.1** for the large-scale RC frame are also used here to establish the numerical models of the two RC frames. For the proposed approach, all the beams of these specimens are assumed rigid in axial and flexure, and the columns in the frames are first expressed as query points by the predictors presented in **Section 4.4**. Then, the LWLS-SVMR is used to predict the response variables based on the 272 training specimens introduced in **Section 5.3** (see Appendix A), forming the hysteretic modeler for each column. Finally, the established hysteretic modelers are incorporated into **Algorithm 5.2** for dynamic response prediction. For both approaches, a damping

ratio of 2% is assigned to the first two modes of both frames, and the time step is set to the one recorded in the ground motions (i.e., 0.005s). Since these two RC frames are not repaired after each ground motion (Schultz 1986), the four ground motions for frame SS1 and the six ground motions for frame SS2 are grouped to be a sequential ground motion that serves as the input ground motion. OpenSees is used to perform the time-history procedures of two fiber models. Matlab 2018a is used to implement **Algorithm 5.2** as presented in **Section 5.2.3** to perform the time-history procedures of the two MDOF models. The input ground motions for frames SS1 and SS2 are presented in Figures 5.5 and 5.6, respectively. Note that frame SS1 collapsed under EQ4, and thus, only the first 2.75s of the experimental results are recorded (Schultz 1986). The time-history results regarding the time versus roof displacement and the floors versus peak story drift ratio are presented in Figures 5.7-5.10.

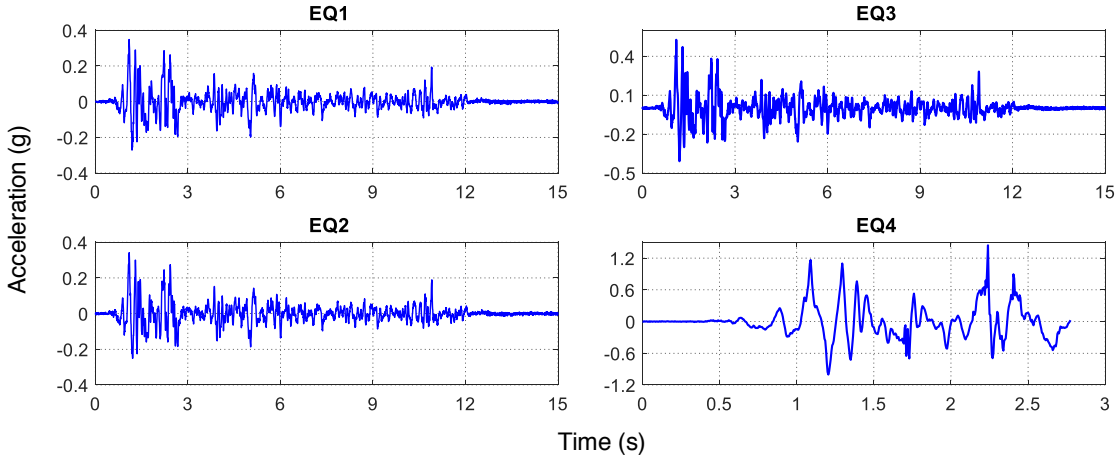


Figure 5.5 Four time versus ground accelerations for frame SS1.

Figure 5.7 presents the comparison of the predicted time-roof displacement results for frame SS1 between the fiber model and the proposed method, with the experimental data serving as the ground truth. By observation, the proposed method achieves better agreement with the

experimental data for all four EQs over the full-time history. Further, the proposed approach nearly captures the peak roof displacements for all four EQs, while the fiber model underestimates those peak roof displacements. Peak story drift ratio is an important engineering demand parameter (EDP) which is typically used to quantify the seismic performance of an RC structure (Bracci et al. 1997; Chopra 2007; Moehle and Deierlein 2004). Figure 5.8 shows the results of floor versus peak story drift ratio for frame SS1. It can be seen that the proposed approach performs better than the fiber model, where the peak story drift ratios predicted by the proposed approach at each floor for all four EQs have closer agreement with the experimental results than those predicted by the fiber model.

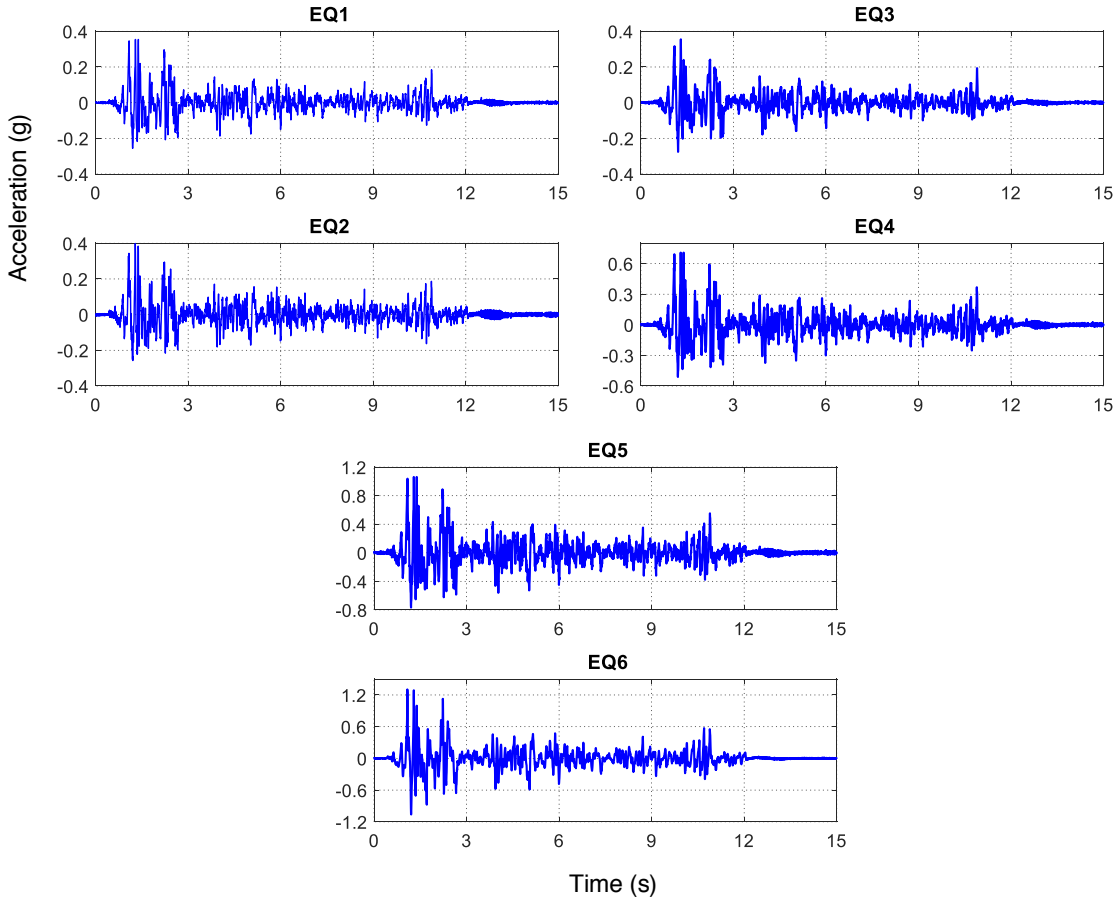


Figure 5.6 Six time versus ground accelerations for frame SS2.

A similar trend is observed by the comparison of the results of the predicted time-roof displacement for frame SS2, as shown in Figure 5.9. The proposed approach also accurately captures the peak roof displacements for all six EQs, while the fiber model underestimates these values. Additionally, for the comparison of the predicted peak drift ratios at the second through ninth floors, the proposed method shows better agreement with the experimental data for all six EQs than the fiber model (Figure 5.10). However, for the predicted peak drift ratios at the first floor, compared to the proposed method, the fiber model achieves a closer agreement with the experimental results for EQ3 through EQ5 and has a comparable performance for EQ1, EQ2, and EQ6. Further, both the fiber model and the proposed approach show discrepancy with the experimental results for the predicted peak drift ratios at the first and seventh through ninth floors for EQ5 and EQ6, where the PGA for EQ5 is 1.06g and for EQ6 is 1.30g. This is because under extreme seismic intensities, the behavior of frame SS2 becomes more irregular, and modes other than the first are seen to have a greater effect on displacement response, as discussed in Schultz (1986). Both the fiber model and the proposed method consider the first two modes more than others, finally leading to significant errors. In addition, for the proposed approach, the number of small-scale column specimens in the training dataset is only 20 and may not be sufficient to eliminate the potential sample bias. Thus, this causes the predicted cyclic backbone curve and hysteretic parameters of the columns in the small-scale RC frame SS2 to be imprecise. Nevertheless, in most cases, the proposed system-level data-driven framework still achieves better agreement with the experimental data than the traditional fiber model.

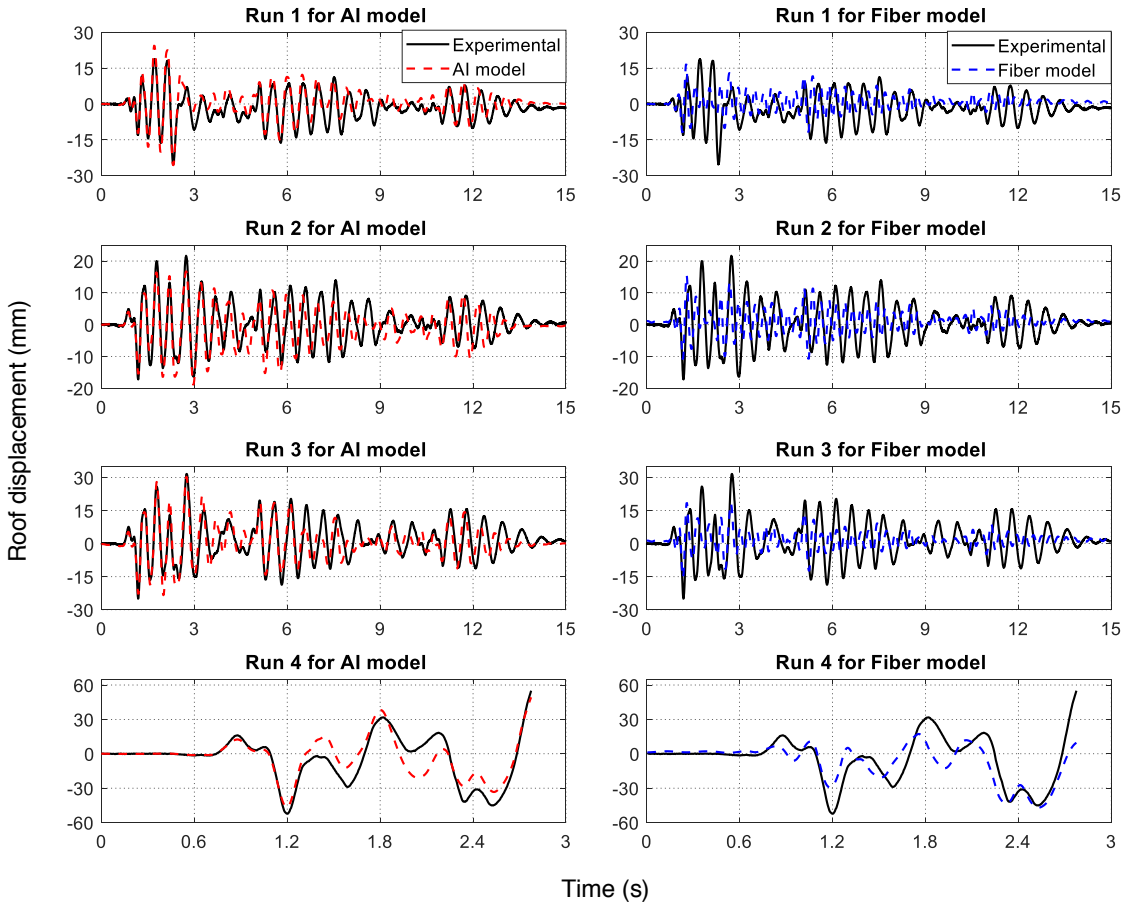


Figure 5.7 Time vs. roof displacement results for the traditional approach (i.e., fiber model) and the proposed AI model, with the experimental data serving as the ground truth.

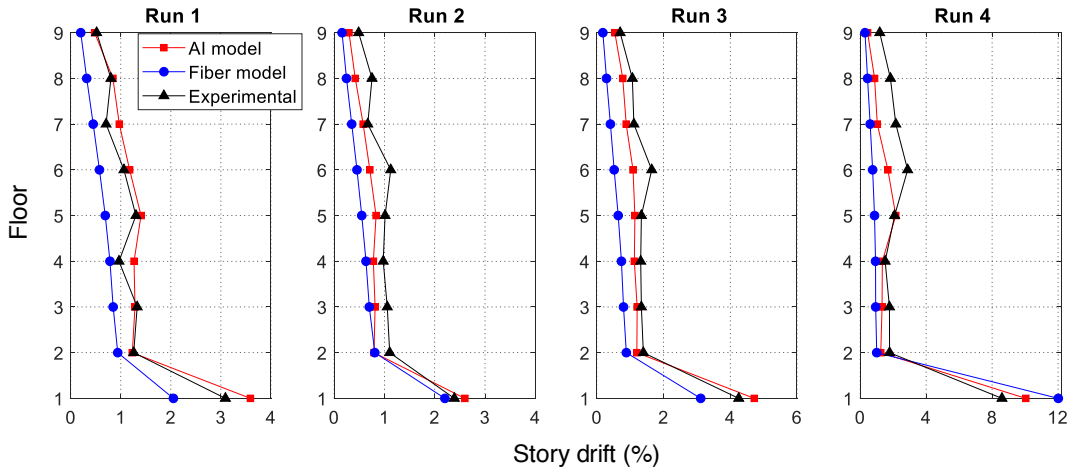


Figure 5.8 Distribution of peak story drift ratio along the floors for the traditional approach (i.e., fiber model) and the proposed AI model, with the experimental data serving as the ground truth.

More importantly, the computational time for all ground motions using the proposed approach only requires *133* (SS1) and *289* (SS2) seconds. This time is substantially diminished when compared to the fiber models which took *972* (SS1) and *1,942* (SS2) seconds. Thus, the proposed approach significantly enhances the computational efficiency while still maintaining (and in most cases, even improving) good prediction performance. Further, the fiber model is implemented using OpenSees, which is developed using compiled language (i.e., C++), while the proposed approach is implemented using Matlab, which is an interpreted language. Thus, OpenSees is inherently faster than Matlab. However, the proposed approach is still much more efficient than the fiber model. Based on these comparisons, the proposed approach presented in this chapter performs significantly better than the traditional method for all seismic response quantities and agrees better with the experimental data.

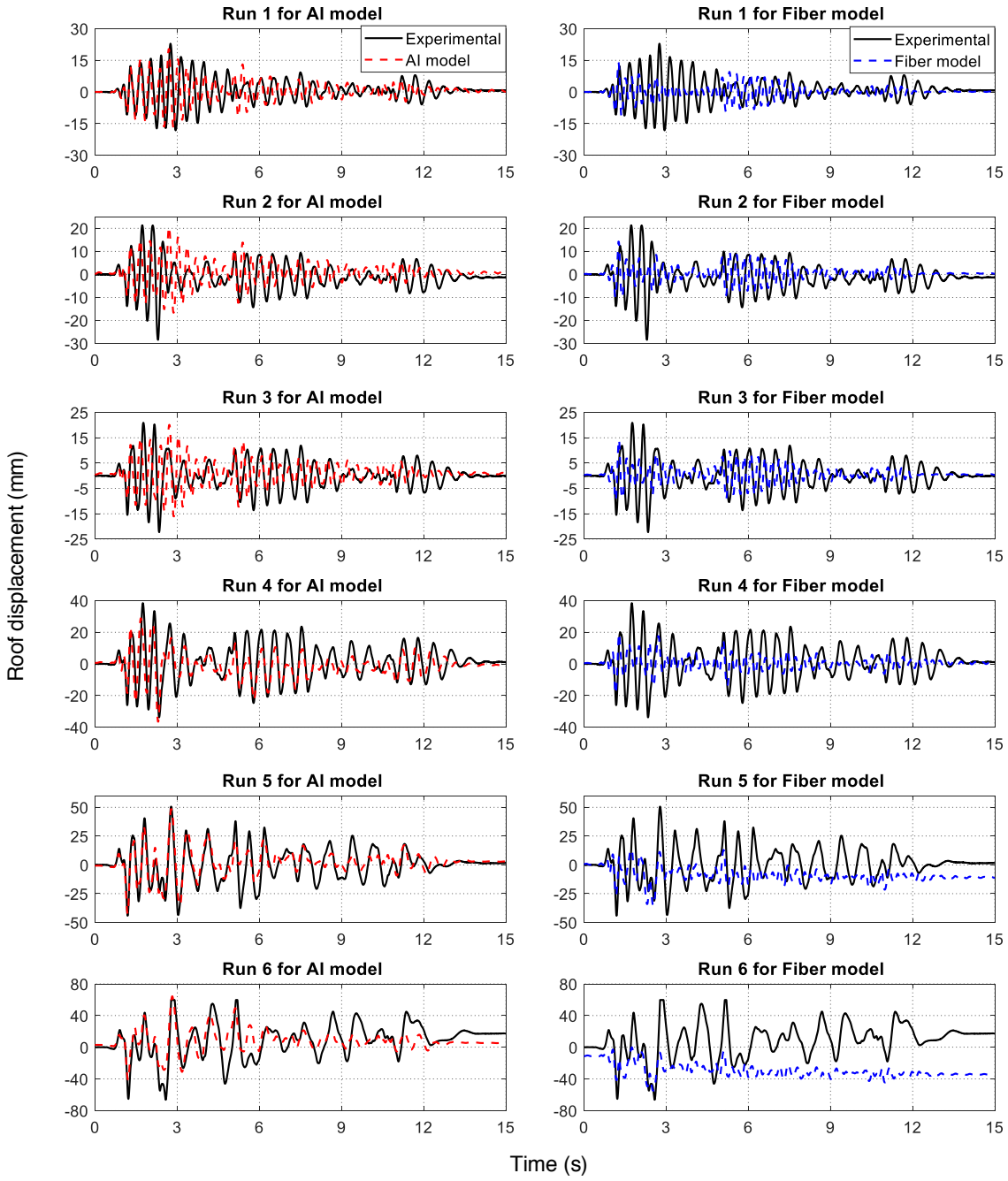


Figure 5.9 Time vs. roof displacement results for the traditional approach (i.e., fiber model) and the proposed AI model, with the experimental data serving as the ground truth.

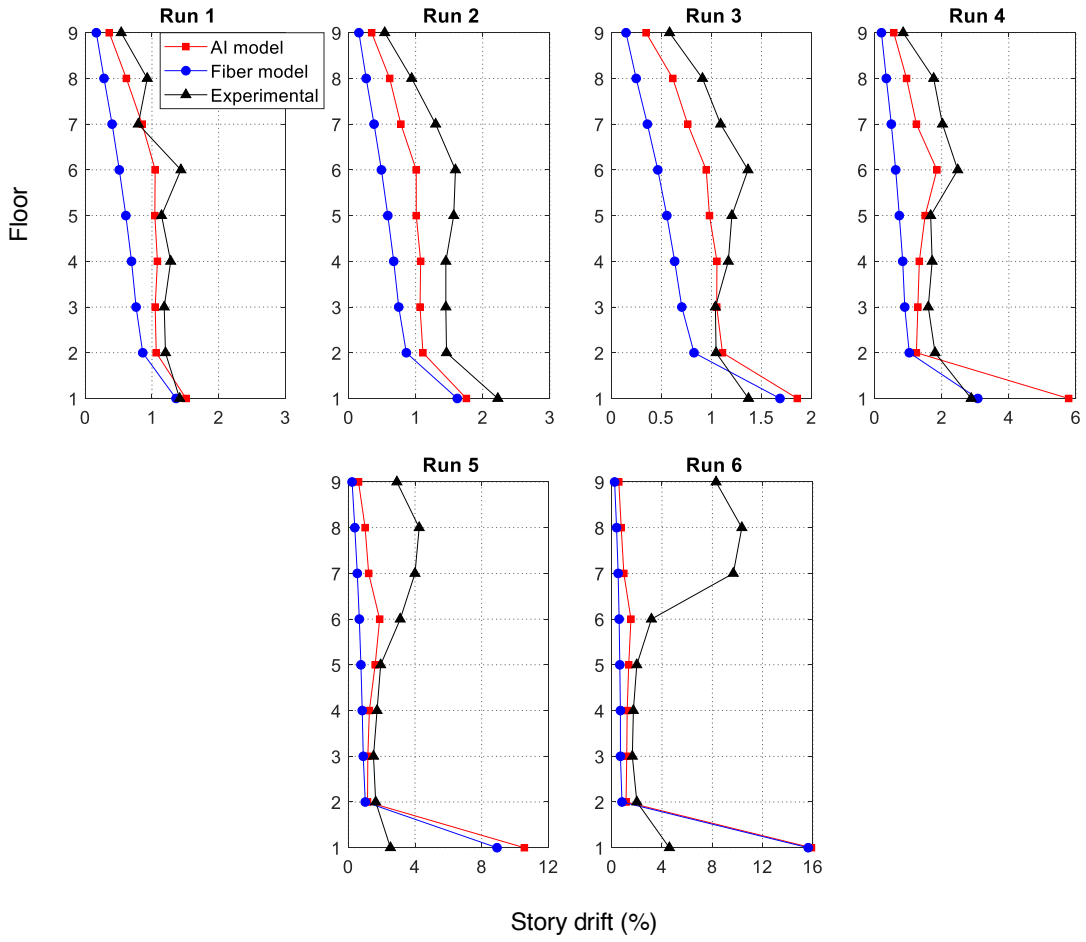


Figure 5.10 Distribution of peak story drift ratio along the floors for the traditional approach (i.e., fiber model) and the proposed AI model, with the experimental data serving as the ground truth.

5.4.3 Discussion of results

From the above results for both displacement-controlled quasi-static cyclic loading and shake table tests, it can be concluded that the proposed system-level computational framework outperforms the widely-used distributed plasticity fiber model in terms of both prediction capability and computational efficiency. In addition, since the data-driven computing procedures are initiated at the component level, the physical experiment data for RC columns is used in developing the hysteretic modeler for each column. Therefore, it is expected that the hysteretic modeler can reasonably reproduce the experimental hysteretic behavior of each column in the target RC frames. However, a theoretical assumption is made at the system level, where the shear building model is used to translate the responses at the component level to the structural response at the system level. Ultimately, it was found that this assumption does not have a significant influence on the predictive performance of the proposed approach. This may be attributed to the phenomenon that the beams in the RC frames utilized are stiffer than the columns, and thus, the shear building model is appropriate for this case.

5.5 Summary

A novel system-level data-driven framework is proposed in this section to predict the seismic response of RC structural systems under displacement-controlled quasi-static cyclic loading and shake table tests. The proposed system-level computational framework is a hybrid ML-physics based approach, which incorporates a novel component-level data-driven framework presented in **Section 4.4** with a shear-building model. This integration efficiently leverages the advantages of both approaches. Two data-driven seismic response solvers are developed to implement the proposed approach for the seismic response prediction of RC structural systems under displacement-controlled quasi-static cyclic loading and shake table tests. To validate the performance of the proposed approach, RC frames are selected as illustrative examples. The numerical results validate that the proposed approach outperforms the widely-used traditional modeling approaches in predicting the seismic response of RC frames under both quasi-static cyclic loading and shake table tests. Moreover, the proposed method significantly enhances the computational efficiency for both cases in comparison with the widely-used traditional approaches, yielding great potential for regional seismic risk quantification and other near-real-time needs.

CHAPTER VI

SOLUTIONS TO DATA-RELATED PROBLEMS*

6.1 Overview

Machine learning (ML) methods have high requirements for the input data in order to achieve high generalization performance. The input data must be high-quality and sufficiently large in size. Otherwise, once trained, the ML models will not be able to accurately predict the target response variables. In real-world scenarios, datasets are most likely corrupted by outliers, contain missing values, and may not be sufficient in size, leading to large sample biases. These data-related problems will significantly degrade the generalization performance of ML methods as introduced in **Section 2.4**, and thus negatively affect the performance of the proposed data-driven frameworks presented in **Chapters IV and V**. This chapter presents novel computational methods which were created to deal with such data-related problems, yielding data-driven frameworks that are extremely robust. First, a novel locally weighted ML model is developed to eliminate the negative effect induced by outliers. Second, a new multiple imputation (MI) method is proposed to deal with missing data problems. Lastly, a novel regression-based transfer learning (TL) method is developed to reduce the negative effect of small sample bias. Each method is assessed and validated by comparing the numerical results with physical experiment data. With the help of these new computational methods, the proposed data-driven frameworks will have good generalization performance even if the dataset is plagued with any or all of these kinds of issues. The detailed information is given below.

*Section 6.4 of this chapter is reprinted with permission from "Reducing the effect of sample bias for small data sets with double-weighted support vector transfer regression" by Huan Luo and Stephanie Paal, 2020. *Computer-Aided Civil and Infrastructure Engineering*, 1-16, Copyright [2020] by John Wiley and Sons.

6.2 Solution to Dataset Corrupted by Outliers

As introduced in **Section 2.4.1**, standard ML methods can be negatively affected by a dataset corrupted by outliers. This section presents a novel ML approach for constructing data-driven procedures that are robust to input data which is corrupted by outliers. The novel ML approach is an extension of locally weighted least squares support vector machines for regression (LWLS-SVMR) presented in **Section 4.3** and thus is called robust LWLS-SVMR (RLWLS-SVMR). A significant drawback of LWLS-SVMR is that it is sensitive to outliers close to query points. To solve this shortcoming, an extra weight that is a function of residuals is introduced into the reformulation of LWLS-SVMR to form RLWLS-SVMR. The major advantage of the proposed method over LWLS-SVMR is not only that it is robust to input data contaminated by various types of outliers (i.e., extreme and non-extreme outliers) but also that it maintains the local nature, where, in order to predict a query point, the entire set of training data does not need to be fit. Instead, it only requires the fitting of a subset of training data nearby (relevant to) the query point. These characteristics yield a model that both overcomes the negative interference of outliers and avoids the potential influence of irrelevant points, achieving a suitable trade-off between the capacity of the learning system and the number of training data points. The development of the proposed RLWLS-SVMR is presented in the next sub-section.

6.2.1 Development of RLWLS-SVMR

Assume a multi-dimensional training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ is collected from a domain of interest and some observations (i.e., data points) have been corrupted by outliers. For the remainder of this section, the following notations are utilized. Let R be the real numbers set; $\boldsymbol{x}_i \in R^p$ is a row vector with p dimensions (i.e., p variables) which can be written as $\boldsymbol{x}_i = \left(x_{i1}, \ldots, x_{ip}\right)$, and $\boldsymbol{x}_i' \in R^p$ represents the transpose of $\boldsymbol{x}_i$ and is a column vector with p dimensions which can be written as

$\boldsymbol{x}_i' = \left(x_{i1}, \ldots, x_{ip}\right)^T$; $y_i \in R$ is a real number; $\boldsymbol{X} \in R^{n \times p}$ is an $n \times p$ matrix which can be written as $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$; the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is an $n \times (p+1)$ matrix which includes n data points and each data point contains p explanatory variables (i.e., $\boldsymbol{x}_i \in R^p$) and one response (i.e., $y_i \in R$).

Given an independent test set $\left\{\left(\boldsymbol{x}_q, \hat{y}_q\right)\right\}_{q=1}^m$ that is not included in the training set, for each query point $\boldsymbol{x}_q, q = 1, \ldots, m$, where the response values $\hat{y}_q$ are to be predicted and thus not considered in the following process. The basic procedure of the RLWLS-SVMR is as follows:

(1) Define a subset $\left\{\left(\boldsymbol{x}_{(s)}, y_{(s)}\right)\right\}_{s=1}^r$ from the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ by a parameter f_q, where f_q can take any value in the range $(0, 1]$; the number of data points r in the subset is equivalent to $Ceil\left(f_q n\right)$, and the points in the subset are determined and sorted by the Euclidean distance metric via the following procedure:

(2) Calculate the Euclidean distance from each data point in the training set to each query point $\left\|\boldsymbol{x}_i - \boldsymbol{x}_q\right\|, i = 1, \ldots, n; q = 1, \ldots, m$, so for each query point, there is a distance vector $\boldsymbol{d}_q = \left(d_{q1}, \ldots, d_{qn}\right), q = 1, \ldots m$;

(3) Sort the entries in each distance vector increasingly so a new sorted distance vector $\boldsymbol{d}_{(q)} = \left(d_{(q1)}, \ldots, d_{(qn)}\right), q = 1, \ldots m$ is obtained;

(4) The data points in the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, corresponding to the first r entries in the sorted distance vector $\boldsymbol{d}_{(q)}$ (i.e., $d_{(q1)}, \ldots, d_{(qr)}$), can be selected as the subset $\left\{\left(\boldsymbol{x}_{(s)}, y_{(s)}\right)\right\}_{s=1}^r$. Note: for different query points, the subset may vary.

(5) After the subset is determined, the learning objective of the RLWLS-SVMR is to find $\boldsymbol{w}' = (w_1, w_2, \ldots, w_h)^T \in R^h$ and $b \in R$ that minimize the following objective function:

$$J(\boldsymbol{w}', e_s) = \frac{1}{2}(\boldsymbol{w}')^T \boldsymbol{w}' + \frac{1}{2}\gamma_q \sum_{s=1}^{r} v_q(\boldsymbol{x}_{(s)})\beta_q(\boldsymbol{x}_{(s)})e_s^2, q = 1, \dots, m \quad (6.1)$$

Subject to: $y_{(s)} = (\boldsymbol{w}')^T \varphi(\boldsymbol{x}'_{(s)}) + b + e_s, s = 1, \dots, r \quad (6.2)$

where $e_s \in R, s = 1, \dots, r$ is the error term; $\gamma_q \in R, q = 1, \dots, m$ is a regularization parameter; $\beta_q(\boldsymbol{x}_{(s)}), v_q(\boldsymbol{x}_{(s)}) \in R, s = 1, \dots, r; q = 1, \dots, m$ are weights that can take any value in the range $[\varepsilon, 1]$, $\beta_q(\boldsymbol{x}_{(s)})$ is a function of Euclidean distance where data points in a subset close to a query point have larger weights and far away from the query point have smaller weights; $v_q(\boldsymbol{x}_{(s)})$ is a function of residual where data points in a subset around the query point having large residuals have smaller weights and having small residuals have larger weights; $\varepsilon \in R$ is a real number approaching 0; $\varphi(\boldsymbol{x}'_{(s)})$ is a feature vector, and $\varphi(\cdot): R^p \to R^h$ is a mapping function from p dimensions to a higher h-dimensional feature space. Note: $\boldsymbol{x}'_{(s)}$ is a column vector, thus $\varphi(\boldsymbol{x}'_{(s)})$ is also a column vector.

If $\beta_q(\boldsymbol{x}_{(s)})$ takes a value approaching ε, it means the point $(\boldsymbol{x}_{(s)}, y_{(s)})$ is far away from the query point $(\boldsymbol{x}_q, \hat{y}_q)$ (relatively large Euclidean distance) and plays a lesser role in the determination of $\hat{y}_q$; while, if $\beta_q(\boldsymbol{x}_{(s)})$ takes a value approaching one, it means the point $(\boldsymbol{x}_{(s)}, y_{(s)})$ is close to the query point $(\boldsymbol{x}_q, \hat{y}_q)$ (relatively small Euclidean distance) and plays an important role in the determination of $\hat{y}_q$.

(6) The Lagrangian function is established to solve Eq. (6.1) and Eq. (6.2):

$$L(\boldsymbol{w}', b, e_s; \alpha_s) = J(\boldsymbol{w}', e_s) - \sum_{s=1}^{r} \alpha_s\{(\boldsymbol{w}')^T \varphi(\boldsymbol{x}'_{(s)}) + b + e_s - y_{(s)}\} \quad (6.3)$$

where $\alpha_s \in R, s = 1, \dots, r$ is a Lagrange multiplier (also called support values).

The Karush-Kuhn-Tucker (KKT) conditions for optimality are used by differentiating the variables in Eq. (6.3) above, which results in the following:

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{w}'} = 0 \rightarrow \boldsymbol{w}' = \sum_{s=1}^{r} \alpha_s \varphi(\boldsymbol{x}'_{(s)}) \\ \dfrac{\partial L}{\partial b} = 0 \rightarrow 0 = \sum_{s=1}^{r} \alpha_s \\ \dfrac{\partial L}{\partial e_s} = 0 \rightarrow e_s = \dfrac{\alpha_s}{\gamma_q v_q(\boldsymbol{x}_{(s)}) \beta_q(\boldsymbol{x}_{(s)})},\ s = 1,\ldots,r;\ q = 1,\ldots,m \\ \dfrac{\partial L}{\partial \alpha_s} = 0 \rightarrow y_{(s)} = (\boldsymbol{w}')^T \varphi(\boldsymbol{x}'_{(s)}) + b + e_s,\ s = 1,\ldots,r \end{cases} \tag{6.4}$$

Rearranging Eq. (6.4) and eliminating $\boldsymbol{w}'$ and e_s, using the kernel function to replace the inner product of the feature vectors, the following matrix equation can be obtained:

$$\begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & K(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(1)}) + \dfrac{1}{\gamma_q v_q(\boldsymbol{x}_{(1)}) \beta_q(\boldsymbol{x}_{(1)})} & K(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(2)}) & \cdots & K(\boldsymbol{x}_{(1)},\boldsymbol{x}_{(r)}) \\ 1 & K(\boldsymbol{x}_{(2)},\boldsymbol{x}_{(1)}) & K(\boldsymbol{x}_{(2)},\boldsymbol{x}_{(2)}) + \dfrac{1}{\gamma_q v_q(\boldsymbol{x}_{(2)}) \beta_q(\boldsymbol{x}_{(2)})} & \cdots & K(\boldsymbol{x}_{(2)},\boldsymbol{x}_{(r)}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & K(\boldsymbol{x}_{(r)},\boldsymbol{x}_{(1)}) & K(\boldsymbol{x}_{(r)},\boldsymbol{x}_{(2)}) & \cdots & K(\boldsymbol{x}_{(r)},\boldsymbol{x}_{(r)}) + \dfrac{1}{\gamma_q v_q(\boldsymbol{x}_{(r)}) \beta_q(\boldsymbol{x}_{(r)})} \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_r \end{bmatrix} = \begin{bmatrix} 0 \\ y_{(1)} \\ y_{(2)} \\ \vdots \\ y_{(r)} \end{bmatrix} \tag{6.5}$$

where $q = 1, \ldots, m$ and the kernel function is $K(\boldsymbol{x}_{(s)}, \boldsymbol{x}_{(t)}) = \varphi^T(\boldsymbol{x}'_{(s)}) \varphi(\boldsymbol{x}'_{(t)}), s = 1, \ldots, r; t = 1, \ldots, r.$

(7) To determine $\beta_q(\boldsymbol{x}_{(s)}) \in R, s = 1, \ldots, r; q = 1, \ldots, m$, for each query point $\boldsymbol{x}_q$, let $d_{(qr)}$ be the distance from $\boldsymbol{x}_q$ to the r^{th} nearest neighbor $\boldsymbol{x}_{(r)}$ (i.e., $d_{(qr)}$ is the maximum distance compared to $d_{(q1)}, \ldots, d_{(q(r-1))}$), and let $\beta_q(\boldsymbol{x}_{(s)}) = T\left(d_{(qr)}^{-1} \left\| \boldsymbol{x}_{(s)} - \boldsymbol{x}_q \right\|\right)$, where $T(\cdot)$ is a tricube weight function (Cleveland 1979), which is defined as the following:

$$T(g) = f(x) = \begin{cases} (1 - |g|^3)^3, & |g| < 1 \\ \varepsilon, & |g| \geq 1 \end{cases} \tag{6.6}$$

where ε can take any values close to 0, and in this work $\varepsilon = 1e-4$ to avoid a zero in the denominator in Eq. (6.5).

The weight $v_q(\boldsymbol{x}_{(s)})$ in Eq. (6.5) is associated with the robustness to outliers close to a query point, and the determination of $v_q(\boldsymbol{x}_{(s)}) \in R, s = 1, \ldots, r; q = 1, \ldots, m$, for

each query point $\boldsymbol{x}_q$ is discussed in detail in the next section. The initial values of $v_q(\boldsymbol{x}_{(s)})$ are set to one. Note that when $v_q(\boldsymbol{x}_{(s)}) = 1, s = 1, \dots, r; q = 1, \dots, m$ and the values are not updated, the proposed RLWLS-SVMR reverts to LWLS-SVMR.

(8) After solving Eq. (6.5) (Suykens et al. 1999; 2002), the Lagrange multiplier $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_r)$ and b can be obtained, which can then be utilized to predict the query point $\boldsymbol{x}_q$ using the following:

$$\hat{y}(\boldsymbol{x}_q) = \sum_{s=1}^{r} \alpha_s K(\boldsymbol{x}_q, \boldsymbol{x}_{(s)}) + b \tag{6.7}$$

The RBF kernel is utilized, which is defined as follows:

$$K(\boldsymbol{x}_q, \boldsymbol{x}_{(s)}) = exp\left(-\frac{\|\boldsymbol{x}_q - \boldsymbol{x}_{(s)}\|_2^2}{2\sigma_q^2}\right) \tag{6.8}$$

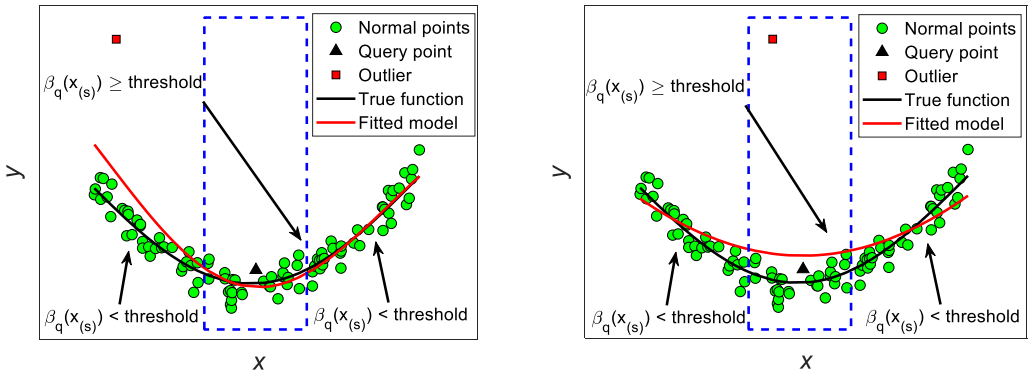
6.2.2 Detection of negative effects due to outliers

In comparison to data-driven procedures established by global ML approaches, where the outliers in the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ must first be detected and removed (Rousseeuw and Leroy 1987) or robust global ML methods must be employed directly for the entire training set, the proposed RLWLS-SVMR is a robust, local ML model. In this sense, all the points of the training set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ are not necessarily considered in the training procedure for prediction of an individual query point $\boldsymbol{x}_q$. Considering the fact that the outliers are just a small portion of the entire training set, it is possible that outliers only exist in certain regions of the training set rather than being distributed across the entire training set. In this case, the advantage of the proposed RLWLS-SVMR model is distinct. This is because, given a query point $\boldsymbol{x}_q$, the selected subset $\{(\boldsymbol{x}_{(s)}, y_{(s)})\}_{s=1}^r$ around this query point may not contain outliers, or the subset may contain outliers but they are sufficiently far away from the query point (see Figure 6.1) such that the outliers have little negative effect on the prediction of the query point.

Figure 6.1 shows a schematic sketch illustrating how an outlier can affect the prediction of a query point. Figure 6.1(a) shows the case where an outlier (red square point) exists in a selected subset $\{(\boldsymbol{x}_{(s)}, y_{(s)})\}_{s=1}^{r}$ but far away from the query point (black triangular point). Figure 6.1(b) shows the case where an outlier occurs close to the query point. Each point $(\boldsymbol{x}_{(s)}, y_{(s)})$ in this subset has a weight $\beta_q(\boldsymbol{x}_{(s)})$, and points close to the query point have larger weights $\beta_q(\boldsymbol{x}_{(s)})$ while points far away from the query point have smaller weights $\beta_q(\boldsymbol{x}_{(s)})$. In this way, points close to the query point have important contributions to the prediction of the query point, while those far away have little influence. If an outlier is far away from the query point, it is possible that the outlier will yield little negative influence on the prediction of the query point (see Figure 6.1(a)). This means the weight $v_q(\boldsymbol{x}_{(s)})$ in Eq. (6.5) does not need to be updated (i.e., RLWLS-SVMR reverts to LWLS-SVMR), since the outlier does not have a significantly negative effect on prediction. Thus, it is necessary to detect these types of negative effects such that a non-robust local model (i.e., LWLS-SVMR) learned via a subset containing outliers can be reliably employed.

This can be achieved by selecting an appropriate region encompassing the query point (e.g., the region enclosed by the blue dashed rectangle in Figure 6.1) by way of imposing a threshold. Then, the residuals between observed and predicted values within this region can be calculated, and a bound (positive number) can be selected. If the absolute average of the calculated residuals is smaller than the bound, it means the outlier has little negative impact on the prediction of the query point (e.g., Figure 6.1(a)); however, if the absolute average is greater than the bound, the outlier is considered to have a sufficiently negative influence (e.g., Figure 6.1(b)). The reasoning for choosing the absolute average of the residuals within the selected region as the judgment criterion is explained here. Considering the observation form $y_i = y_{itrue} + e_i$, if the LWLS-SVMR perfectly fits the true function, the predicted value will equal the true value ($\hat{y}_i = y_{itrue}$).

Thus, the residuals can be obtained by $y_i - \hat{y}_i = y_i - y_{itrue} = e_i$. As e_i in classical statistical learning approaches is assumed zero mean (Rousseeuw and Leroy 1987), the absolute average of residuals within the selected region (i.e., the range within the blue dashed rectangle) will be zero, that is $\left|E\left(\{e_i\}_{i=1}^{l}\right)\right| = 0$ (assume there are l data points within the blue dashed rectangle). The **algorithm 6.1** is developed to realize this detection procedure:



(a) (b)

Figure 6.1 Schematic sketch for detection of negative effects due to an outlier: (a) outlier far away from the query point has a diminished negative effect on prediction of the query point; (b) outlier close to the query point has a significantly negative effect on prediction of the query point.

Algorithm 6.1: Implementation of proposed algorithm for the detection of negative effect due to outliers

For each query point $\boldsymbol{x}_q, q = 1, \ldots, m$, **do**

(a) Given an optimal combination $(f_q, \gamma_q, \sigma_q^2)$, define a subset $\{(\boldsymbol{x}_{(s)}, y_{(s)})\}_{s=1}^{r}$ and weights $\beta_q(\boldsymbol{x}_{(s)})$ using Eq. (6.6);

(b) Set all weights $v_q(\boldsymbol{x}_{(s)})$ in Eq. (6.5) for the subset $\{(\boldsymbol{x}_{(s)}, y_{(s)})\}_{s=1}^{r}$ to 1;

(c) Solve Eq. (6.5) to obtain $\boldsymbol{\alpha}$, b, and compute residuals $e_s = \alpha_s / \left(\gamma_q v_q(\boldsymbol{x}_{(s)}) \beta_q(\boldsymbol{x}_{(s)})\right)$, where $s = 1, \ldots, r$;

(d) Set a threshold value and select the residuals within the region where the points having weights $\beta_q(\boldsymbol{x}_{(s)})$ are greater than the threshold;

(e) Set a bound value and calculate the absolute of average of the selected residuals, and compare the absolute and bound;

If absolute > bound **then**

Flag = 1

else

Flag = 0

end if

end for

In **Algorithm 6.1**, flag = 1 represents the case when a negative influence is detected; while flag = 0 represents the opposite.

6.2.3 Robust regression by iterative RLWLS-SVMR

When outliers exist close to a query point $\boldsymbol{x}_q$, the predicted response value for the query point $\boldsymbol{x}_q$ will be negatively affected by those outliers (see Figure 6.1(b)). Thus, a robust approach is presented here to eliminate the negative influence of outliers by iteratively updating the weights $v_q(\boldsymbol{x}_{(s)})$, as a function of e_s estimated by LWLS-SVMR. These weights are computed via Eq. (6.9) and according to Suykens et al. (2002).

$$v_q(\boldsymbol{x}_{(s)}) = \begin{cases} 1 & if\ |e_s/\delta| \le c_1 \\ \dfrac{c_2 - |e_s/\delta|}{c_2 - c_1} & if\ c_1 \le |e_s/\delta| \le c_2 \\ \varepsilon & otherwise \end{cases} \tag{6.9}$$

where $c_1 = 2.5$, $c_2 = 3$, $\varepsilon = 10^{-4}$, and $\delta = 1.483 MAD(e_s)$ is a robust estimate where MAD is the median absolute deviation and other variables are defined previously.

Algorithm 6.2: Implementation of proposed iterative RLWLS-SVMR

For each query point $\boldsymbol{x}_q, q = 1, \ldots, m$, **do**

1. Initialization stage:
   (a) Given an optimal combination $\left(f_q, \gamma_q, \sigma_q^2\right)$, define a subset $\left\{\left(\boldsymbol{x}_{(s)}, y_{(s)}\right)\right\}_{s=1}^{r}$ and weights $\beta_q\left(\boldsymbol{x}_{(s)}\right)$ using Eq. (6.6);
   (b) Set all weights $v_q\left(\boldsymbol{x}_{(s)}\right)$ in Eq. (6.5) for the subset $\left\{\left(\boldsymbol{x}_{(s)}, y_{(s)}\right)\right\}_{s=1}^{r}$ to 1;
   (c) Solve Eq. (6.5) to obtain $\boldsymbol{\alpha}$, b, and compute $e_s = \alpha_s / \left(\gamma_q v_q\left(\boldsymbol{x}_{(s)}\right)\beta_q\left(\boldsymbol{x}_{(s)}\right)\right)$, where $s = 1, \ldots, r$.
2. Iterative stage:
   Set the maximum iterative number N, tolerance tol, count $i = 0$, and $t = Inf$
   while $t > tol$ && $i < N$ **do**
   (a) Set $\boldsymbol{\alpha}^{(i)} = \boldsymbol{\alpha}$, $b^{(i)} = b$, $e_s^{(i)} = e_s$, and $v_q^{(i)}\left(\boldsymbol{x}_{(s)}\right) = v_q\left(\boldsymbol{x}_{(s)}\right)$;
   (b) Compute the robust estimate $\delta^{(i)} = 1.483MAD\left(e_s^{(i)}\right)$;
   (c) Update the weights $v_q^{(i+1)}\left(\boldsymbol{x}_{(s)}\right)$ from $\delta^{(i)}$ and $e_s^{(i)}$ using Eq. (6.9);
   (d) Solve Eq. (6.5) to obtain the $\boldsymbol{\alpha}^{(i+1)}$ and $b^{(i+1)}$;
   (e) Update the $e_s^{(i+1)} = \alpha_s^{(i+1)} / \left(\gamma_q v_q^{(i+1)}\left(\boldsymbol{x}_{(s)}\right)\beta_q\left(\boldsymbol{x}_{(s)}\right)\right)$;
   (f) Calculate $t = \left\|\boldsymbol{\alpha}^{(i+1)} - \boldsymbol{\alpha}^{(i)}\right\|$;
   (g) Set $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(i+1)}, b = b^{(i+1)}, e_s = e_s^{(i+1)}$, and $v_q\left(\boldsymbol{x}_{(s)}\right) = v_q^{(i+1)}\left(\boldsymbol{x}_{(s)}\right)$;
   (h) Set $i = i + 1$
   end while
3. Output stage:
   (a) Output the final $\boldsymbol{\alpha}$ and b from the procedure 2
   (b) Given $\boldsymbol{\alpha}$ and b, predict the response value $\hat{y}_q$ of the query point $\boldsymbol{x}_q$ using Eq. (6.7).

end for

After $v_q\left(\boldsymbol{x}_{(s)}\right)$ is determined, the iterative RLWLS-SVMR to predict the response value of a query point $\boldsymbol{x}_q$ is achieved by the **algorithm 6.2** above.

6.2.4 Implementation of a hybrid algorithm

This section introduces the implementation procedure of the proposed RLWLS-SVMR by using a hybrid algorithm. As introduced in **Section 6.2.2**, outliers are only representative of a small amount of the training data, and therefore, not all of the regions will necessarily contain outliers. It is true that some query points may be far away from outliers. In this case, the negative effect from outliers can be ignored, the weights $v_q\left(\boldsymbol{x}_{(s)}\right)$ in Eq. (6.5) do not need to be updated (i.e., the RLWLS-SVMR reverts to LWLS-SVMR), and the results predicted by the LWLS-SVMR model can be trusted, as discussed in **Section 6.2.2**. By combining detection of the negative effect of outliers and the iterative version of RLWLS-SVMR, an efficient hybrid algorithm is developed to predict

query points by adaptively using either LWLS-SVMR or the iterative version of RLWLS-SVMR depending on whether or not a negative effect is detected. The hybrid algorithm is implemented in this section as the following:

Algorithm 6.3: implementation of proposed hybrid algorithm

For each query point $\boldsymbol{x}_q, q = 1, \ldots, m$, **do**

Given an optimal combination $\left(f_q, \gamma_q, \sigma_q^2\right)$, detect if there is any negative influence induced by outliers using *Algorithm 6.1*

If flag = 0 **then**

Predict the response $\hat{y}_q$ of the query point $\boldsymbol{x}_q$ according to $\boldsymbol{\alpha}$ and b obtained in *Algorithm 6.1* using Eq. (6.7) and record the predicted result;

else

Perform an iterative procedure using *Algorithm 6.2* and record the final predicted result;

end if

end for

In addition to the implementation of RLWLS-SVMR, other relevant ML approaches are also implemented for performance comparison. The relevant ML approaches are LS-SVMR (Suykens et al. 2002), weighted LS-SVMR (WLS-SVMR) (Suykens et al. 2002), and iterative WLS-SVMR (IWLS-SVMR) (De Brabanter et al. 2009). Note that the LWLS-SVMR is already incorporated into the hybrid algorithm and the disadvantage of LWLS-SVMR for datasets corrupted by outliers has already been discussed in theory (**Section 6.2.2**). Thus, direct implementation of LWLS-SVMR is not included in this dissertation. LS-SVMR serves as the baseline to address the problems associated with input datasets corrupted by outliers (since all other models used here are variants of LS-SVMR). The main difference between the proposed RLWLS-SVMR and WLS-SVMR and IWLS-SVMR is that RLWLS-SVMR is a robust, *local* model, whereas both WLS-SVMR and IWLS-SVMR are robust, *global* models. The detailed formulations for LS-SVMR, WLS-SVMR, and IWLS-SVMR can be found in the original references (Suykens et al. 2002; De Brabanter et al. 2009). The RBF kernel is also utilized for LS-SVMR, WLS-SVMR, and IWLS-

SVMR. The optimal hyper-parameter combinations for all four models are obtained using five-fold cross-validation on the training data (De Brabanter et al. 2002).

6.2.5 Numerical results

This section presents illustrative examples for validating the proposed approach. In order to assess the proposed approach for a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ corrupted by outliers, two examples are carried out. The two examples vary in terms of the type of data: one employs simulated datasets, and the second utilizes a multi-dimensional, real-world dataset. The proposed method is compared with LS-SVMR, WLS-SVMR, and IWLS-SVMR for both examples. The generalization performance for the simulated datasets is quantified by the coefficient of determination (R^2), mean absolute error (MAE), and root mean square error (RMSE) metrics introduced in **Section 3.4.4**. Since R^2, MAE, and RMSE are sensitive to outliers, and the real world datasets may contain outliers, the performance for real world datasets is quantified by a robust variant of R^2 (R_R^2) (Kvalseth 1985) which was also presented in **Section 3.4.4**.

A very simple example is used to illustrate that original R^2, MAE and RMSE are sensitive to outliers in a test set, but the robust variant of R^2 is robust to such outliers. Assume a response variable in the test set is corrupted by one outlier, $\boldsymbol{y} = (2, 4, 6, 8, 100, 12, 14)$ (i.e., the fifth element (100) is corrupt, and the actual value is 10). A robust model is applied to predict the response values for the test set, and the predicted response is $\hat{\boldsymbol{y}} = (2, 4, 6, 8, 10, 12, 14)$, which means the robust model perfectly predicts the response in the test set. However, if we use the original R^2, MAE, RMSE and R_R^2 to quantify the performance of the robust model, one can obtain the performance of this robust model is -0.09, 12.86, 34.01, and 1, respectively. Therefore, only the robust variant of R^2 reflects the actual performance of the robust model, and the other statistics are sensitive to outliers and fail to quantify the actual performance. Note that if more outliers exist

in the test set, the robust variant of R^2 may also fail to reflect the actual performance, but it is still more robust than RMSE, MAE, and the original R^2 (Kvalseth 1985).

6.2.5.1 Results for simulated datasets

In this example, four synthetic datasets corrupted by four combinations of two types of random error terms and two types of outliers are generated to show the robustness of RLWLS-SVMR. In the real world, the random error term reflects data noise that cannot be avoided as purely clean data is impossible (Rousseeuw and Leroy 1987) (note that noise is not necessarily representative of an outlier as introduced in Rousseeuw and Leroy 1987). The error model proposed by Huber (1964) is used to generate these four synthetic datasets. Specifically, the random error terms are simulated using a Gaussian distribution with zero mean and either constant or non-constant variance. The outliers are simulated by either a Gaussian distribution with higher variance or a standard Cauchy distribution with heavy tails. In this setting, a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ not corrupted by outliers is simulated from a *sinc* function, which is defined in this way:

$$y_i = \frac{sin(x_i)}{x_i} + e_i \tag{6.10}$$

where x_i is drawn from a uniform distribution $x_i \sim U[-10, 10]$, e_i is a random error term that is drawn from a Gaussian distribution using either constant variance, i.e., $e_i \sim N(0, 0.01^2)$ or non-constant variance, i.e., $e_i \sim N(0, \sigma_i^2)$ and $\sigma_i \sim U[0.01, 0.05]$.

The smaller variance is selected to distinguish noise from outliers in the regression setting (Figure 6.2). The number of normal data points following the definition above is 162. Another 38 points are defined as the potential outliers, where e_i is drawn from either a Gaussian distribution with higher variance, i.e., $e_i \sim N(0, 1^2)$ or a standard Cauchy distribution with heavy tails, i.e., $e_i \sim C(0,1)$. A total of 200 data points are drawn from the mixture procedure introduced above to form the training set. By setting different random number seeds, four combinations of error terms

and outliers are established to form four synthetic training datasets where the locations of outliers differs in order to more extensively evaluate the robustness of these four ML models, as shown in Figure 6.2 (a,c,e,g).

In Figure 6.2, the four synthetic training datasets are shown on the left (subfigures a,c,e,g) which differ according to the error and outlier distributions, while the corresponding test sets are shown on the right (subfigures b,d,f,h). The variations are as follows: Figure 6.2(a,b), *Synthetic 1:* the error terms for normal points are drawn from $e_i \sim N(0, 0.01^2)$ and the potential outliers are drawn from $e_i \sim N(0, 1^2)$; Figure 6.2 (c,d), *Synthetic 2:* the error terms for normal points are drawn from $e_i \sim N(0, \sigma_i^2)$ and $\sigma_i \sim U[0.01, 0.05]$, and the potential outliers are drawn from $e_i \sim N(0, 1^2)$; Figure 6.2(e,f), *Synthetic 3*: the error terms for normal points are drawn from $e_i \sim N(0, 0.01^2)$ and the potential outliers are drawn from $e_i \sim C(0,1)$; and, Figure 6.2(g,h). *Synthetic 4:* the error terms for normal points are drawn from $e_i \sim N(0, \sigma_i^2)$ and $\sigma_i \sim U[0.01, 0.05]$ and the potential outliers are drawn from $e_i \sim C(0,1)$. Note that the potential outliers are only applied to the four synthetic training datasets. It is clearly observed that not all of the potential outliers are real outliers, and only the points far from the bulk of the data points are true outliers (i.e., y-outliers). Another 200 independent test data points (i.e., Figure 6.2 (b,d,f,h)) not corrupted by outliers corresponding to four different synthetic training datasets are drawn to test the performance of the data-driven procedures. The scatter plots of training and test data as well as the predictions on the test data by the LS-SVMR, WLS-SVMR, IWLS-SVMR, and RLWLS-SVMR models are presented in Figure 6.2.

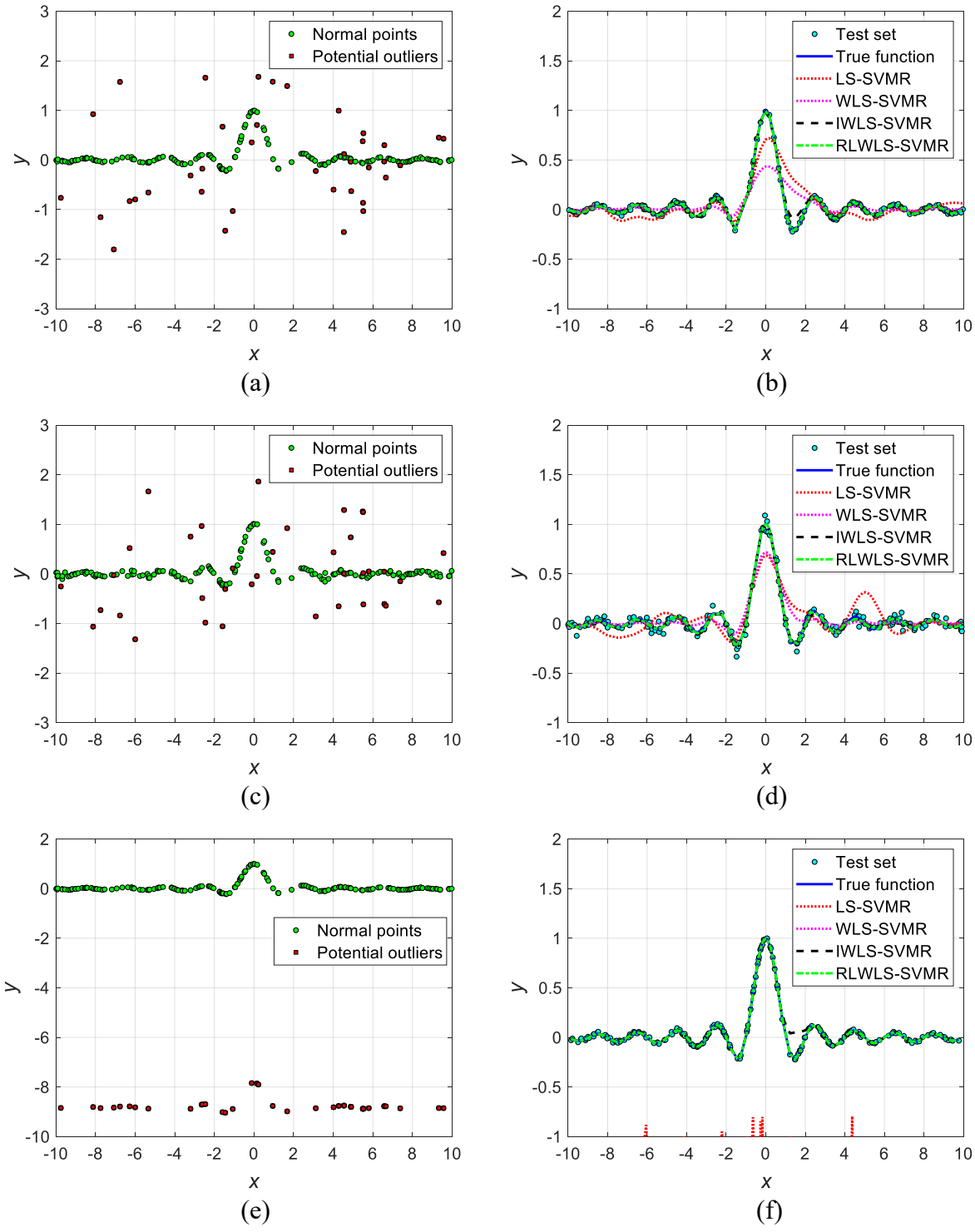
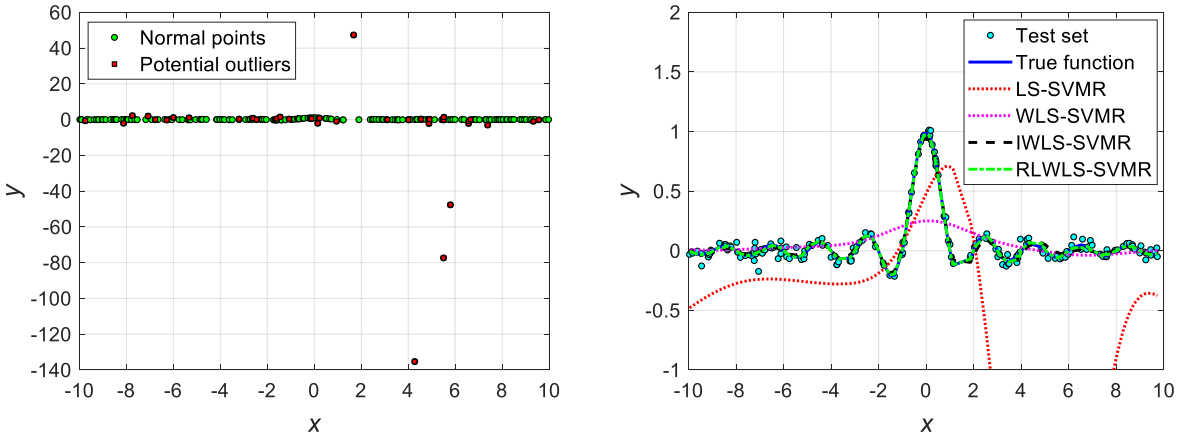


Figure 6.2 Left subfigures (a,c,e,g): Training of a *sinc* function with four synthetic training datasets (with various error and simulated outlier characteristics employed to plague the training data); Right subfigures (b,d,f,h): Testing (estimation of the *sinc* function) by LS-SVMR, WLS-SVMR, IWLS-SVMR, and RLWLS-SVMR.



(g) (h)

Figure 6.2 Continued.

It should be noted that for LS-SVMR, WLS-SVMR and IWLS-SVMR, a global model is formed using the entire training dataset before predicting the query points in the test dataset. For the proposed RLWLS-SVMR, different query points in the test dataset are predicted by distinct, individual local models. Each model is formed by training different subsets of training data to achieve an adequate trade-off between prediction capacity and the number of input data for different query points. A comparison of the results between LS-SVMR, WLS-SVMR, IWLS-SVMR, and the proposed RLWLS-SVMR on the four test datasets is shown in Figure 6.2(b,d,f,h). By observation, compared to the true function, LS-SVMR is negatively affected by outliers, especially by those produced by the standard Cauchy distribution with heavy tails. The LS-SVMR is influenced heavily in the direction of outliers, leading to a significant deviation from the true function (Figures 6.2(f) and 6.2(h)). The WLS-SVMR model improves the performance of LS-SVMR but still suffers negative effects. By contrast, both IWLS-SVMR and the proposed RLWLS-SVMR models perform much more robustly to outliers, where both overcome the negative interference from outliers and very closely fit the true function.

Table 6.1 presents the metrics of original R^2, RMSE, and MAE for LS-SVMR, WLS-SVMR, IWLS-SVMR, and R-LWLS-SVMR based on the test datasets. Since these datasets are simulated and we know the true values, these metrics can give correct quantifications for the actual performance of these four ML models. Thus, it can be concluded that both IWLS-SVMR and RLWLS-SVMR do adequately capture the true function, and the proposed RLWLS-SVMR has the highest R^2 and lowest RMSE and MAE values, which deem it as the best model for these types of datasets among the four ML models.

Table 6.1 Performance comparison between LS-SVMR, WLS-SVMR, IWLS-SVMR, and RLWLS-SVMR in terms of original R^2, RMSE, and MAE. The synthetic datasets represent the training data corrupted by outliers and the original R^2, RMSE, and MAE are computed on corresponding test datasets between predicted and true values. The bold values represent the best performance.

Datasets	Models	RMSE	MAE	R^2
Synthetic dataset 1	LS-SVMR	0.1163	0.0742	0.5182
	WLS-SVMR	0.1115	0.0589	0.5576
	IWLS-SVMR	0.0213	0.0070	0.9839
	RLWLS-SVMR	**0.0052**	**0.0040**	**0.9990**
Synthetic dataset 2	LS-SVMR	0.1380	0.0992	0.5834
	WLS-SVMR	0.0847	0.0515	0.8428
	IWLS-SVMR	0.0137	0.0106	0.9959
	RLWLS-SVMR	**0.0083**	**0.0051**	**0.9985**
Synthetic dataset 3	LS-SVMR	1.7270	1.6301	-43.5127
	WLS-SVMR	1.7160	1.6964	-42.9457
	IWLS-SVMR	0.0490	0.0164	0.9642
	RLWLS-SVMR	**0.0019**	**0.0011**	**0.9999**
Synthetic dataset 4	LS-SVMR	1.6499	1.0328	-42.4528
	WLS-SVMR	0.1997	0.1062	0.3633
	IWLS-SVMR	0.0229	0.0179	0.9916
	RLWLS-SVMR	**0.0085**	**0.0050**	**0.9989**

6.2.5.2 Example 2: results for real-world datasets

To further investigate the robustness of the proposed RLWLS-SVMR for multi-dimensional problems and demonstrate its practical application in the real-world, we employ eight real-world multi-dimensional datasets across different engineering and science domains to test the model performance and compare it with LS-SVMR, WLS-SVMR and IWLS-SVMR. These eight benchmark datasets (and associated tasks) are the following: (1) *Circular RC columns* (predicting the lateral strength) as presented in Chapter III (see Appendix B); (2) *Concrete slump specimens* (predicting concrete flow) (Yeh 2007); (3) *Automobile characteristics* (predicting the fuel consumption) (Quinlan 1993); (4) *Servo* (predicting the rising time of a servomechanism) (Quinlan 1993); (5) *Crabs* (predicting the body depth of crabs) (Campbell and Mahon 1974); (6) *Boston housing* (predicting the median value of home price in the greater Boston area) (Harrison and Rubinfeld 1978); (7) *Nelson* (predicting the dielectric breakdown strength) (Nelson 1981); and (8) *Bodyfat* (predicting the body fat of human beings) (Penrose et al 1985). The detailed information for all eight real-world datasets can be found in the provided websites in the references. The final results are reported for all eight datasets to demonstrate the broad application of the proposed appraoch. A detailed discussion of how the models peform is carried out for the *Circular RC column* dataset presented in **Chapter III** (see Appendix B) to thoroughly explain the proposed approach and its' performance.

Accurate modeling of lateral strength of RC columns is a very important topic in structural and earthquake engineering, as the strength is an important factor for the design of buildings. In this specific example, the prediction performance of LS-SVMR, WLS-SVMR, IWLS-SVMR, and the proposed RLWLS-SVMR is evaluated for lateral strength prediction of the circular columns. Detailed information regarding this dataset can be found in **Chapter III** (see Appendix B).

The leave-one-out (LOO) cross-validation procedure presented in **Section 3.4.3** is employed to evaluate the performance of LS-SVMR, WLS-SVMR, IWLS-SVMR, and RLWLS-SVMR on lateral strength prediction of these 160 RC columns as well as for the other seven real-world datasets. The performance of these ML models on prediction in these eight real-world datasets is quantified by the robust variant of R^2 defined in **Section 3.4.4**. Note that the true values of the response variables in the real-world datasets are unknown. This is because the observed values of the response variables in real-world datasets contain a random error term (i.e., $y = y_{true} + e$), and the random error is unknown. If outliers exist in the real-world dataset, the original R^2, RMSE, and MAE will be sensitive to these outliers and fail to reflect the prediction performance of these four ML models based on the LOO cross-validation procedure, while the robust variant of R^2 is more robust to outliers and can give a more objective evaluation, as discussed in **Section 6.2.5**. Additionally, it is worth noting that a robust estimator is able to detect outliers where points possess large residuals, while a non-robust estimator cannot be used for this purpose, because the outliers may possess very small residuals (Rousseeuw and Leroy 1987).

A comparison of results is presented in Figure 6.3. By observation of this figure, the green points in all four ML models flock around the red lines which indicates that the predicted and observed values are equal (i.e., near-perfect prediction). However, compared to IWLS-SVMR (Figure 6.3(c)) and RLWLS-SVMR (Figure 6.3(d)), the green points in LS-SVMR (Figure 6.3(a)) and WLS-SVMR (Figure 6.3(b)) are much more scattered. Additionally, there are three red square points in all four ML models which are distant from the red lines. Compared to LS-SVMR and WLS-SVMR, the two red points (i.e., values more than 1000 kN in the observed value direction in Figure 6.3) in IWLS-SVMR and RLWLS-SVMR are much further from the red lines, which lead to higher residuals (i.e., difference between observed and predicted values). The other

remaining red point (i.e., value less than 1000 kN in the observed value direction in Figure 6.3) appears to maintain nearly the same deviation in all four ML models (i.e., the residuals for this red point in all four ML models are almost equivalent).

By analysis of the dataset, it is found that these two red points (i.e., values more than 1000 kN in the observed value direction in Figure 6.3) correspond to two full-scale column tests conducted by Stone and Cheok (1989), where the section dimensions (explanatory variables) and lateral strength (response variable) of these two columns are extreme values which are far larger than all other remaining columns in the dataset. It is also found that the other remaining red point corresponds to a column test performed by Priestley et al. (1981) where the applied axial load (explanatory variable) on this column is an extreme value which is much larger than all other columns in the dataset. Thus, these three red points are detected and identified as high leverage points (i.e., extreme values in the x direction; note that this does not take y into accoun,t and if a high leverage point is also an outlier, it will negatively affect the performance of a non-robust estimator). By observation of Figure 6.3, it is evident that the LS-SVMR is heavily influenced by these two high leverage points (i.e., values more than 1000 kN (outliers)). This negative effect for the LS-SVMR model is exhibited by smaller residuals for the two high leverage points (outliers) but greater scatter in the remaining points than the results for WLS-SVMR, IWLS-SVMR, and RLWLS-SVMR. The WLS-SVMR slightly reduces the negative interference from these points where the residuals are slightly larger, and the green points are slightly less scattered in comparison to the LS-SVMR. However, both IWLS-SVMR and RLWLS-SVMR improve the prediction on all green points by significantly reducing the negative interference, where the green points are much less scattered and those two red points are far away from the red lines. The proposed RLWLS-SVMR performs better than IWLS-SVMR where the green points in RLWLS-SVMR are

less scattered than those in IWLS-SVMR. Since the other remaining red point does not deleteriously change the prediction for all four ML models, it can be concluded that this leverage point is a good leverage point while the other two red points mentioned above are bad leverage points that are also outliers. The final results for the RC column dataset as well as for the other seven datasets mentioned previously are reported in Table 6.2. From Table 6.2, it is observed that the proposed RLWLS-SVMR performs best across all eight benchmark real-world datasets.

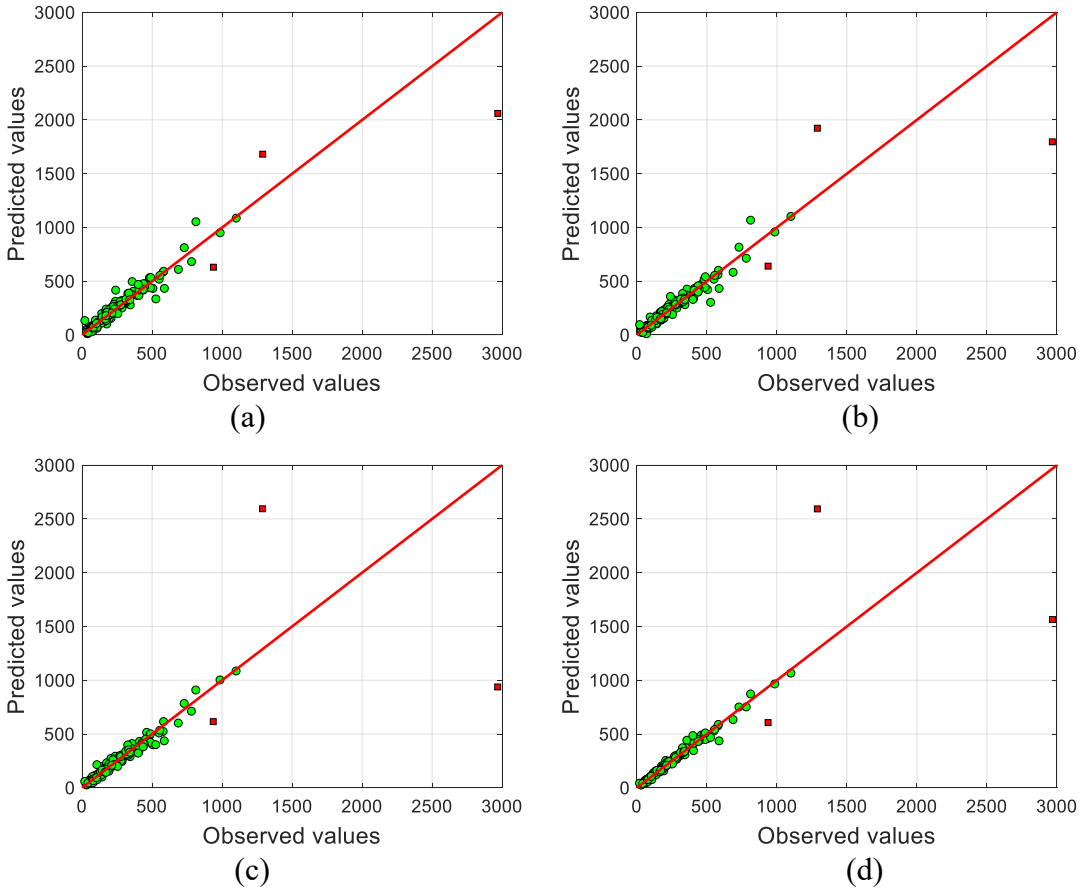


Figure 6.3 Comparison of results using leave-one-out (LOO) cross validation procedure on 160 RC columns of: (a) LS-SVMR, (b) WLS-SVMR, (c) IWLS-SVMR, and (d) RLWLS-SVMR.

Table 6.2 Performance comparison between LS-SVMR, WLS-SVMR, IWLS-SVMR and RLWLS-SVMR on eight benchmark real world datasets in terms of the robust variant of R^2 using LOO cross validation procedure. The bold values represent the best performance.

Datasets	Number of observations	Number of predictors	LS-SVMR	WLS-SVMR	IWLS-SVMR	RLWLS-SVMR
Columns	160	10	0.9747	0.9756	0.9837	**0.9928**
Concrete slump	103	7	0.4675	0.4456	0.4338	**0.6419**
Auto MPG	392	7	0.9393	0.9427	0.9434	**0.9723**
Boston Housing	506	13	0.8691	0.8820	0.8854	**0.9231**
Bodyfat	252	14	0.9973	0.9994	0.9995	**0.9999**
Crabs	200	7	0.9928	0.9924	0.9921	**0.9937**
Servo	167	4	0.7367	0.8326	0.8789	**0.9265**
Nelson	128	2	0.8626	0.8657	0.8675	**0.9012**

6.3 Solution to Missing Data

As introduced in **Section 2.4.2**, any standard machine learning (ML) methods fail to construct a data-driven model when a dataset is incomplete and contains missing data. This section presents a new multiple imputation (MI) method to address the missing data problem in ML models. The approach works by filling in each missing value with multiple realistic, valid candidates, accounting for the uncertainty due to missing data. The proposed method, called sequential regression-based predictive mean matching (SRB-PMM), utilizes Bayesian parameter estimation to consecutively infer the model parameters for variables with missing values, conditionally based on the fully observed and imputed variables. Given the model parameters, a hybrid approach integrating PMM with a cross-validation algorithm is developed to obtain the most plausible imputed dataset. Two case studies are carried out to validate the usefulness of the SRB-PMM approach based on the rectangular RC column dataset presented in **Chapter III**. The results from both case studies suggest that the proposed SRB-PMM approach is an effective means to handle missing data problems prominent in the earthquake engineering field.

6.3.1 Development of SRB-PMM

This section presents the formulation of the proposed *SRB-PMM method*. The proposed method couples the sequential regression, predictive mean matching (PMM), and cross-validation (CV) algorithms to generate multiple plausible and realistic candidates for each missing value with consideration of the uncertainty due to missing data. The detailed procedure for the proposed method is presented below.

Assume a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i \in R^{p+q}$ and $y_i \in R$ is collected from a domain of interest. In this dataset, there are n observations, and each observation has $(p+q)$ explanatory variables (i.e., $\boldsymbol{x}_i \in R^{p+q}$) and one response variable (i.e., $y_i \in R$). However, some data points

(i.e., observations) have one or more explanatory variables with missing values, making the collected dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ incomplete. For the remainder of this section, we assume there are no missing values in the response variable (as this is not relevant in the proposed application domain) and the following notations are used. Let $\boldsymbol{X}^{obs} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p) \in R^{n \times p}$ be a matrix with n observations, and each observation has p fully observed explanatory variables (i.e., there are no missing values for all n observations in these p explanatory variables, such as $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p$ shown in Table 2.1). Let $\boldsymbol{X}^{miss} = (\boldsymbol{Z}_{(1)}, \ldots, \boldsymbol{Z}_{(q)}) \in R^{n \times q}$ be a matrix with n observations and each observation has q partially observed explanatory variables (i.e., there is at least one missing value for each of these q partially observed explanatory variables, such as $\boldsymbol{Z}_{(1)}, \boldsymbol{Z}_{(2)}, \boldsymbol{Z}_{(3)}$ shown in Table 2.1), and $\boldsymbol{Z}_{(1)}, \ldots, \boldsymbol{Z}_{(q)}$ have been ordered increasingly in terms of the missing data ratios. Let $\boldsymbol{y} \in R^n$ be a vector. Thus, the dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ can also be written as $\boldsymbol{D} = (\boldsymbol{X}, \boldsymbol{y})$, where $\boldsymbol{X} = (\boldsymbol{X}^{obs}, \boldsymbol{X}^{miss}) \in R^{n \times (p+q)}$. A schematic format of this incomplete dataset is presented in Table 2.1. Let $\boldsymbol{O} = (\boldsymbol{O}_1, \ldots, \boldsymbol{O}_q) \in R^{n \times q}$ be the indicator matrix where $o_{ij} = 1$ if x_{ij} is observed and $o_{ij} = 0$ if x_{ij} is missing. Note that the indicator matrix $\boldsymbol{O}$ is only applied to $\boldsymbol{X}^{miss}$. Thus, for the jth explanatory variable, where $j = 1, \ldots, q$, the vector $\boldsymbol{Z}_{(j)}$ can be thought of as consisting of two parts: $\boldsymbol{Z}_{(j)}^{obs} = \{x_{ij} : o_{ij} = 1\}$, the data that is observed, and $\boldsymbol{Z}_{(j)}^{miss} = \{x_{ij} : o_{ij} = 0\}$, the data that is not observed. We assume that the missing data are missing at random (MAR) (Hoff 2009), which means that $\boldsymbol{O}$ and $\boldsymbol{X}^{miss}$ are statistically independent and the distribution of $\boldsymbol{O}$ does not depend on the model parameter $\boldsymbol{\Theta}$.

From a probability perspective, missing values can be reasonably imputed only when a multivariate imputation model $p(\boldsymbol{X}^{miss} | \boldsymbol{X}^{obs}, \boldsymbol{\Theta})$ is specified correctly (Schafer 1997), where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_q)$ is the model parameter (e.g., regression coefficients, dispersion parameter). The

multivariate imputation model $p\left(\boldsymbol{X}^{miss}|\boldsymbol{X}^{obs},\boldsymbol{\Theta}\right)$ can be factored as follows (Raghunathan et al. 2001):

$$\begin{aligned} p\left(\boldsymbol{X}^{miss}|\boldsymbol{X}^{obs},\boldsymbol{\Theta}\right) &= p\left(\boldsymbol{Z}_{(1)},\dots,\boldsymbol{Z}_{(q)}|\boldsymbol{X}^{obs},\boldsymbol{\Theta}\right) \\ &= p_q\left(\boldsymbol{Z}_{(q)}|\boldsymbol{Z}_{(1)},\dots,\boldsymbol{Z}_{(q-1)},\boldsymbol{X}^{obs},\boldsymbol{\theta}_q\right) \\ &\times p_{q-1}\left(\boldsymbol{Z}_{(q-1)}|\boldsymbol{Z}_{(1)},\dots,\boldsymbol{Z}_{(q-2)},\boldsymbol{X}^{obs},\boldsymbol{\theta}_{q-1}\right)\times\dots\times p_1\left(\boldsymbol{Z}_{(1)}|\boldsymbol{X}^{obs},\boldsymbol{\theta}_1\right) \end{aligned} \quad (6.11)$$

where $p_j, j=1,\dots,q$ are the conditional density functions.

Eq. (6.11) is initiated by regression of the variable with the fewest number of missing values (i.e., $\boldsymbol{Z}_{(1)}$), $\boldsymbol{Z}_{(1)}$ on $\boldsymbol{X}^{obs}$**.** The missing values are imputed by *PMM* based on the regression results to form an *imputed,* complete data vector $\boldsymbol{Z}_{(1)}$. Then, the complete $\boldsymbol{Z}_{(1)}$ vector is appended with $\boldsymbol{X}^{obs}$ to impute variable $\boldsymbol{Z}_{(2)}$ with the next fewest number of missing values using the univariate model $p_2\left(\boldsymbol{Z}_{(2)}|\boldsymbol{Z}_{(1)},\boldsymbol{X}^{obs},\boldsymbol{\theta}_2\right)$. This means, $\boldsymbol{Z}_{(1)}$ is imputed on $\boldsymbol{U}_1=\boldsymbol{X}^{obs}$, $\boldsymbol{Z}_{(2)}$ is imputed on $\boldsymbol{U}_2=\left(\boldsymbol{X}^{obs},\boldsymbol{Z}_{(1)}\right)$ where $\boldsymbol{Z}_{(1)}$ has imputed values, $\boldsymbol{Z}_{(3)}$ is imputed on $\boldsymbol{U}_3=\left(\boldsymbol{X}^{obs},\boldsymbol{Z}_{(1)},\boldsymbol{Z}_{(2)}\right)$ where $\boldsymbol{Z}_{(1)}$ and $\boldsymbol{Z}_{(2)}$ have imputed values, and others (i.e., $\boldsymbol{Z}_{(4)},\dots,\boldsymbol{Z}_{(q)}$) are imputed in a similarly sequential manner. The detailed imputation procedure for imputing each partially observed explanatory variable using *SRB-PMM* is presented below.

Since missing values exist in $\boldsymbol{Z}_{(j)}$, the model for $\boldsymbol{Z}_{(j)}$ cannot be established directly. For the model $p_1\left(\boldsymbol{Z}_{(1)}|\boldsymbol{X}^{obs},\boldsymbol{\theta}_1\right)$ (which can be written as $p_1\left(\boldsymbol{Z}_{(1)}^{obs},\boldsymbol{Z}_{(1)}^{miss}|\boldsymbol{X}^{obs},\boldsymbol{\theta}_1\right)$), the unknown quantities include the model parameter $\boldsymbol{\theta}_1$ and missing values $\boldsymbol{Z}_{(1)}^{miss}$. According to Bayes rule, the following equation can be given:

$$\begin{aligned} &p_1\left(\boldsymbol{Z}_{(1)}^{obs},\boldsymbol{Z}_{(1)}^{miss}|\boldsymbol{X}^{obs},\boldsymbol{\theta}_1\right) \\ &\quad= p_1\left(\boldsymbol{Z}_{(1)}^{miss}|\boldsymbol{Z}_{(1)}^{obs},\boldsymbol{X}^{obs},\boldsymbol{\theta}_1\right)\times p_1\left(\boldsymbol{Z}_{(1)}^{obs}|\boldsymbol{X}^{obs},\boldsymbol{\theta}_1\right) \end{aligned} \quad (6.12)$$

We specify a normal linear model for $p_1\left(\boldsymbol{Z}_{(1)}^{obs}|\boldsymbol{X}^{obs},\boldsymbol{\theta}_1\right)$ as well as for all other conditional density functions. For a linear model, the regression of $\boldsymbol{Z}_{(1)}^{obs}$ from $\boldsymbol{X}^{obs}$ depends only on $\boldsymbol{X}^{obs1} = \{\boldsymbol{X}^{obs}: o_{i1} = 1\}$, which is given by:

$$\boldsymbol{Z}_{(1)}^{obs} = \boldsymbol{X}^{*obs1}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \tag{6.13}$$

where $\boldsymbol{\beta}_1 = \left(\beta_{11}, \ldots \beta_{1p}\right)$ is a regression coefficient vector; $\boldsymbol{X}^{*obs1}$ is the design matrix including the column corresponding to the intercept term in the regression model (i.e., the column with unity entries), $\boldsymbol{\varepsilon}_1 = \left(\varepsilon_{11}, \ldots, \varepsilon_{1(n_{ob1})}\right)$ is an error vector, $n_{ob1} = length\left(\boldsymbol{Z}_{(1)}^{obs}\right)$ is the number of observed data in $\boldsymbol{Z}_{(1)}$ (note that the number of observations in $\boldsymbol{X}^{*obs1}$ is also n_{ob1}, i.e., $size\left(\boldsymbol{X}^{*obs1}, 1\right) = size\left(\{\boldsymbol{X}^{obs}: o_{i1} = 1\}, 1\right) = n_{ob1}$) and $\varepsilon_{11}, \ldots, \varepsilon_{1(n_{ob1})} \sim$ i.i.d. $N(0, \sigma_1^2)$ or $\boldsymbol{\varepsilon}_1 \sim N(\boldsymbol{0}, \sigma_1^2\boldsymbol{I})$, and $\boldsymbol{I}$ is the identity matrix.

Thus, in this case, the model parameter $\boldsymbol{\theta}_1 = (\boldsymbol{\beta}_1, \sigma_1^2)$ and the posterior distributions need to be determined. Given this setting, the likelihood function is a multivariate normal function $(\boldsymbol{X}^{*obs1}\boldsymbol{\beta}_1, \sigma_1^2\boldsymbol{I})$ (Hoff 2009), which includes unknown model parameters $\boldsymbol{\beta}_1$ and σ_1^2. The posterior joint distribution of these two unknown model parameters can be written as follows:

$$\begin{aligned} &p\left(\boldsymbol{\beta}_1, \sigma_1^2|\boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right) \\ &\quad = p\left(\boldsymbol{\beta}_1|\sigma_1^2, \boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right) \times p\left(\sigma_1^2|\boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right) \end{aligned} \tag{6.14}$$

The posterior joint distribution of unknown model parameters $(\boldsymbol{\beta}_1, \sigma_1^2)$ can be made via a Monte Carlo approximation by sampling from these two conditional distributions $p\left(\sigma_1^2|\boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right)$ and $p\left(\boldsymbol{\beta}_1|\sigma_1^2, \boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right)$. Throughout this section, a g-prior distribution (Zellner 1986) is used for these unknown model parameters $(\boldsymbol{\beta}_1, \sigma_1^2)$. With the use of the g-prior distribution**,** the resulting conditional distributions for $p\left(\boldsymbol{\beta}_1|\sigma_1^2, \boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right)$ and $p\left(\sigma_1^2|\boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right)$ are obtained as follows (Hoff 2009):

$$\left\{\sigma_1^2|\boldsymbol{X}^{obs1},\boldsymbol{Z}_{(1)}^{obs}\right\}\sim\text{inverse-gamma}\left(\frac{1+n_{ob1}}{2},\frac{\hat{\sigma}_1^2+SSR}{2}\right) \tag{6.15}$$

$$\left\{\boldsymbol{\beta}_1|\sigma_1^2,\boldsymbol{X}^{obs1},\boldsymbol{Z}_{(1)}^{obs}\right\}\sim N\left(\frac{g}{g+1}\widehat{\boldsymbol{\beta}}_1,\frac{g}{g+1}\sigma_1^2((\boldsymbol{X}^{*obs1})^{\boldsymbol{T}}\boldsymbol{X}^{*obs1})^{-1}\right) \tag{6.16}$$

where $\widehat{\boldsymbol{\beta}}_1=\left(\left(\boldsymbol{X}^{*obs1}\right)^{\boldsymbol{T}}\boldsymbol{X}^{*obs1}\right)^{-1}\left(\boldsymbol{X}^{*obs1}\right)^{\boldsymbol{T}}\boldsymbol{Z}_{(1)}^{obs}$ is a regression coefficient vector estimated by ordinary least squares (OLS); $\hat{\sigma}_1^2=sum\left(\left(\boldsymbol{Z}_{(1)}^{obs}-\boldsymbol{X}^{*obs1}\widehat{\boldsymbol{\beta}}_1\right)^{\boldsymbol{2}}\right)/(n_{ob1}-p)$ is an unbiased estimate of σ_1^2, $SSR=\left(\boldsymbol{Z}_{(1)}^{obs}\right)^T\left(\boldsymbol{I}-g\boldsymbol{X}^{*obs1}\left(\left(\boldsymbol{X}^{*obs1}\right)^{\boldsymbol{T}}\boldsymbol{X}^{*obs1}\right)^{-1}\left(\boldsymbol{X}^{*obs1}\right)^{\boldsymbol{T}}/(g+1)\right)\boldsymbol{Z}_{(1)}^{obs}$ is the sum of squared residuals (SSR).

Since we can sample from both of these two conditional distributions, a value of $(\boldsymbol{\beta}_1,\sigma_1^2)$ sampled from the posterior joint distribution $p\left(\boldsymbol{\beta}_1,\sigma_1^2|\boldsymbol{X}^{obs1},\boldsymbol{Z}_{(1)}^{obs}\right)$ can be extracted by first sampling σ_1^2 from Eq. (6.15) and then sampling $\boldsymbol{\beta}_1$ from Eq. (6.16) given the drawn σ_1^2. Thus, multiple independent sample values from $p\left(\boldsymbol{\beta}_1,\sigma_1^2|\boldsymbol{X}^{obs1},\boldsymbol{Z}_{(1)}^{obs}\right)$ can be made by independently repeating the procedure. Suppose we obtain S sample values $\{(\boldsymbol{\beta}_1,\sigma_1^2)_s\}_{s=1}^S$ from $p\left(\boldsymbol{\beta}_1,\sigma_1^2|\boldsymbol{X}^{obs1},\boldsymbol{Z}_{(1)}^{obs}\right)$. So, the mean of the model parameters given the S samples can be obtained by a Monte Carlo approximation, where $\overline{\boldsymbol{\beta}}_1=(1/S)\sum_{s=1}^S(\boldsymbol{\beta}_1)_s$ and $\bar{\sigma}_1^2=(1/S)\sum_{s=1}^S(\sigma_1^2)_s$. Given the sampled model parameters $\left(\overline{\boldsymbol{\beta}}_1,\bar{\sigma}_1^2\right)$, a regression model can be established by inserting the model parameters into Eq. (6.13). Next, a hybrid procedure to generate the realistic candidates for missing values using *PMM* incorporated with a k-fold cross-validation procedure based on an ML model is described.

In contrast to other imputation approaches, the goal of the regression model in the *PMM* approach is not to actually generate the imputed values. Instead, the aim is to establish a metric for matching cases with missing values to similar cases with observed values (Schenker and Taylor 1996; Little 1988; Morris et al. 2014; Rubin 1986). The similarity is measured by the Euclidean

distance between the fitted values for the observed data and the predicted values for the missing data. For each missing case, the *PMM* algorithm first identifies a set of cases of observed data whose fitted values are close to the predicted value for the case with missing data in terms of the measured similarities. From those close cases, one case is randomly sampled and assigned its observed value as a substitute for the missing value. Therefore, the *PMM* imputes the missing values based on the realistic observed values, and thus, never generates imputations outside the observed value ranges. In this way, *PMM* overcomes the problems associated with meaningless imputations generated by aforementioned MI approaches (**Section 2.4.2**). However, in this procedure, the randomly selected case may not be the most plausible case, since there is no standard method to evaluate whether or not the selected one is the most plausible.

To solve this problem, a hybrid approach is developed to select the most plausible cases based on the k-fold cross-validation (CV) algorithm (James et al. 2013). The purpose of this hybrid method is not to evaluate if a randomly selected single case for one missing value in one partially observed explanatory variable is the most plausible. Instead, it evaluates the *imputed*, complete dataset where the missing values in all the partially observed explanatory variables are imputed. The evaluation criterion is based on an ML model's performance estimated by the CV algorithm on the *imputed*, complete dataset, where the most plausible cases should be those that result in the ML model with the best performance. We denote that $\boldsymbol{X}^{obs0} = \{\boldsymbol{X}^{obs}: o_{i1} = 0\}$, $\boldsymbol{X}^{*obs0}$ is the design matrix for $\boldsymbol{X}^{obs0}$ as explained for $\boldsymbol{X}^{*obs1}$ previously, n_{ob0} is the number of cases with missing values in $\boldsymbol{Z}_{(1)}$ (note that the number of missing data in $\boldsymbol{X}^{obs0}$ is also n_{ob0}, i.e., $size(\boldsymbol{X}^{obs0}, 1) = size(\{\boldsymbol{X}^{obs}: o_{i1} = 0\}, 1) = n_{ob0}$), and $n_{ob0} + n_{ob1} = n$. The detailed procedure regarding the donor pool generation (i.e., selected close cases) for the missing values in $\boldsymbol{Z}_{(1)}^{miss}$ using the *PMM* algorithm is summarized in **Algorithm 6.4**.

Algorithm 6.4: Generate realistic candidates for missing values using *PMM*

1) Calculate the fitted and predicted values for $\boldsymbol{Z}_{(1)}^{obs}$ and $\boldsymbol{Z}_{(1)}^{miss}$, respectively:
   $\widehat{\boldsymbol{Z}}_{(1)}^{obs} = \boldsymbol{X}^{*obs1}\overline{\boldsymbol{\beta}}_1$
   $\widehat{\boldsymbol{Z}}_{(1)}^{miss} = \boldsymbol{X}^{*obs0}\overline{\boldsymbol{\beta}}_1$
2) Select r cases as the plausible candidates for each missing value $Z_{(1),i}^{miss}$ in $\boldsymbol{Z}_{(1)}^{miss}$:
   for all $i = 1, \dots, n_{ob0}$ **do**
   2.1) Calculate the Euclidian distance vector $\boldsymbol{d}_i = \left\| \widehat{\boldsymbol{Z}}_{(1)}^{obs} - \hat{Z}_{(1),i}^{miss} \right\|$.
   2.2) Sort $\boldsymbol{d}_i$ increasingly to obtain an increasingly ordered vector $\boldsymbol{d}_i = \left(d_{i(1)}, \dots, d_{i(n_{ob1})}\right)$.
   2.3) Select r cases from $\boldsymbol{Z}_{(1)}^{obs}$ corresponding to the first r close entries (i.e., $d_{i(1)}, \dots, d_{i(r)}$) in $\boldsymbol{d}_i$.
   2.4) Assign their observed values as the r candidates for the missing value $Z_{(1),i}^{miss}$.
   end for *i*

Using **Algorithm 6.4**, each missing value in $\boldsymbol{Z}_{(1)}^{miss}$ has r candidates to impute. For each missing value, randomly sample one of the r candidates and impute the missing value. After all the missing values in $\boldsymbol{Z}_{(1)}^{miss}$ are imputed in the same way, an imputed vector is obtained, which is denoted as $\widehat{\boldsymbol{Z}}_{(1)}^{miss}$. Then, continue this procedure within the remaining $r - 1$ candidates for each missing value until all candidates are used. Finally, there will be r imputed $\widehat{\boldsymbol{Z}}_{(1)}^{miss}$, which is denoted as $\left\{\widehat{\boldsymbol{Z}}_{(1),l}^{miss}\right\}_{l=1}^{r}$. Each combination $\left(\boldsymbol{Z}_{(1)}^{obs}, \widehat{\boldsymbol{Z}}_{(1),l}^{miss}\right)$, $l = 1, \dots, r$ forms an imputed $\boldsymbol{Z}_{(1)}$ vector, which is denoted as $\widehat{\boldsymbol{Z}}_{(1),l}$. Therefore, r imputed $\widehat{\boldsymbol{Z}}_{(1)}$ vectors are formed, which is denoted as $\left\{\widehat{\boldsymbol{Z}}_{(1),l}\right\}_{l=1}^{r}$. To impute the missing values in $\boldsymbol{Z}_{(2)}$, $\boldsymbol{U}_1 = \boldsymbol{X}^{obs}$ is updated by $\boldsymbol{U}_{2,l} = \left(\boldsymbol{X}^{obs}, \widehat{\boldsymbol{Z}}_{(1),l}\right)$, $l = 1, \dots, r$. Then, **Algorithm 6.5** is developed to impute $\boldsymbol{Z}_{(j)}$, $j = 2, \dots, q$ in a sequential way.

Algorithm 6.5: Sequentially impute the missing values for $\boldsymbol{Z}_{(j)}, j = 2, \ldots, q$

Given the $\{\boldsymbol{U}_{2,l}\}_{l=1}^{r}$, where $\boldsymbol{U}_{2,l} = (\boldsymbol{X}^{obs}, \widehat{\boldsymbol{Z}}_{(1),l})$.
for all $l = 1, \ldots, r$ **do**
  for all $j = 2, \ldots, q$ **do**
    1) Compute the model parameters $(\bar{\boldsymbol{\beta}}_j, \bar{\sigma}_j^2)$ using Eqs. (6.12-6.16) with the replacement of variables and parameters for $\boldsymbol{Z}_{(j)}$, i.e., $p_j(\boldsymbol{Z}_{(j)}|\boldsymbol{U}_{j,l}, \boldsymbol{\beta}_j, \sigma_j^2)$.
    2) Generate r realistic candidates for each missing value in $\boldsymbol{Z}_{(j)}^{miss}$ using **algorithm 6.4** with the replacement of variables and parameters for $\boldsymbol{Z}_{(j)}$.
    3) Randomly select a candidate for imputing each missing value in $\boldsymbol{Z}_{(j)}^{miss}$.
    4) Denote the finally imputed $\boldsymbol{Z}_{(j)}$ as $\widehat{\boldsymbol{Z}}_{(j),l}$ and update the $\boldsymbol{U}_{j+1,l} = (\boldsymbol{U}_{j,l}, \widehat{\boldsymbol{Z}}_{(j),l})$.
  end for *j*
  5) Set $\widehat{\boldsymbol{D}}_l = (\boldsymbol{X}_l^{impute}, \boldsymbol{y})$, where $\widehat{\boldsymbol{D}}_l$ is an *imputed*, complete dataset and $\boldsymbol{X}_l^{impute} = \boldsymbol{U}_{q+1,l} = (\boldsymbol{U}_{q,l}, \widehat{\boldsymbol{Z}}_{(q),l})$.
end for *l*

By implementing **Algorithm 6.5**, one can obtain r *imputed*, complete datasets $\{\widehat{\boldsymbol{D}}_l\}_{l=1}^{r}$. Next, we use a k-fold cross-validation (CV) algorithm to minimize a cost function and determine which imputed dataset is the most plausible based on a data-driven model (ML technique). The following procedure is used to select the most plausible imputed dataset, which is defined as the one capable of minimizing the cost function $CF\left(\boldsymbol{y}, f\left(\boldsymbol{X}^{impute}\right)\right)$ by a k-fold cross-validation procedure, where $CF(\cdot)$ represents the cost function and $f(\cdot)$ represents an ML technique:

Algorithm 6.6: Selection of the most plausible imputed dataset by K-fold CV procedure

Given the r imputed datasets $\{\widehat{\boldsymbol{D}}_l\}_{l=1}^{r}$, where $\widehat{\boldsymbol{D}}_l = (\boldsymbol{X}_l^{impute}, \boldsymbol{y})$, cost function $CF(\cdot)$, ML technique $f(\cdot)$.
for all $l = 1, \ldots, r$ **do**
  1) Compute the cost by K-fold CV procedure:
  $CV_{K-fold}(\widehat{\boldsymbol{D}}_l) = \frac{1}{K}\sum_{k=1}^{K} CF\left(\boldsymbol{y}_{n_k}, f\left(\boldsymbol{X}_{n_k,l}^{impute}\right)\right)$.
end for *l*
2) Choose the imputed dataset that has the $min\left(\{CV_{K-fold}(\widehat{\boldsymbol{D}}_l)\}_{l=1}^{r}\right)$.

In **Algorithm 6.6**, n_k is the size of the kth group (i.e., $n_k = floor(n/K)$); $\boldsymbol{y}_{n_k}$ is the observed response variable for the kth group in terms of the lth *imputed*, complete dataset $\widehat{\boldsymbol{D}}_l$; $f\left(\boldsymbol{X}_{n_k,l}^{impute}\right)$ is the predicted response for the kth group by an ML technique $f(\cdot)$ trained on $\left(\boldsymbol{X}_{-n_k,l}^{imputed}, \boldsymbol{y}_{-n_k}\right)$ in terms of $\widehat{\boldsymbol{D}}_l$; $\left(\boldsymbol{X}_{-n_k,l}^{imputed}, \boldsymbol{y}_{-n_k}\right)$ is the complementary set of $\left(\boldsymbol{X}_{n_k,l}^{imputed}, \boldsymbol{y}_{n_k}\right)$ in $\widehat{\boldsymbol{D}}_l$.

Using **Algorithms 6.4 – 6.6**, the most plausible *imputed*, complete dataset can be determined. The m most plausible *imputed*, complete datasets to constitute an ensemble can be created for MI analyses to account for the uncertainty of missing data by independently repeating **Algorithms 6.4 – 6.6** m times. Each *imputed*, complete dataset can be used to develop an analytical model, and thus m analytical models forming an ensemble can be developed for predictions. The final predicted results are the average of the predicted results of m models. A schematic flowchart is presented in Figure 6.4 to illustrate this procedure.

6.3.2 Design of two case studies

This section presents the details of the numerical experiment design and validation for the performance of the *SRB-PMM* in practical applications in CE. Two case studies are performed. The first case study is to evaluate the capabilities of the proposed *SRB-PMM* in improving the lateral strength prediction performance of a data-driven model based on an RC column dataset subjected to ten different missing data ratios. The second one serves to illustrate the practical application of the *SRB-PMM* in post-earthquake structural analysis when the target damaged building is missing critical structural information. The detailed information is introduced below.

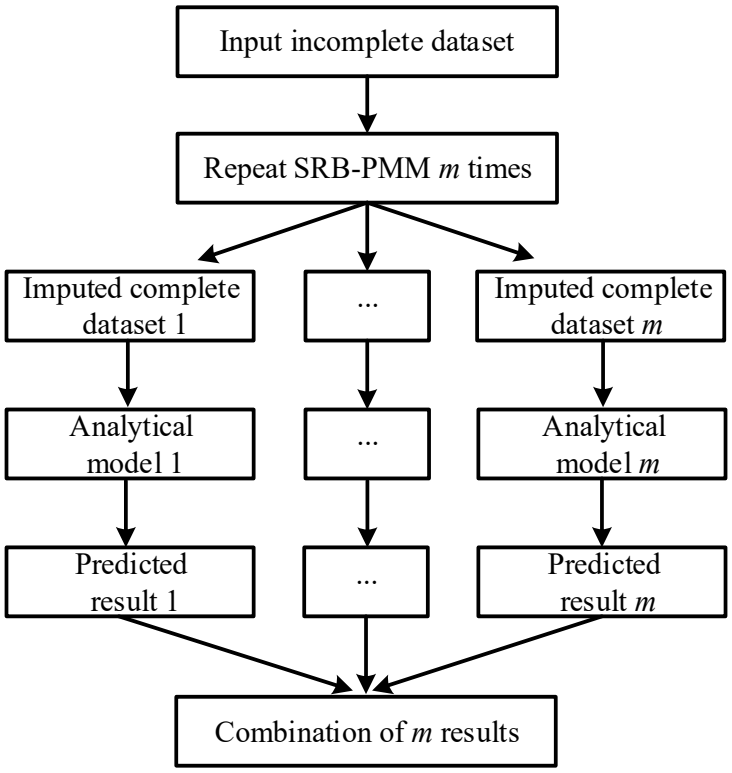


Figure 6.4 Schematic flowchart for the prediction based on an ensemble of m data-driven models.

In structural engineering, RC columns are the primary lateral load-carrying structural member to effectively resist earthquake loads. The lateral strength of an RC column is a critical factor when quantifying the seismic performance of the overall structure. Thus, it is important to accurately predict the lateral strength of RC columns in structural engineering. The *RC column* dataset presented in **Chapter III** (with full details in Appendix A) is used to perform the two case studies. There are ten features or predictors used in this study: the column gross sectional area A_g (calculated by $b \times h$, where b is column section width and h is column section depth), concrete compressive strength f_c', column cross-sectional effective depth d, longitudinal reinforcement yield stress f_{yl}, longitudinal reinforcement area A_{sl}, transverse reinforcement yield stress f_{yt}, transverse reinforcement area A_{st}, stirrup spacing s, shear span a, and applied axial load P. The

response variable is the lateral strength V_m, which is defined as the maximum shear force in the hysteretic force-displacement curve.

Since the RC column dataset does not contain missing values, the case studies are performed on synthetic incomplete datasets. For Case Study 1, synthetic incomplete datasets with ten different missing data ratios are generated from the complete column dataset to comprehensively test the performance of the proposed *SRB-PMM* approach. The performance of the proposed approach is also compared with the two widely used MI methods mentioned previously: *JM* and *FCS*. For Case Study 2, an RC column randomly sampled from the RC column dataset serves as an example of the target damaged building which is hypothetically missing some critical structural information when surveyed in a post-earthquake state. The randomly sampled RC column's critical feature information regarding the material strength and reinforcement details is necessary to build the numerical model for further seismic analysis; however, in this case study, this information is removed and thus assumed unknown. The proposed *SRB-PMM* approach will be used to impute this critical feature information. The seismic analysis results obtained from the imputed information will be compared with experimentally observed results to illustrate the practical application of the *SRB-PMM* approach. The detailed information regarding the designs of these two case studies is presented in **Sections 6.3.2.1** and **6.3.2.2**.

6.3.2.1 Design of case study 1

The purpose of this first case study is to evaluate the capability of the *SRB-PMM* approach in improving the lateral strength prediction performance of a data-driven model based on the mentioned RC column dataset subjected to ten different missing data ratios and thus, to investigate how the missing data ratio affects its performance. The synthetic incomplete datasets are generated in the following way. First, for the original complete RC column dataset, we use the 10-fold cross-

validation procedure (**Section 3.4.2**) to generate ten different training and test sets where the ten test sets are mutually exclusive. Then, we select ten missing data ratios: 5%, 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50%. For each missing data ratio, we generate an incomplete training set from each original complete training set by randomly sampling observations. The number of sampled observations equals the $ceil(missing\ ratio \times n_{tr})$, where n_{tr} is the size of the training set. Given the sampled observations, we randomly sample half of the column features (or predictors) (i.e., five features), which serve as the fully observed explanatory variables. The remaining half of the column features serve as the partially observed explanatory variables (e.g., it could be the concrete compressive strength, reinforcement yield stress, or other features). The number of explanatory variables with missing values for each sampled observation is set randomly between 1 and 5. Following these steps, a synthetic incomplete training set can be generated from the complete training set. Note that the 10 mutually exclusive test sets for each missing data ratio are held constant (i.e., missing data is only applied to the training set).

The least squares support vector machines for regression (LS-SVMR) technique is used to construct the data-driven model employed in this work. Five types of data-driven models are designed. *Delete-LS-SVMR* is established as a baseline, where the data-driven model is developed based on the *reduced,* complete training set formed by deleting the observations with missing values in the incomplete training set. *SRB-PMM-LS-SVMR* is the data-driven model developed using the proposed *SRB-PMM* method where the incomplete training set is first imputed using the *SRB-PMM* approach presented in **Section 6.3.1** and then the *SRB-PMM-LS-SVMR* is developed based on the *imputed,* complete training set. The third and fourth data-driven models developed in this work are established to thoroughly compare the performance of the proposed approach with existing, popular MI approaches. *JM-LS-SVMR* and *FCS-LS-SVMR* are developed using the *JM*

(with a multivariate normal model) (Schafer 1997) and *FCS* (with a univariate normal model) (Buuren and Groothuis-Oudshoorn 2010) imputation methods, respectively. The final data-driven model, *Complete-LS-SVMR,* is employed as an experimental benchmark (or ground truth), where the original complete training set is used to develop the data-driven model. The ten test sets for all five data-driven models are the same, as introduced above. For each developed data-driven model, the final performance is evaluated by taking the average of the ten tests.

6.3.2.2 Design of case study 2

In the second case study, the objective is to illustrate the practical application of the *SRB-PMM* approach in expediting post-earthquake structural evaluations when critical structural information required for seismic analysis is missing. The rectangular RC column dataset (**Appendix A**) is also used in this case study. Specifically, we first randomly sample an RC column from the 262 column specimens, and this column then serves as the target structure with missing critical feature information. The critical feature information considered in this case study is the concrete compressive strength f_c', longitudinal reinforcement yield stress f_{yl}, longitudinal reinforcement area A_{sl}, transverse reinforcement yield stress f_{yt}, and transverse reinforcement area A_{st}. This is because these features may easily be missing from field surveys, whereas the feature information regarding the column geometry may more easily be extracted in a routine evaluation. Thus, the information pertaining to these five features is assumed unknown for the sampled column and requires imputation before a seismic analysis can be carried out. The synthetic incomplete datasets are generated based on the remaining 261 column specimens in a similar way as in *Case Study 1* but with two differences. The first difference is that this case study only has one incomplete dataset for each missing data ratio and does not have the split of training and test sets. The second

difference is regarding the partially observed explanatory variables. In this case study, the partially observed explanatory variables are restricted to the aforementioned five features.

In this case study, we limit the missing data ratio to 5% and 10%. Therefore, in total, there are two synthetic incomplete datasets. The sampled column missing the information pertaining to the five critical features is then added to these two synthetic, incomplete datasets. Then, the *SRB-PMM* method is used to impute the missing values in the synthetic, incomplete datasets. After all the missing values are imputed, we then use the imputed feature information along with the known feature information (e.g., column geometry) to perform a seismic analysis of this sampled column. The performance of the *SRB-PMM* method is evaluated by comparing the imputed sampled column's simulated seismic response with its experimentally observed response in terms of the hysteretic force-displacement relation.

6.3.3 Case study implementation

To implement the *SRB-PMM* method, some parameters introduced in **Section 6.3.1** need to be established. The number of close cases r is set to five. The m most plausible candidates is set to three. The cost function (i.e, $CF(\cdot)$) is mean squared error (MSE), which is evaluated by the LS-SVMR based on the 10-fold cross-validation procedure (**Section 3.4.2**). The detailed implementation of the *JM* and *FCS* methods can be found in Schafer (1997) and Buuren and Groothuis-Oudshoorn (2010). The m candidates to account for the uncertainty of missing data for the *JM* and *FCS* methods is also set to three. All codes are implemented in Matlab. To illustrate the post-earthquake structural evaluation, OpenSees (Mazzoni et al. 2006) is used to perform the seismic analysis of the sampled column with the imputed feature information.

6.3.4 Numerical results

In this section, the experimental results of the two case studies are presented. For the first case study, results pertaining to the performance of the five data-driven models, *SRB-PMM-LS-SVMR, FCS-LS-SVMR, JM-LS-SVMR, Delete-LS-SVMR,* and *Complete-LS-SVMR* are all presented. Further, the investigation of how the missing data ratio affects the performance of these data-driven models in terms of R^2, RMSE, and MAE is presented. For the second case study, the hysteretic force-displacement relation of the sampled RC column obtained with the imputed critical feature information is compared with the experimentally observed results for the same column. At last, a discussion regarding the proposed *SRB-PMM* approach is presented.

6.3.4.1 Results for case study 1

The results for each missing data ratio are averaged to reflect the performance of *SRB-PMM-LS-SVMR, FCS-LS-SVMR, JM-LS-SVMR,* and *Delete-LS-SVMR* in terms of the average R^2, RMSE, and MAE metrics. The average R^2, RMSE, and MAE values across ten different missing data ratios are reported in Figure 6.5. Note that the results for *Complete-LS-SVMR* do not fluctuate with the variation of missing data ratios since the *Complete-LS-SVMR* is developed based on the original complete training set and serves as the benchmark for this work. By observation of Figure 6.5, the results for *SRB-PMM-LS-SVMR, FCS-LS-SVMR, JM-LS-SVMR,* and *Delete-LS-SVMR* show that the average RMSE and MAE values increase globally (though some values decrease locally) with increasing missing data ratios, and the average R^2 values decrease globally (though some values increase locally) with increasing missing data ratios. This phenomenon suggests that the performance of all imputation methods are inversely related to the missing data ratio, which is to be expected. Additionally, compared to the results of *Delete-LS-SVMR*, the proposed *SRB-PMM-LS-SVMR* improves the prediction performance for all ten missing data ratios, while both

JM-LS-SVMR and *FCS-LS-SVMR* degrade the prediction performance in some cases. Moreover, the obvious difference between *Delete-LS-SVMR* and *Complete-LS-SVMR* suggests that directly deleting the observations with missing values is not an effective way to handle the missing data since it reduces the prediction performance of the data-driven modeling procedure substantially.

To further investigate these findings, the following criteria (Kang 2013) are used to quantify the R^2, RMSE, and MAE improvements (%) versus discarding the observations with missing values, for each imputation method, across the ten different missing data ratios. The R^2, RMSE, and MAE improvements (%) are calculated as the following:

$$R^2 \text{ improvement (\%)} = 100 \times \left(\frac{R^2 \text{ with imputation}}{R^2 \text{ without imputation}} - 1\right) \tag{6.17}$$

$$\text{RMSE improvement (\%)} = 100 \times \left(1 - \frac{\text{RMSE with imputation}}{\text{RMSE without imputation}}\right) \tag{6.18}$$

$$\text{MAE improvement (\%)} = 100 \times \left(1 - \frac{\text{MAE with imputation}}{\text{MAE without imputation}}\right) \tag{6.19}$$

Note that the improvement is not calculated using the average R^2, RMSE, and MAE values for each missing data ratio. The improvement for each missing data ratio is first calculated based on the original R^2, RMSE, and MAE values. Then, the calculated improvements are averaged to reflect the average prediction performance improvements of *SRB-PMM-LS-SVMR, FCS-LS-SVMR*, and *JM-LS-SVMR* in comparison to *Delete-LS-SVMR*. The average improvements in terms of R^2, RMSE, and MAE are reported in Table 6.3. Then, the average improvements are employed to compare the prediction performance of *SRB-PMM-LS-SVMR, FCS-LS-SVMR,* and *JM-LS-SVMR*. The greater the average improvements, the better the performance of imputation methods. By observation of Table 6.3, it is found that, in most cases, the proposed *SRB-PMM-LS-SVMR* outperforms both *JM-LS-SVMR* and *FCS-LS-SVMR* and achieves the best improvement in prediction performance, meaning that the proposed *SRB-PMM* imputation method possesses the

best performance in most cases. Further, the proposed *SRB-PMM* method always improves the prediction performance, which is demonstrated by all positive values in Table 6.3. Both *JM* and *FCS* occasionally degrade the prediction performance, which is illustrated by the appearance of some negative values in Table 6.3. The performance degradation of both *JM* and *FCS* may be attributed to the meaningless imputations induced by simulated candidates outside of the observed data range.

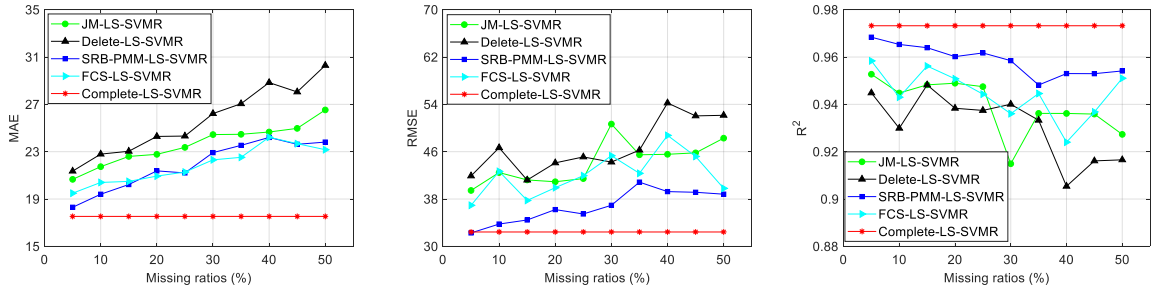


Figure 6.5 The performance comparison of *SRB-PMM-LSSVMR, FCS-LS-SVMR, JM-LS-SVMR, Delete-LS-SVMR,* and *Complete-LS-SVMR* in terms of the average R^2, RMSE, and MAE metrics versus ten missing data ratios.

Table 6.3 The average performance improvement versus discarding observations with missing values across ten missing data ratios in terms of R^2, RMSE, and MAE. The bold values represent the best performance improvements.

Indicators	Models	5%	10%	15%	20%	25%	30%	35%	40%	45%	50%
R^2	JM-LS-SVMR	0.83	1.60	0.00	1.13	1.07	-2.67	0.31	3.40	2.16	1.17
	FCS-LS-SVMR	1.43	1.41	0.84	1.33	0.73	-0.43	1.21	2.06	2.26	3.76
	SRB-PMM-LS-SVMR	**2.49**	**3.81**	**1.66**	**2.33**	**2.59**	**1.96**	**1.60**	**5.26**	**4.02**	**4.10**
RMSE	JM-LS-SVMR	5.86	8.99	-0.05	7.24	8.11	-14.52	1.71	15.99	12.01	7.41
	FCS-LS-SVMR	11.84	8.56	8.35	9.51	7.02	-2.43	8.49	10.10	13.17	23.67
	SRB-PMM-LS-SVMR	**22.99**	**27.64**	**16.37**	**17.90**	**21.37**	**16.52**	**11.71**	**27.62**	**24.81**	**25.57**
MAE	JM-LS-SVMR	3.25	4.65	1.90	6.23	3.94	6.77	9.55	14.48	10.98	12.43
	FCS-LS-SVMR	8.73	10.50	11.06	**13.80**	12.51	**14.92**	**16.74**	16.02	15.59	**23.52**
	SRB-PMM-LS-SVMR	**14.30**	**14.87**	**12.20**	11.99	**12.81**	12.58	13.01	**16.07**	**15.78**	21.40

6.3.4.2 Results for case study 2

An RC column (specimen No. 6 in Tanaka and Park 1990) is randomly sampled from the column dataset. The column's critical feature information introduced in **Section 6.3.2.2** is assumed unknown and requires imputation prior to any seismic analysis. The missing data ratios considered in this case study are limited to 5% and 10%, as introduced in **Section 6.3.2.2**. After the synthetic, incomplete datasets are generated, the *SRB-PMM* approach is run independently three times for each missing data ratio to account for uncertainty due to the missing data. For each run, a group of the most plausible candidates for the five missing values can be generated. The seismic analysis for the sampled column is then based on the imputed feature information. Figures 6.6(a,b,c) and 6.7(a,b,c) present the imputed values and the seismic analysis results of the sampled column generated from the synthetic incomplete column datasets with 5% and 10% missing data ratios, respectively. Figures 6.6(d) and 6.7(d) show the average of the simulated results to account for the uncertainty due to the missing data.

By observation of Figures 6.6(a,b,c) and 6.7(a,b,c), it is evident that it is necessary to account for the uncertainty due to the missing data. This is because, although a single run may produce a good result, it can also produce significant bias. For example, for the 5% missing data ratio, Figures 6.6(a,c) show that the seismic analysis results underestimate the actual seismic performance of the sampled column, while the results presented in Figure 6.6(b) overestimates the true seismic performance; and for the 10% missing data ratio, the results in Figures 6.7(a,c) overestimate the actual seismic performance in spite of Figure 6.7(b) showing a good estimation. Thus, it is hard to judge which single run is a reasonable estimation before knowing the actual seismic performance. However, once considering the uncertainty, the estimation can be justified even if the actual seismic performance is unknown. Both Figures 6.6(d) and 6.7(d) account for the

uncertainty of the missing data, and these results show reasonable estimations. Therefore, this case study demonstrates that the proposed *SRB-PMM* method performs well for the incomplete column datasets with 5% and 10% missing data ratios, which in turn illustrates its practical application in post-earthquake structural evaluation subjected to missing data problems.

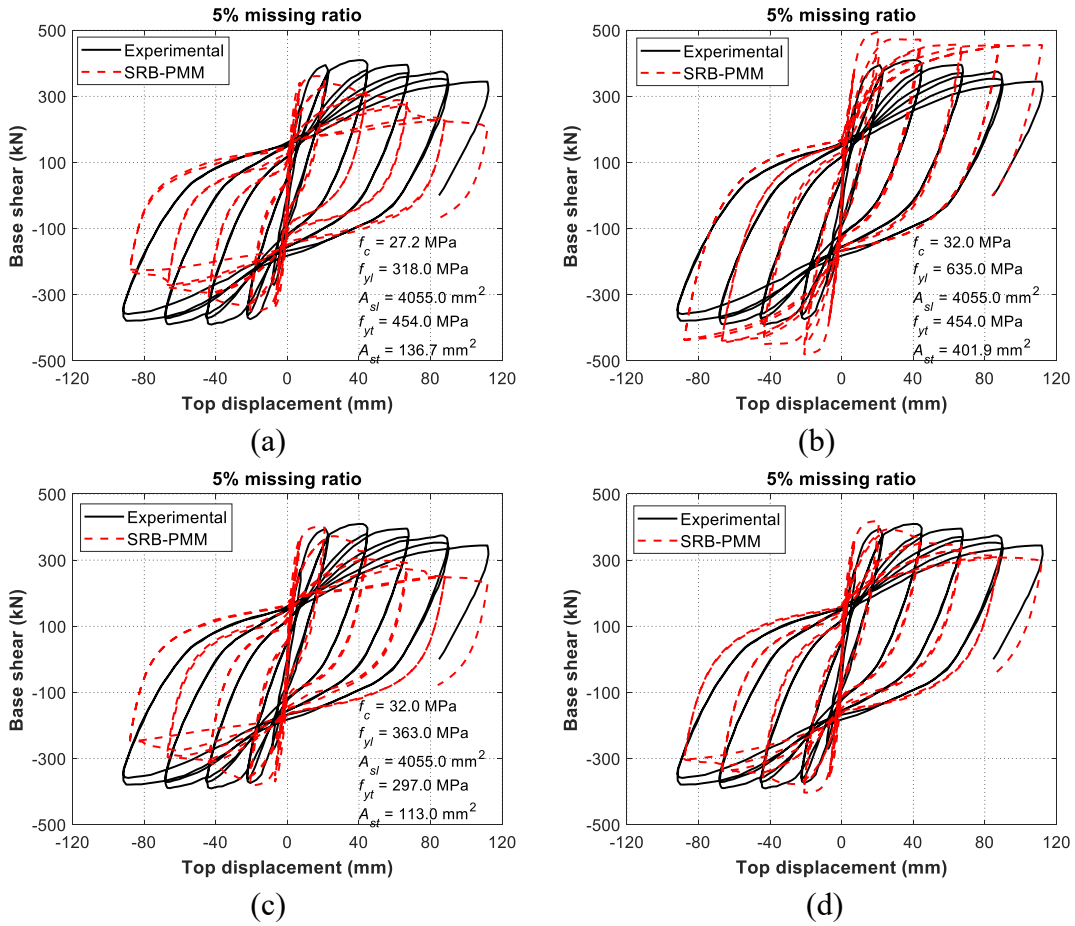


Figure 6.6 Seismic analysis result for the sampled RC column missing critical feature information. (a), (b), and (c) are the three results comparison between the experimental and simulated results, and the three simulated results are obtained from the three imputed information presented on the figures using the *SRB-PMM* based on the column dataset with 5% missing data ratio. The simulated result in (d) is taking the mean of the three simulated results to account for the uncertainty due to the missing data.

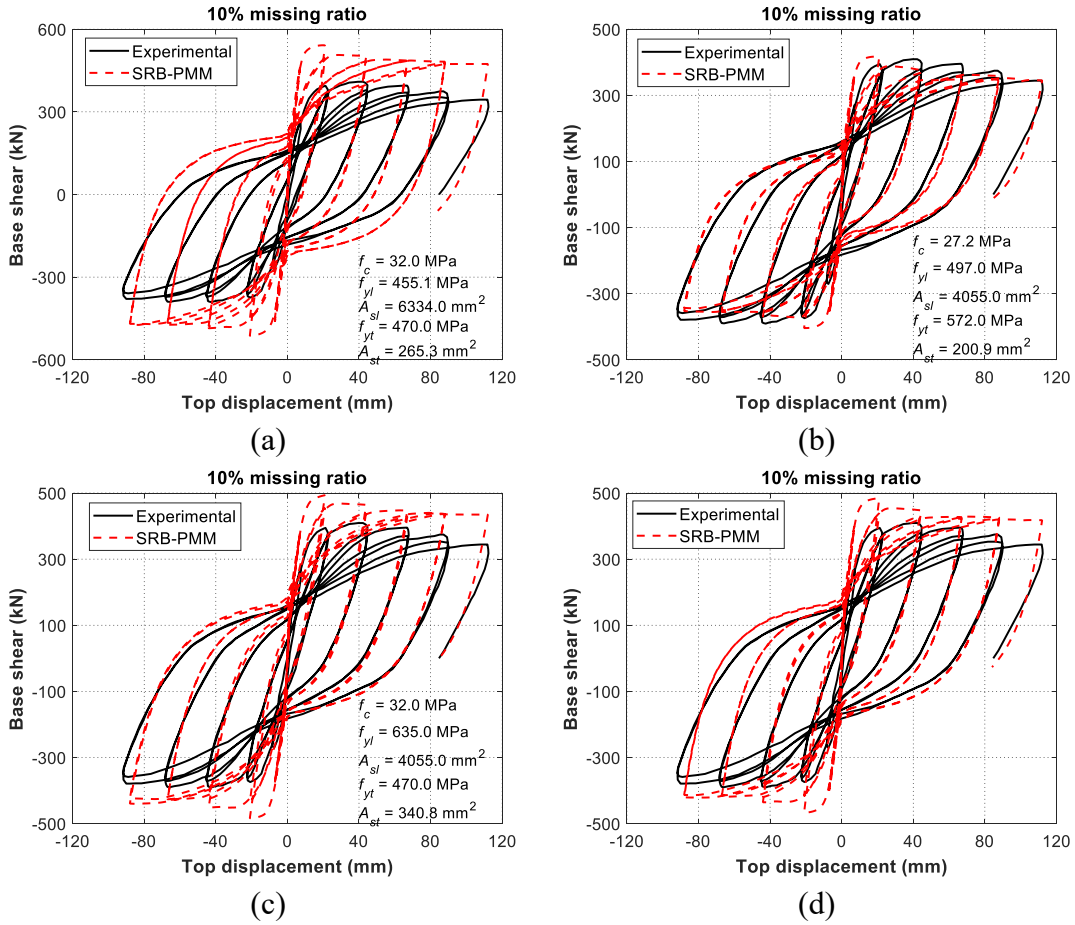


Figure 6.7 Seismic analysis result for the sampled RC column missing critical feature information. (a), (b), and (c) are the three results comparison between the experimental and simulated results, and the three simulated results are obtained from the three imputed information presented on the figures using the SRB-PMM based on the column dataset with 10% missing data ratio. The simulated result in (d) is taking the mean of the three simulated results to account for the uncertainty due to the missing data.

Results from these two case studies demonstrate that the proposed *SRB-PMM* method is able to generate realistic, valid candidates for the missing values, without risking meaningless imputations as is characteristic of existing, popular imputation approaches. The first case study further illustrates that the proposed *SRB-PMM* method enhances the lateral strength prediction performance of the data-driven model when compared to the baseline model (*Delete-LS-SVMR*). It can also be concluded that when the missing data ratio is less than 10%, the proposed *SRB-PMM*

method can generate valid candidates, which yields the *SRB-PMM-LS-SVMR model,* trained on the imputed dataset, having comparable performance to the model formed using the original complete training set (i.e., *Complete-LS-SVMR*). The second case study validates the practical application of the *SRB-PMM* in seismic performance estimation of RC columns when information regarding critical features is missing. These results demonstrate the wide-scale capabilities of the proposed data-driven modeling framework towards expediting post-disaster structural evaluations, where all critical structural properties may not be known in the field.

As the *SRB-PMM* method is a multiple imputation (MI) method, the uncertainty due to the missing data is also incorporated into the final structural analyses. On the basis of these two case studies, the results show that by independently running the method three times, it is sufficient to cover the variation induced by the uncertainty of the missing data. Therefore, based on the two case studies, it can be concluded that the proposed *SRB-PMM* method is a useful and effective tool to handle missing data problems in CE applications.

6.4 Solution to Small Datasets

As introduced in **Section 2.4.3**, small datasets typically have unignorable sample bias and can lead to a fully-trained machine learning (ML)model that has a large bias. This section presents a novel regression-based transfer learning (TL) model to address small sample bias. The proposed TL model is termed double-weighted support vector transfer regression (DW-SVTR), as it couples least squares support vector machines for regression (LS-SVMR) with two weight functions. The first weight function uses kernel mean matching (KMM) to reweight the source domain data such that the means of the source and target domain data in a reproduced kernel Hilbert space (RKHS) are close. In this way, the source domain data points relevant to the target domain points have a larger weight than irrelevant source domain points. The second weight is a function of estimated residuals, which aims to further reduce the negative interference of irrelevant source domain points. The proposed approach is assessed and validated by simulated and real datasets, showing that the proposed DW-SVTR can even reduce sample bias and improve prediction performance between two irrelevant domains. The detailed information is presented as follows.

6.4.1 Development of DB-SVTR

Suppose two datasets, $\{(\boldsymbol{x}_j^S, y_j^S)\}_{j=1}^n$ and $\{(\boldsymbol{x}_k^T, y_k^T)\}_{k=1}^m$ where $\boldsymbol{x}_j^S$ and $\boldsymbol{x}_k^T$ are both $\in R^p$ and y_j^S and y_k^T are both $\in R$, are sampled from the source domain distribution $p^S(\boldsymbol{x}, y)$ and the target domain distribution $p^T(\boldsymbol{x}, y)$ respectively, where $\boldsymbol{x} \in R^p, y \in R$. In the proposed TL method, we do not have the pre-assumption that the source and target domains are related. Therefore, the source and target domains could be unrelated (e.g., both the marginal and posterior distributions of the two domains are different). Since the source and target domains could be unrelated and arbitrarily far apart, this means that the units of the predictors and response variables between these two domains may vary greatly, leading to a significant discrepancy in numeric values. In this case,

there is no way to utilize the information from the source domain to improve the prediction for the target domain. Thus, the first step is to eliminate the impact of the different units. For both domains, we first transform the predictors $\boldsymbol{x}_t \in R^p$ and response $y_t \in R$ of the dataset $\{(\boldsymbol{x}_t, y_t)\}_{t=1}^{d}$ to zero mean and unit variance space respectively using the following formulas:

$$\boldsymbol{x}_t = (\boldsymbol{x}_t - \overline{\boldsymbol{x}})./\boldsymbol{\sigma}_{\boldsymbol{x}} \tag{6.20}$$

$$y_t = \frac{y_t - \bar{y}}{\sigma_y} \tag{6.21}$$

where the *"./"* operator means element division of two vectors, as explained in Matlab, $\overline{\boldsymbol{x}} \in R^p$ is the mean of the predictors, $\boldsymbol{\sigma}_{\boldsymbol{x}} \in R^p$ is the standard deviation of the predictors, $\bar{y} \in R$ is the mean of the response variable, $\sigma_y \in R$ is the standard deviation of the response variable.

After successfully transforming the data, the data in both domains will be within the space with zero mean and unit variance. Denote $\boldsymbol{z}_j^S = \left(\boldsymbol{x}_j^S, y_j^S\right)$ is a point from the transformed dataset in the source domain and $\boldsymbol{z}_k^T = (\boldsymbol{x}_k^T, y_k^T)$ is a point from the transformed dataset in the target domain. Since the dataset in the target domain is small and not sufficient in size, it cannot be directly employed to train a good ML model due to the potential sample bias. Thus, we need to borrow the data from the source domain to augment the small dataset in the target domain to reduce the sample bias. Denote $\boldsymbol{z}_i = (\boldsymbol{x}_i, y_i)$ is a point from the augmented dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m+n}$ that is formed by combining the transformed source domain dataset $\left\{\left(\boldsymbol{x}_j^S, y_j^S\right)\right\}_{j=1}^{n}$ and the transformed target domain dataset $\{(\boldsymbol{x}_k^T, y_k^T)\}_{k=1}^{m}$. Given the augmented dataset, the learning objective of the proposed DW-SVTR is to find optimal model parameters $\boldsymbol{w} = (w_1, w_2, \dots, w_h)^T \in R^h$ and $b \in R$ that minimize the following objective function:

$$\text{Minimize: } J(\boldsymbol{w}, e_i) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \frac{1}{2}\gamma \sum_{i=1}^{m+n} \beta(\boldsymbol{z}_i) v(\boldsymbol{x}_i) e_i^2 \tag{6.22}$$

$$\text{Subject to: } \mathrm{y}_i = \boldsymbol{w}^T \varphi(\boldsymbol{x}_i) + b + e_i, i = 1, \dots, (m+n) \tag{6.23}$$

where $e_i \in R, s = 1, \ldots, r$ is the error term; $\gamma \in R$ is a regularization parameter; $\beta(\boldsymbol{z}_i), v(\boldsymbol{x}_i) \in R, i = 1, \ldots, m+n$ are weights that can take any value in the range $[\varepsilon, 1]$, $\beta(\boldsymbol{z}_i)$ is a weight to determine the importance of each data point in the augmented dataset and $v(\boldsymbol{x}_i)$ is a weight, which is a function of residual where data points having large residuals have smaller weights and having small residuals have larger weights; the determination of these two types of weight function will be introduced in detail; $\varepsilon \in R$ is a real number approaching 0; $\varphi(\boldsymbol{x}_i)$ is a feature vector, and $\varphi(\cdot): R^p \to R^h$ is a mapping function from p dimensions to a higher h-dimensional feature space.

If $\beta(\boldsymbol{z}_i)$ takes a value approaching ε, it means that the point $\boldsymbol{z}_i$ is irrelevant to the data points in the target domain and plays a lesser role in the prediction for the target domain; while, if $\beta(\boldsymbol{z}_i)$ takes a value approaching one, it means the point $\boldsymbol{z}_i$ is highly relevant to the target domain and plays an important role in the prediction for the target domain.

The Lagrangian function is established to solve Eq. (6.22) and Eq. (6.23):

$$L(\boldsymbol{w}, b, e_i; \alpha_i) = J(\boldsymbol{w}, e_i) - \sum_{i=1}^{m+n} \alpha_i((\boldsymbol{w})^T \varphi(\boldsymbol{x}_i) + b + e_i - y_i) \tag{6.24}$$

where $\alpha_i \in R, i = 1, \ldots, m+n$ is a Lagrange multiplier (also called support values).

The Karush-Kuhn-Tucker (KKT) conditions for optimality are used by differentiating the variables in Eq. (6.24) above, which results in the following:

$$\begin{cases} \dfrac{\partial L}{\partial \boldsymbol{w}} = 0 \to \boldsymbol{w} = \sum_{i=1}^{m+n} \alpha_i \varphi(\boldsymbol{x}_i) \\ \dfrac{\partial L}{\partial b} = 0 \to 0 = \sum_{i=1}^{m+n} \alpha_i \\ \dfrac{\partial L}{\partial e_i} = 0 \to e_i = \dfrac{\alpha_i}{\gamma v(\boldsymbol{x}_i)\beta(\boldsymbol{z}_i)}, \ i = 1, \ldots, m+n \\ \dfrac{\partial L}{\partial \alpha_i} = 0 \to y_i = (\boldsymbol{w})^T \varphi(\boldsymbol{x}_i) + b + e_i, \ i = 1, \ldots, m+n \end{cases} \tag{6.25}$$

Rearranging Eq. (6.25) and eliminating $\boldsymbol{w}$ and e_i, using a kernel function to replace the inner product of feature vectors, the following matrix equation can be obtained:

$$\begin{bmatrix} 0 & 1 & 1 & \cdots & 1 \\ 1 & K(\boldsymbol{x}_1,\boldsymbol{x}_1)+\dfrac{1}{\gamma v(\boldsymbol{x}_1)\beta(\boldsymbol{z}_1)} & K(\boldsymbol{x}_1,\boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_1,\boldsymbol{x}_{m+n}) \\ 1 & K(\boldsymbol{x}_2,\boldsymbol{x}_1) & K(\boldsymbol{x}_2,\boldsymbol{x}_2)+\dfrac{1}{\gamma v(\boldsymbol{x}_2)\beta(\boldsymbol{z}_2)} & \cdots & K(\boldsymbol{x}_2,\boldsymbol{x}_{m+n}) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & K(\boldsymbol{x}_{m+n},\boldsymbol{x}_1) & K(\boldsymbol{x}_{m+n},\boldsymbol{x}_2) & \cdots & K(\boldsymbol{x}_{m+n},\boldsymbol{x}_{m+n})+\dfrac{1}{\gamma v(\boldsymbol{x}_{m+n})\beta(\boldsymbol{z}_{m+n})} \end{bmatrix} \begin{bmatrix} b \\ \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_{m+n} \end{bmatrix} = \begin{bmatrix} 0 \\ y_1 \\ y_2 \\ \vdots \\ y_{m+n} \end{bmatrix} \tag{6.26}$$

where the kernel function is $\mathrm{K}(\mathbf{x}_\mathrm{i},\mathbf{x}_\mathrm{t}) = \varphi^\mathrm{T}(\mathbf{x}_\mathrm{i})\varphi(\mathbf{x}_\mathrm{t})$, $\mathrm{i} = 1,\dots,\mathrm{m}+\mathrm{n};\ \mathrm{t} = 1,\dots,\mathrm{m}+\mathrm{n}$.

For the determination of $\beta(\boldsymbol{z}_i) \in R, i = 1,\dots,m+n$, for each data point in the augmented dataset, we wish to accord points relevant to the points in the target domain more weight than irrelevant points. In conjunction with the use of the kernel function, the relevance is evaluated by the Euclidean distance in a reproduced kernel Hilbert space (RKHS). Specifically, in a feature space, data points (e.g.,$\varphi(\boldsymbol{z}_i)$) close to the points (e.g., $(\boldsymbol{z}_k^T)$) in the target domain will acquire more weights than distant points. Since the small dataset in the target domain has already been included in the augmented dataset, the $\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_k^T)$ will be one. Thus, the problem is changed to determine the $\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_j^S)$. To obtain the $\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_j^S)$ for each data point in the source domain, the data points in the source domain are reweighted such that the mean (i.e., $\frac{1}{n}\sum_{j=1}^{n}\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_j^S)\varphi(\boldsymbol{z}_j^S)$) of the weighted data points in the source domain is close to the mean ($\frac{1}{m}\sum_{k=1}^{m}\varphi(\boldsymbol{z}_k^T)$) of the data points in the target domain. Denote $\boldsymbol{\beta} = \{\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_j^S)\}_{j=1}^{n}$ as a weight vector containing the weight for each data point in the source domain. According to the kernel mean matching (KMM) algorithm (Huang et al. 2007; Gretton et al. 2009), the weight vector $\boldsymbol{\beta}$ can be obtained by minimizing the discrepancy between the mean of the weighted source domain data and the mean of the target domain data subjected to two constraints as shown in the following:

$$\boldsymbol{\beta} = arg\min_{\beta}\left\|\frac{1}{n}\sum_{j=1}^{n}\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_j^S)\varphi(\boldsymbol{z}_j^S) - \frac{1}{m}\sum_{k=1}^{m}\varphi(\boldsymbol{z}_k^T)\right\|^2 \tag{6.27}$$

By reformulating Eq. (6.27) and using the kernel function to replace the inner product of feature vectors, the following quadratic programming (QP) problem concerning the two constraints can be formulated:

$$\text{Minimize: } J(\boldsymbol{\beta}) = \frac{1}{2}\,\boldsymbol{\beta}^T \boldsymbol{K}_1 \boldsymbol{\beta} - \boldsymbol{\kappa}^T \boldsymbol{\beta} \tag{6.28}$$

$$\text{Subject to: } \left|\frac{1}{n}\sum_{j=1}^{n} \beta\left(\boldsymbol{z}_i \cap \boldsymbol{z}_j^S\right) - 1\right| \leq \epsilon \qquad 0 \leq \beta\left(\boldsymbol{z}_i \cap \boldsymbol{z}_j^S\right) \leq B, j = 1, \dots, n \tag{6.29}$$

where $\boldsymbol{K}_1 = \boldsymbol{K}_{jt} \coloneqq K\left(\boldsymbol{z}_j^S, \boldsymbol{z}_t^S\right) \in R^{n\times n}, j, t = 1, \dots, n$ is a kernel matrix calculated based on the data in the source domain, $B = 1000$ is the upper boundary to reflect the scope of discrepancy between the source domain distribution $p^S(\boldsymbol{z})$ and the target domain distribution $p^T(\boldsymbol{z})$, $\epsilon = \left(\sqrt{n} - 1\right)/\sqrt{n}$ is the normalization precision, $\boldsymbol{\kappa} \coloneqq \frac{n}{m}\boldsymbol{K}_2 \mathbf{1}_{m\times 1} \in R^n$ where $\boldsymbol{K}_2 = \boldsymbol{K}_{jk} \coloneqq K\left(\boldsymbol{z}_j^S, \boldsymbol{z}_k^T\right) \in R^{n\times m}, j = 1, \dots, n$ and $k = 1, \dots, m$ is a kernel matrix calculated based on the source and target domain data.

After solving the QP problem and normalizing the weight $\boldsymbol{\beta} = \boldsymbol{\beta}/max(\boldsymbol{\beta})$, each data point in the source domain will have a weight $\beta\left(\boldsymbol{z}_i \cap \boldsymbol{z}_j^S\right)$. Since the weight $\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_k^T)$ for each data point in the target domain has been determined, the remaining is to determine the weight $\beta(\boldsymbol{z}_i)$ for each data point in the augmented dataset. The points having a large weight in the augmented dataset will be more relevant to the target domain points than points having a small weight. Additionally, irrelevant data points are equivalent to outliers, as they are distant from the target domain data points (De Brabanter et al. 2009; Mu and Yuen 2015; Rousseeuw and Leroy 1987; Suykens et al. 2002; Yuen and Mu 2012; Yuen and Ortiz 2017). Although these 'outliers' already have small weights, we wish to further reduce their negative effect. Thus, another weight $v(\boldsymbol{x}_i)$, which is a function of residuals, is incorporated too, as presented in Eq. (6.30). By imposing a weight $\beta(\boldsymbol{z}_i)$ to each data point in the augmented dataset, the relevant points will have small residuals while the

irrelevant points or 'outliers' will have large residuals. Points having large residuals will have a small weight $v(\boldsymbol{x}_i)$, whereas points having small residuals will have a large weight $v(\boldsymbol{x}_i)$. Therefore, in this sense, the importance of the relevant points is further emphasized, while that of the irrelevant points is further diminished. According to Suykens et al. (2002), $v(\boldsymbol{x}_i)$ is determined by the following:

$$v(\boldsymbol{x}_i) = \begin{cases} 1 & if\ |e_i/\ \delta| \leq c_1 \\ \frac{c_2 - |e_i/\ \delta|}{c_2 - c_1} & if\ c_1 \leq |e_i/\ \delta| \leq c_2 \\ \varepsilon & otherwise \end{cases} \tag{6.30}$$

where $c_1 = 2.5$, $c_2 = 3$, $\varepsilon = 10^{-4}$, and $\delta = 1.483\text{MAD}(e_i)$ is a robust estimate where MAD is the median absolute deviation and other variables are defined previously.

After solving Eq. (6.26) (Suykens et al. 1999; 2002), the Lagrange multiplier $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_r)$ and b can be obtained, which can then be utilized for prediction in the target domain (e.g., $\boldsymbol{x}^T$) using the following:

$$\hat{y}(\boldsymbol{x}^T) = \sum_{i=1}^{m+n} \alpha_i\, K(\boldsymbol{x}^T, \boldsymbol{x}_i) + b \tag{6.31}$$

The RBF kernel presented in Eq. (4.11) is utilized.

6.4.2 Implementation

The implementation procedure of the proposed DW-SVTR approach for reducing the sample bias of small datasets is summarized as follows:

Algorithm 6.7: Implementation of proposed DW-SVTR model

Require: Training datasets in the source domain $\{\boldsymbol{z}_j^S\}_{j=1}^n = \{(\boldsymbol{x}_j^S, y_j^S)\}_{j=1}^n$ and target domain $\{\boldsymbol{z}_k^T\}_{k=1}^m = \{(\boldsymbol{x}_k^T, y_k^T)\}_{k=1}^m$, test data in the target domain $\boldsymbol{x}^T$, and optimal hyper-parameter combination $(\gamma, \sigma^2, \sigma_\beta^2)$.

1. Initialization stage:
   (a) Transform the training datasets in the source and target domains individually using Eq. (6.20-6.21).
   (b) Record the means $\bar{\boldsymbol{x}}_{tr}^T$, $\bar{y}_{tr}^T$ and standard deviations $\boldsymbol{\sigma}_{\boldsymbol{x}_{tr}^T}$, $\sigma_{y_{tr}^T}$ for the target domain training dataset $\{(\boldsymbol{x}_k^T, y_k^T)\}_{k=1}^m$.
   (c) Combine the transformed datasets in the source and target domains as an augmented dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m+n}$.
2. Reweighting stage:
   (a) Calculate the $\boldsymbol{K}_1$ and $\boldsymbol{\kappa}$ in Eq. (6.28) using Eq. (4.11) with the parameter σ_β^2.
   (b) Set $\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_k^T) = 1, k = 1, \dots, m$.
   (c) Solve Eqs. (6.28-6.29) to obtain $\boldsymbol{\beta} = \{\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_j^S)\}_{j=1}^n$ and normalize it as $\boldsymbol{\beta} = \boldsymbol{\beta}/max(\boldsymbol{\beta})$.
   (d) Combine $\{\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_k^T)\}_{k=1}^m$ and $\{\beta(\boldsymbol{z}_i \cap \boldsymbol{z}_j^S)\}_{j=1}^n$ as $\{\beta(\boldsymbol{z}_i)\}_{i=1}^{m+n}$.
   (e) Set weight $v(\boldsymbol{x}_i)$ in Eq. (6.30) for each data point in the augmented dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{m+n}$ to 1;
   (f) Solve Eq. (6.26) to obtain $\boldsymbol{\alpha}$, b, and compute $e_i = \alpha_i/(\gamma v(\boldsymbol{x}_i)\beta(\boldsymbol{z}_i)), i = 1, \dots, m+n$.
3. Iterative stage:
   Set the maximum iterative number S, tolerance tol, count $s = 0$, and $t = Inf$
   while $t > tol$ && $s < S$ **do**
      (a) Set $\boldsymbol{\alpha}^{(s)} = \boldsymbol{\alpha}$, $b^{(s)} = b$, $e_i^{(s)} = e_i$, and $v^{(s)}(\boldsymbol{x}_i) = v(\boldsymbol{x}_i)$;
      (b) Compute the robust estimate $\delta^{(s)} = 1.483MAD(e_i^{(s)})$;
      (c) Update the weight $v^{(s+1)}(\boldsymbol{x}_i)$ from $\delta^{(s)}$ and $e_i^{(s)}$ using Eq. (6.30);
      (d) Solve Eq. (6.26) to obtain the $\boldsymbol{\alpha}^{(s+1)}$ and $b^{(s+1)}$;
      (e) Update the $e_i^{(s+1)} = \alpha_i^{(s+1)}/\left(\gamma v^{(s+1)}(\boldsymbol{x}_i)\beta(\boldsymbol{z}_i)\right)$;
      (f) Calculate $t = \left\|\boldsymbol{\alpha}^{(s+1)} - \boldsymbol{\alpha}^{(s)}\right\|$;
      (g) Set $\boldsymbol{\alpha} = \boldsymbol{\alpha}^{(s+1)}, b = b^{(s+1)}, e_i = e_i^{(s+1)}$, and $v(\boldsymbol{x}_i) = v^{(s+1)}(\boldsymbol{x}_i)$;
      (h) Set $s = s + 1$
   end while
4. Output stage:
   (a) Transform the data $\boldsymbol{x}^T$ with the recorded mean $\bar{\boldsymbol{x}}_{tr}^T$ and standard deviation $\boldsymbol{\sigma}_{\boldsymbol{x}_{tr}^T}$ using Eq. (6.20);
   (b) Output the final $\boldsymbol{\alpha}$ and b from the stage 3;
   (c) Given $\boldsymbol{\alpha}$ and b, predict the response value $\hat{y}(\boldsymbol{x}^T)$ of the transformed data $\boldsymbol{x}^T$ using Eq. (6.31).
   (d) Transform the predicted $\hat{y}(\boldsymbol{x}^T)$ back by $\hat{y}(\boldsymbol{x}^T) = \hat{y}(\boldsymbol{x}^T) \times \sigma_{y_{tr}^T} + \bar{y}_{tr}^T$.

6.4.3 Numerical results

To thoroughly assess the performance of the proposed DW-SVTR approach, two examples are carried out. First, a simulated example is used to illustrate the general performance for the most challenging case where both the marginal and posterior distributions of the two domains are different. Then, the proposed approach is employed to predict the shear strength of non-ductile reinforced concrete (RC) columns to illustrate the real-world utilization of the approach when sufficient large datasets are not available.

6.4.3.1 Results for simulated datasets

This example is designed to illustrate how the proposed DW-SVTR works in an especially challenging case. In this example, the datasets in the source and target domains are respectively generated from different joint distributions, where both the posterior and the marginal distributions are different i.e., $p^S(\boldsymbol{x}) \neq p^T(\boldsymbol{x})$ and $p^S(y|\boldsymbol{x}) \neq p^T(y|\boldsymbol{x})$. The source domain has a sufficiently large number of data points, while the target domain has a small number of data points. Thus, the target domain has a potentially large sample bias. This case is more challenging as both the predictor and the response values in the datasets between the source and the target domains may be significantly different, more likely leading to the case where there is no relevance between the source and the target domains. It is commonly thought that there is no way to use an ML model trained in one domain to improve the prediction on another, seemingly, completely irrelevant domain. However, the theory presented in the previous section along with the following experimental results demonstrates that the proposed DW-SVTR can still reduce the negative effect induced by the sample bias of small data and improve the predictive performance in this case.

The marginal distributions of the datasets in the source and target domains are assumed as normal and uniform distributions, respectively, where $x^S \sim Normal(8, 3^2)$ and

$x^T \sim Uniform(-5,5)$. The responses for the dataset in the source domain are generated from $y^S = -6x^S + (x^S)^3 + \varepsilon^S$, while those for the dataset in the target domain are generated according to $y^T = x^T + (x^T)^2 + (x^T)^3 + \varepsilon^T$. The error term distribution for the source domain dataset is $\varepsilon^S \sim Normal(0, 200^2)$ and for the target domain dataset is $\varepsilon^T \sim Normal(0, 12^2)$. Thus, in this sense, the posterior and the marginal distributions between the source and the target domains are different. Ten points (red square points in Figure 6.8a) randomly sampled from the target domain serve as the training data in the target domain and 600 points (blue circle points in Figure 6.8b) randomly sampled from the source domain are the training data in the source domain. An individual test dataset including 200 points (green square points in Figure 6.8d) is randomly generated from the target domain. In this example, three analytical cases are considered. In these analytical cases, the training dataset is varied, but the test dataset holds constant: (1) *Target only:* the 10 training sample points in the target domain (red square points in Figure 6.8a) are used to train an ML model, and this trained ML model is then used to predict the 200 test sample points in the target domain (green square points in Figure 6.8d); (2) *Source only:* the 600 training sample points in the source domain (blue circle points in Figure 6.8b) is used to train an ML model, and this ML model is used to predict the 200 test sample points in the target domain; and, (3) *DW-SVTR:* both the 10 and the 600 training sample points are used as the training dataset for the proposed DW-SVTR, and the trained DW-SVTR model is then utilized to predict the 200 test sample points in the target domain. The least squares support vector machines for regression (LS-SVMR) technique is employed for cases (1) and (2). All the hyper-parameters for both LS-SVMR and proposed DW-SVTR in these three cases are obtained using 10-fold cross-validation on the corresponding training data sets.

The experiment is run 10 distinct times by setting different random seeds to comprehensively reflect the performance of the proposed DW-SVTR. A typical representative of the results for 10 runs is presented in Figure 6.8. Figure 6.8(a) shows the small training dataset in the target domain, which only includes 10 training sample points and thus has a potentially large sample bias. Figure 6.8(b) presents the training datasets in the source and target domains that are combined in a coordinate system. It is found that, in Figure 6.8(b), only three points in the target domain are surrounded with the points in the source domain in the original space. This illustrates the significant lack of relevance between the two domains since the data points in the source domain only potentially can reduce the sample bias within the range represented by the three points in the target domain. Figure 6.8(c) shows the combined training dataset in the transformed space. Note that the transformation for the datasets in the source and target domains is first performed separately using Eqs. (6.20-6.21). Then the transformed datasets in the source and target domains are combined, as described in **Algorithm 6.7** in **Section 6.4.2**. It is observed in Figure 6.8(c) that the relevance between the two domains significantly increases after transformation. Figure 6.8(d) shows the results for all three analytical cases. For analytical case 1, from Figure 6.8(d), it is observed that the LS-SVMR model trained with 10 training sample points has a large bias in some areas where the training sample points are not available. This is demonstrated in Figure 6.8(d) by the apparent discrepancy of the blue dashed line (i.e., target only) and the black solid line (i.e., true function) in the areas where the training sample points are not available as shown in Figure 6.8(a). For analytical case 2, as the source domain training dataset is not significantly relevant to the target domain, the LS-SVMR model trained with the 600 source domain training sample points has a significantly large bias for prediction on the target domain. This is illustrated by the significant discrepancy between the magenta dotted line (i.e., source only) and black solid line across almost

all the areas represented by the test dataset in the target domain. For analytical case 3, the proposed DW-SVTR model is trained with the combined training dataset in the transformed space. Since the proposed DW-SVTR model accords more weight to the source domain sample points that are close to the 10 target domain training sample points than the distant source domain points, the proposed approach can borrow more relevant source domain sample points to augment the 10 target domain training sample points, effectively reducing the sample bias without suffering significant negative effects from those distant source domain points. Also, the negative interferences of these distant source domain points are further diminished by the second weight in the proposed DW-SVTR model, as introduced previously.

The obtained three hyper-parameters for the proposed DW-SVTR for this typical representative are presented in Figure 6.8(d). The results predicted by the DW-SVTR (red dash-dot line in Figure 6.8d) agree well with the true function (black solid line in Figure 6.8d), demonstrating that the proposed DW-SVTR can reasonably predict all test sample points in the target domain regardless of the unrelated nature of the two domains, and further, illuminating the powerful TL capabilities of the proposed DW-SVTR approach. The results of 10 random trials for these three analytical cases are presented in Figure 6.9. Figure 6.9(a) and (b) show the predictive performance comparison over the 10 random trials using box plots in terms of R^2 and RMSE, respectively. By observation of Figure 6.9, it is evident that the proposed DW-SVTR (i.e., *Analytical Case 3*) statistically performs the best over the other two analytical cases, and the *Analytical Case 2* (i.e., source only) statistically has the worst performance.

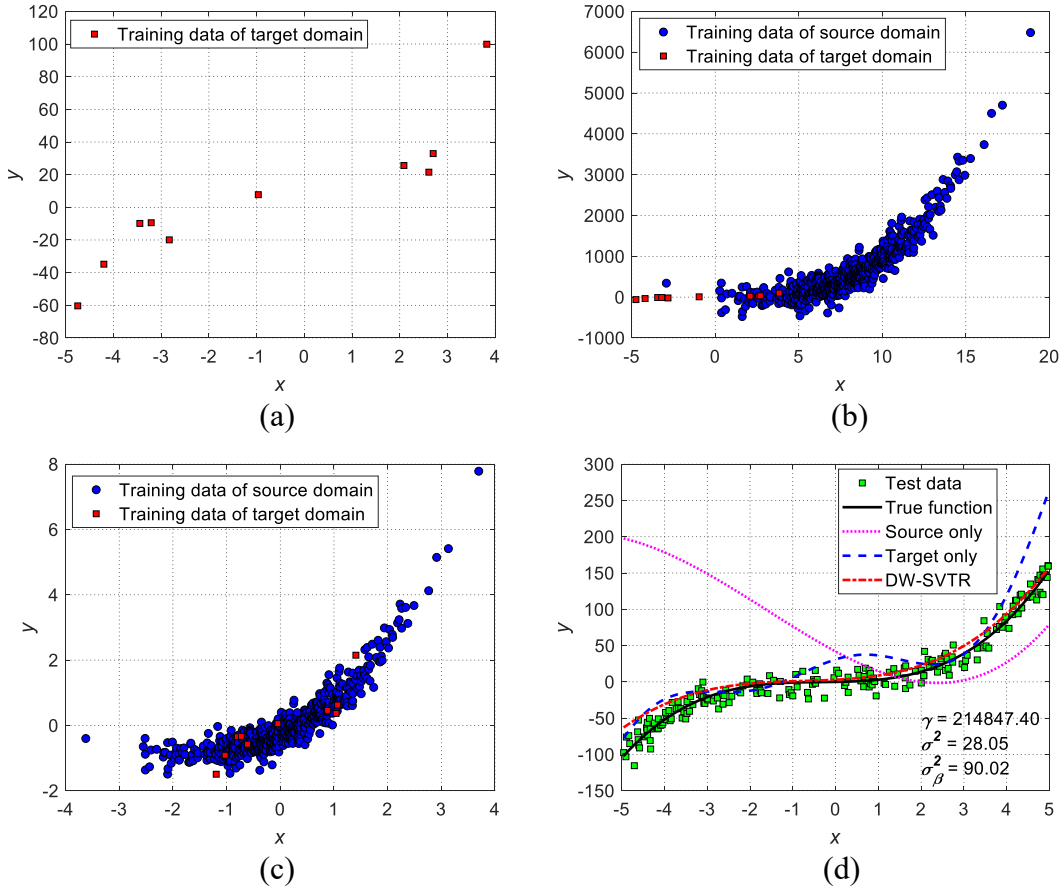


Figure 6.8 A typical representative of 10 random trials for the comparison of the results among three analytical cases. (a) target domain training sample points in the original space; (b) combined source and target domain training datasets in the original space; (c) combined source and target domain training datasets in the transformed space; (d) result comparison of three analytical cases in the original space.

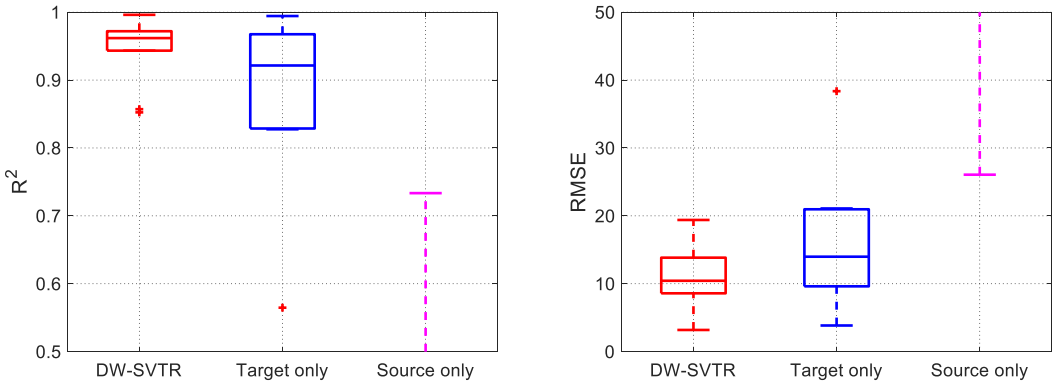


Figure 6.9 Result comparison among three analytical cases over the 10 random trials using box plots in terms of R^2 and RMSE.

6.4.3.2. Results for shear strength prediction of non-ductile RC columns

In structural and earthquake engineering, ductile columns typically have good seismic performance and deformation capacity and will most likely experience flexure failures under large earthquakes, while non-ductile columns often have relatively worse seismic-resistant capacity, leading to flexure-shear and shear failures under earthquakes (Moehle 2014). Non-ductile columns will easily cause the global collapse of RC frame buildings under large earthquakes due to the shear strength deficiency. Thus, it is critical and necessary to identify the shear strength of non-ductile columns before the occurrence of large earthquakes such that these non-ductile columns can be reinforced and retrofitted to enhance their seismic performance avoiding the global collapse of RC frame buildings.

In this example, two-column datasets including rectangular RC columns and circular RC columns (presented in **Chapter III**) are used to further assess the proposed DW-SVTR model in a real-world application. For the rectangular RC column dataset, there are a total of 262 sample points where 208 of them are flexure-critical columns (ductile columns) and the remaining 54 are shear- and flexure-shear-critical columns (non-ductile columns). For the circular RC columns,

there are a total of 160 sample points where 98 of them are ductile columns (i.e., flexure-critical columns) and the remaining 62 are non-ductile columns (i.e., flexure-shear- and shear-critical columns). For each dataset, the input predictors (i.e., explanatory variables) are column gross sectional area (X_1), concrete compressive strength (X_2), column cross-sectional effective depth (X_3), longitudinal reinforcement yield stress (X_4) and cross-sectional area (X_5), transverse reinforcement yield stress (X_6) and cross-sectional area (X_7), stirrup spacing to effective depth ratio (X_8), shear span to effective depth ratio (X_9), and applied axial load (X_{10}), and the response variables are lateral strength (y_1) and drift capacity (y_2) respectively. Thus, for either rectangular or circular section RC columns, the dataset is comprised of the same predictors with two different response variables. More detailed information for the 262 rectangular RC column dataset and for the 160 circular RC column dataset can be found in **Section 3** and **Appendices A and B**.

For each dataset, we select the non-ductile columns as the target domain and the ductile columns as the source domain. The main difference between ductile and non-ductile columns is that the lateral strength for the ductile columns is governed by flexural strength while that for non-ductile columns is dominated by shear strength (Moehle 2014). The lateral strength is defined as the maximum shear force (kN) in the hysteretic force-deformation curve. Ten numerical experiments are designed to sufficiently assess the performance of the proposed DW-SVTR approach based on these two datasets. For each dataset, the task for the target domain will always be the shear strength prediction of non-ductile columns. But the source domain training dataset will vary. The detailed information is as follows.

In this validation, the target domain data is the 54 non-ductile RC rectangular columns with shear strength as the response variable. Five numerical experiments are designed to explore the impact of four different transfer strategies in comparison to one baseline. In *Experiment 1,* the

source domain training dataset is the 208 rectangular ductile columns with flexural strength as the response variable. In *Experiment 2,* the source domain training dataset is the 208 rectangular ductile columns with drift capacity as the response variable. In *Experiment 3,* the source domain training dataset is the 98 circular ductile columns with shear strength as the response variable. In *Experiment 4,* the source domain training dataset is the 98 circular ductile columns with the drift capacity as the response variable. Experiment 5 corresponds to the baseline, where only the target domain training dataset is used and no transfer strategy is applied. It should be noted that the units between lateral strength (kN) and drift capacity (%) are different, which causes the numeric values between them to have a significantly large discrepancy. Further, the reinforcement layouts between rectangular and circular columns are also different. In this sense, *Experiment 1* could be analogous to when the source and target domains have related joint distributions; *Experiment 2* could be analogous to related marginal distributions but unrelated posterior distributions (i.e., $p^S(y|\boldsymbol{x}) \neq p^T(y|\boldsymbol{x})$); *Experiment 3* could be analogous to unrelated marginal distributions (i.e., $p^S(\boldsymbol{x}) \neq p^T(\boldsymbol{x})$) but related posterior distributions; and, *Experiment 4* could be analogous to unrelated marginal and posterior distributions (i.e., $p^S(\boldsymbol{x}) \neq p^T(\boldsymbol{x})$ and $p^S(y|\boldsymbol{x}) \neq p^T(y|\boldsymbol{x})$). Therefore, these five numerical experiments can thoroughly and effectively assess the performance of the proposed DW-SVTR model.

For each experiment, the availability of the target domain training data is apportioned as 10%, 15%, 20%, 25%, 30%, 35%, 40%, 45%, and 50% of total target domain data, and the test set for the target domain will always be 50% of the total target domain data (mutually exclusive from the target domain training data). For each case of data availability, each experiment is run 10 times with different random seeds to measure the performance variability of the proposed DW-SVTR model and ensure the results statistically reliable. It should be noted that for each run, the same

target domain training and test sets are applied for all four experiments and the baseline. For Experiments 1 to 4, the proposed DW-SVTR model is used, while for Experiment 5 (i.e., baseline) an LS-SVMR model is used. The results for these transfer strategies and baseline in each case of target domain data availability are shown in Figure 6.10, where both R^2 and RMSE are taken as the averages of the R^2 and RMSE over the 10 random trials.

From Figure 6.10, it is observed that, compared to the baseline, both R^2 and RMSE (here, R^2 and RMSE are taken as the average of 10 random trials) suggest that the proposed DW-SVTR model significantly improves the prediction performance when the target domain training data is very small (i.e., only has 10% availability). The RMSE is decreased from 109.59 kN (i.e., *baseline*) to 72.34 kN (i.e., *Experiment 1*), 96.22 kN (i.e., *Experiment 2*), 91.02 kN (i.e., *Experiment 3*), and 97.52 kN (i.e., *Experiment 4*), which is equivalent to a reduction of 34%, 12%, 17%, and 11%, respectively. The R^2 is increased from 0.19 (i.e., *baseline*) to 0.62 (i.e., *Experiment 1*), 0.40 (i.e., *Experiment 2*), 0.46 (i.e., *Experiment 3*), and 0.35 (i.e., *Experiment 4*), enhancing the values by 229%, 110%, 142%, and 84%, respectively. With the increase in size of the target domain training data, the prediction performance in terms of averages of RMSE and R^2 over 10 random trials for all five experiments globally increases, and the improved performance by proposed DW-SVTR globally decreases. This is because, with the increase of available target domain training data, the target domain sample bias decreases and thus, the performance difference between the baseline and the proposed approach also decreases. According to different transfer strategies, the improved performance by the proposed DW-SVTR also varies. The most significant performance improvement in terms of both RMSE and R^2 is for *Experiment 1*, followed by *Experiment 3*, and *Experiment 2* is comparable to *Experiment 4*, but both *Experiments 2 and 4* are outperformed by *Experiment 3*. It is worth noting that the proposed DW-SVTR model also works for *Experiment 2*

where the posterior distributions between the source and target domains are unrelated and for *Experiment 4* where both the marginal and posterior distributions are unrelated as introduced previously. This further demonstrates that the proposed approach is effective even if the source and target domains are unrelated. The performance variability over the 10 random trials for all five cases is reported in the boxplots in Figures 6.11 and 6.12. By observation of Figures 6.11 and 6.12, it is observed that, compared to the baseline, the proposed DW-SVTR statistically improves the performance in terms of the median of 10 random trials for all four transfer strategies.

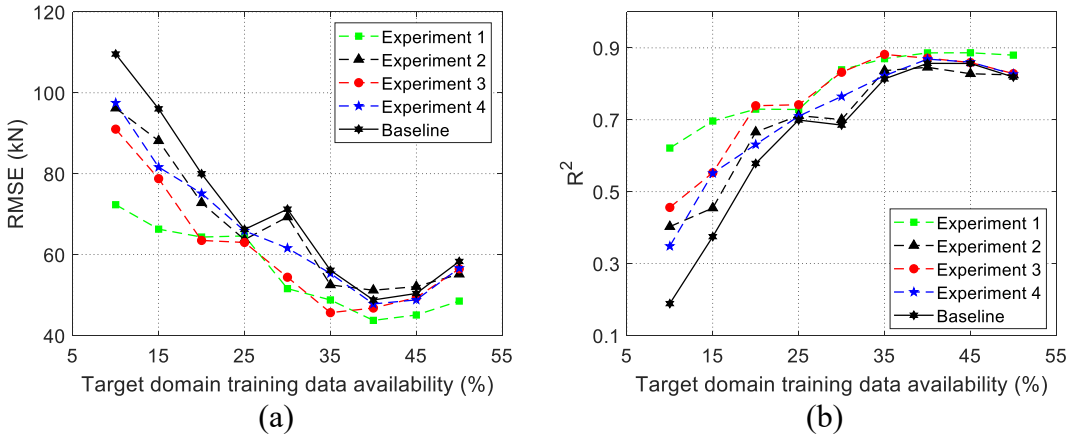


Figure 6.10 Performance versus size of target domain training data availability curve in terms of (a) mean RMSE and (b) mean R^2 for rectangular columns over the 10 random trials.

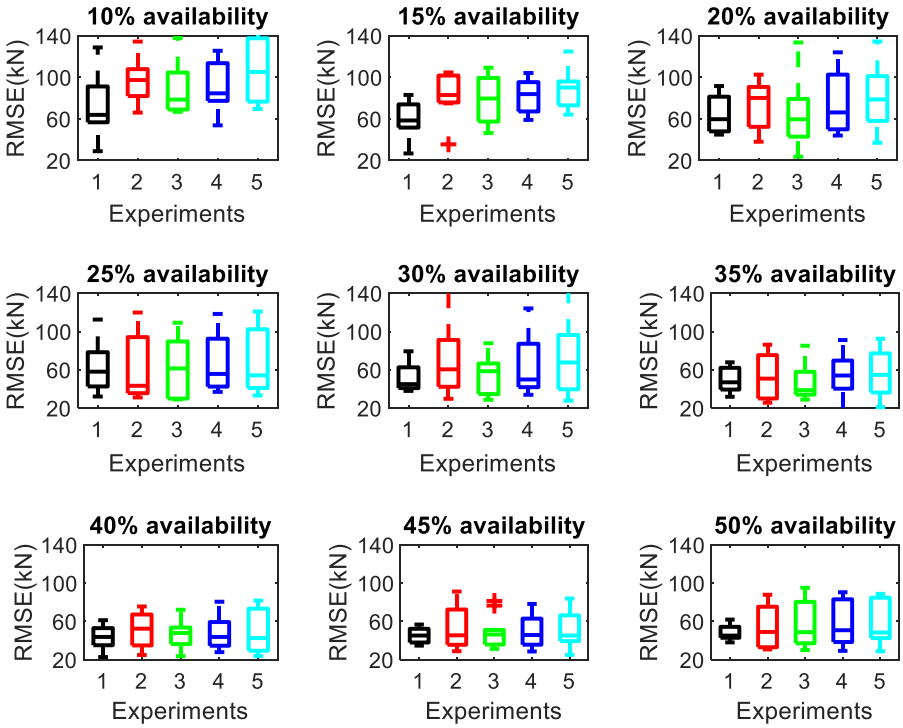


Figure 6.11 Boxplots for rectangular columns over 10 random trials based on four different transfer situations and one baseline in terms of RMSE and the 1, 2, 3, 4, and 5 in the *x*-axis represent the Experiment 1, Experiment 2, Experiment 3, Experiment 4, and Experiment 5 (i.e., Baseline).

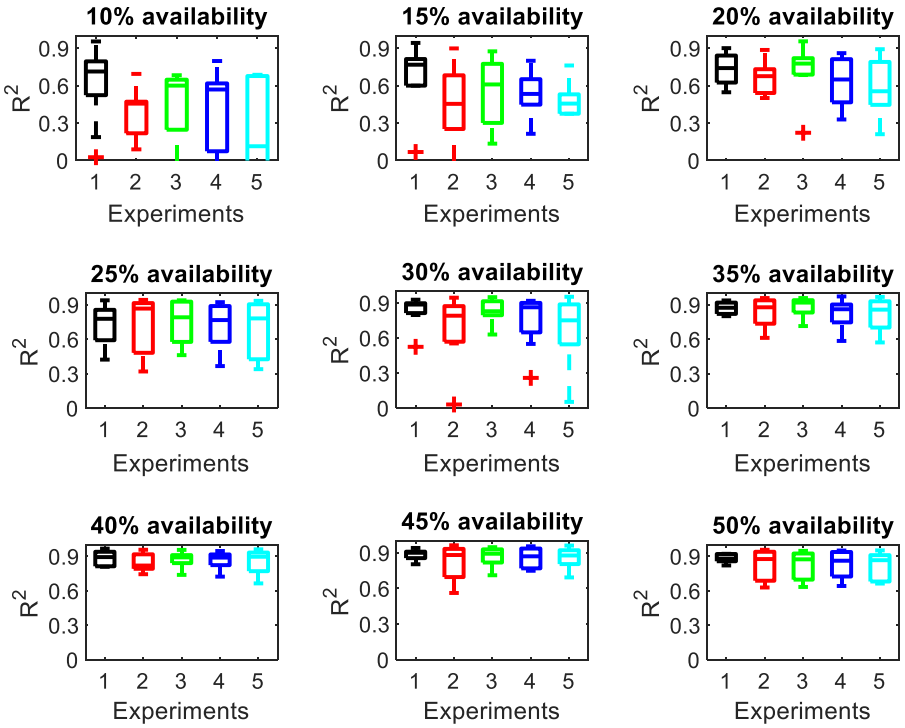


Figure 6.12 Boxplots for rectangular columns over 10 random trials based on four different transfer situations and one baseline in terms of R^2 and the 1, 2, 3, 4, and 5 in the *x*-axis represent *Experiment 1, Experiment 2, Experiment 3, Experiment 4, and Experiment 5* (i.e., *Baseline*).

For the circular columns, the target domain dataset is comprised of the 62 non-ductile circular columns with shear strength as the response variable. *Experiment 1* corresponds to the scenario where the source domain training dataset consists of the 98 circular ductile columns with flexural strength as the response variable. *Experiment 2* corresponds to the scenario where the source domain training dataset consists of the 98 circular ductile columns with drift capacity as the response variable. *Experiment 3* corresponds to the scenario where the source domain training dataset consists of the 208 rectangular ductile columns with flexural strength as the response variable. *Experiment 4* corresponds to the scenario where the source domain training dataset consists of the 208 rectangular ductile columns with drift capacity as the response variable.

Experiment 5 also corresponds to the baseline, as introduced in the validation of RC rectangular column dataset. The same validation procedure described in the rectangular column's validation is also utilized here. The four transfer strategies are also the same. The results for these four transfer strategies and baseline in each case of target domain training data availability are shown in Figure 6.13, where both R^2 and RMSE are taken as the averages of the R^2 and RMSE over the 10 random trials.

From Figure 6.13, it is observed that when the availability of target domain training data is 10%, the R^2 for the baseline is negative, which means the well-trained LS-SVMR model for the baseline has a significantly large bias and thus breaks down. In this case, the proposed DW-SVTR can still improve the performance of the baseline. Additionally, when the availability of target domain training data is 15%, both R^2 and RMSE suggest that the proposed DW-SVTR approach significantly improves the prediction performance of the baseline. The RMSE is decreased from 143.38 kN (i.e., baseline) to 110.48 kN (i.e., *Experiment 1*), 129.07 kN (i.e., *Experiment 2*), 128.57 kN (i.e., *Experiment 3*), and 135.66 kN (i.e., *Experiment 4*), which is equivalent to a reduction of roughly 23%, 10%, 10%, and 5%, respectively. The R^2 is increased from 0.16 (i.e., *baseline*) to 0.49 (i.e., *Experiment 1*), 0.33 (i.e., *Experiment 2*), 0.31 (i.e., *Experiment 3*), and 0.19 (i.e., *Experiment 4*), enhancing the values by roughly 206%, 106%, 94%, and 19%, respectively. With the increase in size of the target domain training data, a similar tendency reflected in the rectangular columns is also exhibited by the circular columns. According to the different transfer strategies, the improved performance by the proposed DW-SVTR also varies. The most significant improvement for both RMSE and R^2 is again *Experiment 1*, followed by *Experiment 3*. *Experiment 2* is slightly better than *Experiment 4*, but both are outperformed by *Experiment 3*. This investigation agrees well with that for the rectangular columns. The performance variability over

the 10 random trials for all five experiments is also reported in boxplots in Figures 6.14 and 6.15. By observation of Figures 6.14 and 6.15, it is evident that, compared to the baseline, the proposed DW-SVTR statistically improves the performance in terms of the median of the 10 random trials for all four transfer strategies, as observed in the rectangular column validation.

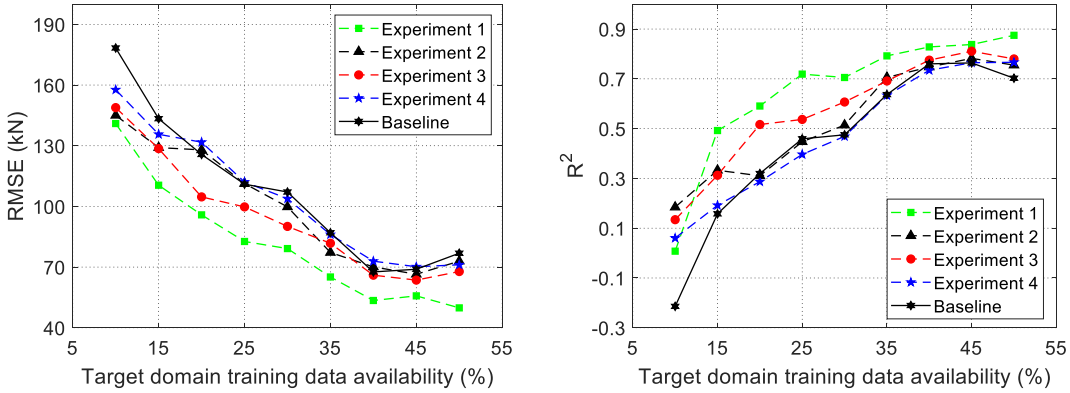


Figure 6.13 Performance versus size of target domain training data availability curve in terms of (a) mean RMSE and (b) mean R^2 for circular columns over the 10 random trials.

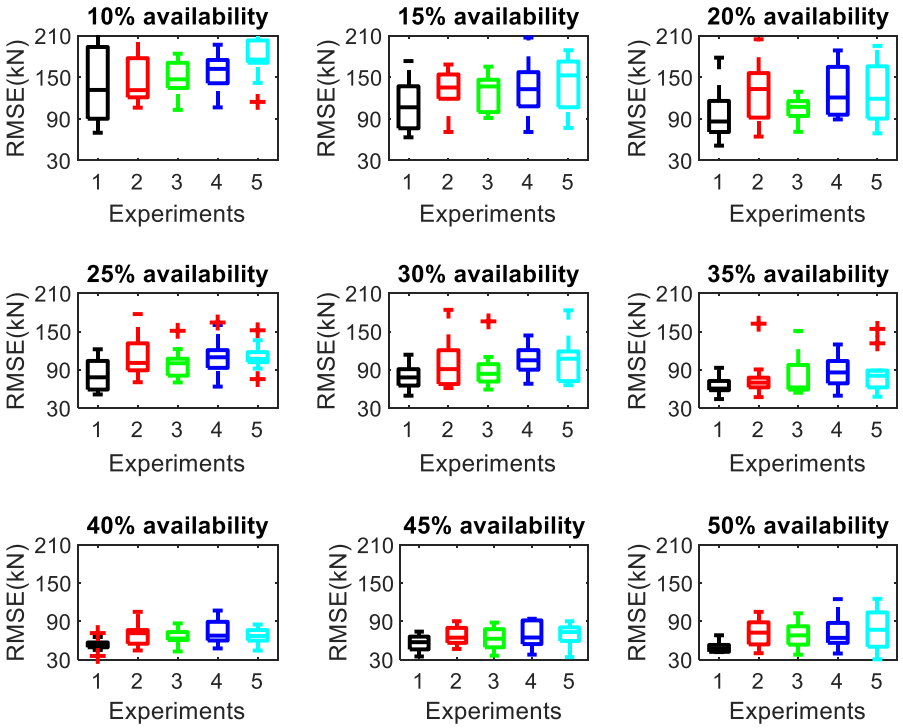


Figure 6.14 Boxplots for circular columns over 10 random trials based on four different transfer situations and one baseline in terms of RMSE and the 1, 2, 3, 4, and 5 in the *x*-axis represent the Experiment 1, Experiment 2, Experiment 3, Experiment 4, and Experiment 5 (i.e., Baseline).

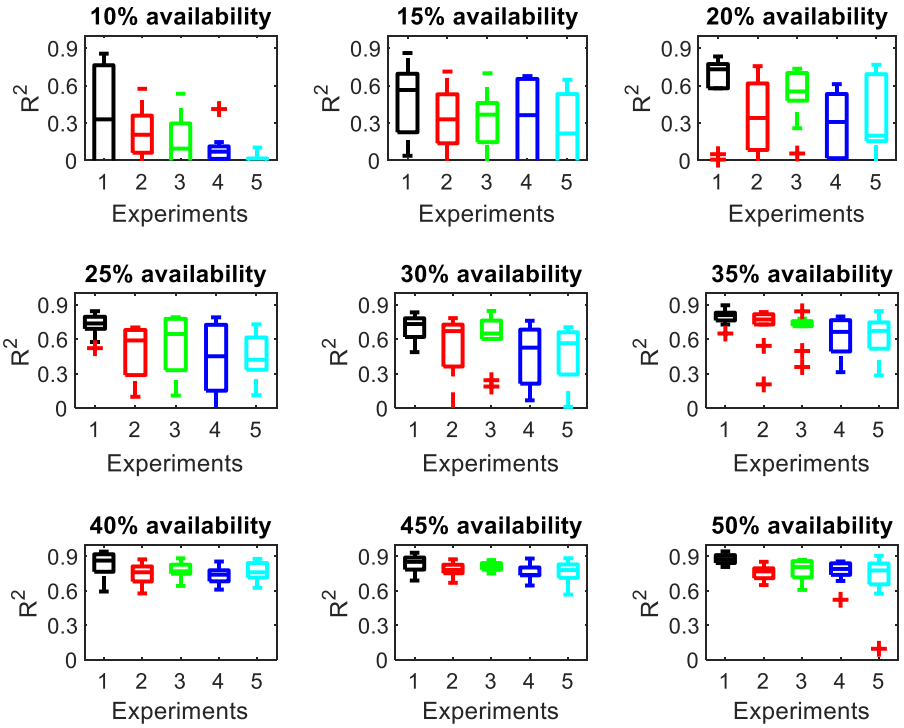


Figure 6.15 Boxplots for circular columns over 10 random trials based on four different transfer situations and one baseline in terms of R^2 and the 1, 2, 3, 4, and 5 in the *x*-axis represent the Experiment 1, Experiment 2, Experiment 3, Experiment 4, and Experiment 5 (i.e., Baseline).

The results obtained by both the simulated and multi-dimensional real-world datasets presented above suggest that the proposed DW-SVTR approach can reduce the sample bias induced by a small dataset and then improve the prediction performance. Further, the proposed DW-SVTR model is also validated effectively by the results for the most challenging case: two unrelated domains where both their marginal and posterior distributions are different (i.e., $p^S(\boldsymbol{x}) \neq p^T(\boldsymbol{x})$ and $p^S(y|\boldsymbol{x}) \neq p^T(y|\boldsymbol{x})$).

The simulated example gives a clear and direct explanation to illustrate how the proposed approach reduces the sample bias and improves the prediction performance for two unrelated domains. The real-world example explicitly investigates the performance of the proposed approach

in terms of target domain training data availability and different transfer strategies. For the relation between performance variability and target domain training data availability over 10 random trials, as shown in Figures 6.11, 6.12, 6.14 and 6.15 (i.e., boxplots), it is observed that the variability occurs when the size of the target domain training data is small (e.g., 10% availability). Further, with the increase of the target domain training data availability, the performance variability decreases in general, though there are several results not following this trend. This is because, when the target domain training data is small (e.g., 10% availability), different random seeds (i.e., 10 random trials) produce the target domain training data that has different levels of sample bias for the corresponding test data. This causes the variation of performance improvement, leading to apparent performance variability. When the size of target domain training data increases, the difference among these levels of sample bias is decreased, producing the relatively lower performance variability. For the relation between performance variability and different transfer strategies, all of the numerical results suggested that *Experiment 1* produces the best performance improvement. This could be explained by the fact that, compared to other transfer strategies, *Experiment 1* is associated with the source domain that is related to the target domain, which makes that the source domain can provide more useful information for the proposed DW-SVTR model to reduce the sample bias and improve the prediction performance. Notably, even for two irrelevant domains, the proposed approach is still able to seek limited useful information from the source domain to reduce the sample bias and enhance the prediction performance.

6.5 Summary

This chapter has presented the development and validation of novel computational methods to address three popular and challenging data-related problems. Specifically, a novel machine learning (ML) approach is proposed that is robust to input data corrupted by outliers. The proposed model is a modification of LWLS-SVMR to overcome its noted drawback regarding lack of robustness to outliers close to query points. The formulation and implementation of the proposed method is introduced in detail. Furthermore, this method is a robust, local model, where prediction of a query point only requires fitting of a subset (not the entire training set) where the data points are relevant with the query point. In comparison to other robust, global approaches, this characteristic enables avoidance of a potential negative influence from irrelevant points and achieves a suitable trade-off between capacity of the learning system and number of training data. Four simulated datasets and eight multi-dimensional real-world datasets are employed to verify that the proposed approach is able to significantly reduce the negative effects of outliers. The proposed RLWLS-SVMR exhibits robustness to outliers and performs best in comparison to all other approaches.

Subsequently, a novel multiple imputation (MI) method called sequential regression-based predictive mean matching (SRB-PMM) is proposed to address missing data problems. The SRB-PMM method imputes the missing values for the partially observed explanatory variables sequentially, starting from the variable with the fewest number of missing values to that with the greatest number of missing values. The use of PMM ensures the imputed values are always inside the observed data range and thus, overcomes the problems associated with meaningless imputations due to the misspecification of the imputation model. Further, a hybrid approach coupling PMM and a K-fold cross-validation algorithm is developed to select the most plausible

imputed dataset. To validate the usefulness of SRB-PMM, two case studies in CE are performed based on the RC column dataset presented in Chapter III. The aim of the first case study is to compare the SRB-PMM method with existing MI methods and to investigate how the missing data ratio affects their performance. For this case study, five data-driven models (i.e., *Delete-LS-SVMR, SRB-PMM-LS-SVMR, JM-LS-SVMR, FCS-LS-SVMR, and Complete-LS-SVMR*) are developed to predict the lateral strength of RC columns. The results reveal that with increasing missing data ratios, the performances of SRB-PMM, joint modeling (JM), and fully conditional specification (FCS) decrease globally. Compared to the baseline (i.e., *Delete-LS-SVMR*), the SRB-PMM method improves prediction performance across all ten missing data ratios, while both JM and FCS occasionally degrade the prediction performance. Additionally, discarding observations with missing values is not always applicable, most relevant, in the case of post-earthquake damage survey data. The second case study aims to illustrate this point by estimating the seismic performance of RC columns, where these columns are missing critical feature information. The results show that the proposed SRB-PMM method can generate realistic, valid candidates for the critical feature information, resulting in reasonable seismic performance estimation.

Finally, a novel regression-based, transfer learning (TL) approach is proposed to reduce the sample bias induced by small datasets. The proposed TL model is termed double-weighted support vector transfer regression (DW-SVTR), as it couples least squares support vector machines for regression (LS-SVMR) with two weight functions. The model formulation and implementation are introduced in detail. Numerical experiments including simulated and multi-dimensional real data are performed to assess and validate the performance of the proposed DW-SVTR model, showing that the proposed approach can transfer the useful information of a large source domain dataset to reduce the sample bias of a small target domain dataset. Further, the results also demonstrated that the proposed approach is valid even for transfer between two irrelevant domains.

CHAPTER VII

CONCLUSIONS

7.1 Summary and Conclusions

Accurate and rapid seismic response prediction of reinforced concrete (RC) structures in earthquake-prone regions is an important topic of research in structural and earthquake engineering. However, existing physics-based modeling approaches do not exhibit good compromise between predictive performance and computational efficiency and overall do not have good generalization performance. To address these problems, this dissertation has proposed a novel data-driven computational paradigm, which can provide a generalized, accurate, robust and efficient way to predict structural seismic response. Further, new computational approaches have been developed to deal with three popular data-related problems: outliers, missing values, and small datasets. To be specific, the contributions in this dissertation are summarized in the following:

- Two RC column datasets, one for rectangular and another for circular columns, were developed for use in this research. Each column specimen in the datasets was tested under cyclic loading reversals. Each data point was composed of the column's features (e.g., geometry, material properties, and design details) that serve as predictors and critical parameters (e.g., backbone bone curve and hysteretic parameters) that serve as response variables. The critical parameters quantify the nonlinear hysteretic properties of the RC column subjected to cyclic loading. To extract the critical parameters from the experimental force-deformation data, a modified three-parameter hysteretic model and a hybrid optimization algorithm were proposed. The proposed hysteretic model

allowed the definition of the softening branch in the monotonic backbone curve. A global metaheuristic algorithm, called simulated annealing (SA), and a downhill simplex method were integrated to optimize three hysteretic parameters, effectively avoiding local minima. A high-quality dataset which is large (relative to the number of features, response variables, and application domain) is essential for the development of accurate and reliable machine learning models and thus, is an important contribution of this dissertation.

- A new machine learning (ML)-based backbone curve model (ML-BCV) was developed for rapid prediction of the bi-linear cyclic backbone curve of RC flexure-, shear-, and flexure-shear-critical columns based on the developed RC rectangular column dataset. The proposed approach integrates a multi-output least squares support vector machine for regression (MLS-SVMR) with a grid search algorithm. The model was tested using cross-validation approaches. Further, the model was compared to the traditional distributed plasticity fiber model in predicting the cyclic backbone curve of three columns (one for flexure-, one for shear-, and one for flexure-shear-critical RC columns). The results showed that the proposed ML-BCV reduced the root-mean-square error (RMSE) for the four values governing the shape of the backbone curve by 80% (drift ratio at yield shear), 61% (yield shear force), 58% (drift ratio at maximum shear), and 67% (maximum shear force), demonstrating that the ML-BCV is increasingly robust and accurate compared to traditional modeling approaches.
- A novel locally weighted ML model (LWLS-SVMR) was developed by combining LS-SVMR and a locally weighted learning algorithm for generalized drift capacity prediction of RC flexure-, shear-, and flexure-shear-critical columns based on the RC

circular column dataset. The proposed LWLS-SVMR was validated by comparison with global LS-SVMR and locally weighted quadratic regression (LWQR) using cross-validation approaches. Finally, the model was also compared with a traditional empirical equation, and the results demonstrated that the proposed LWLS-SVMR is superior to all other approaches and thus, is a promising ML-based technique for enhancing the prediction of drift capacity, universally across RC flexure-, shear-, and flexure-shear-critical columns.

- A new component-level data-driven framework was developed for generalized, accurate, and efficient seismic response history prediction of structural components subjected to both displacement-controlled cyclic loading and dynamic ground motions. The proposed framework is a hybrid ML-physics based approach, where ML was used to directly link the experimental data to nonlinear properties of a target component and a physical model that meets universal laws was utilized to perform the seismic analysis. The framework was illustrated via an RC column. The proposed hysteretic model, LWLS-SVMR, and the RC circular column dataset were used to establish the framework. Two data-driven seismic response solvers were developed to implement the established framework. The two solvers were utilized for seismic response history prediction of RC flexure-, shear-, and flexure-shear-critical columns under cyclic loads as well as a full-scale RC bridge column subjected to six consecutive ground motions. When compared to the experimental data, the results demonstrated that the proposed data-driven framework significantly outperformed the widely used distributed plasticity fiber model in terms of overall accuracy, generalized prediction capabilities, and computational efficiency.

- The component-level data-driven framework was extended to the system-level by coupling it with a shear building model. The proposed system-level framework was illustrated via an RC frame building. The lateral nonlinear force-deformation characteristics of the RC columns in each story were determined by the component-level framework using the column dataset, while the system-level seismic response history was obtained by the shear building model (satisfying equilibrium and compatibility under earthquake loads). Two data-driven seismic response solvers were developed to implement the proposed system-level framework. The two solvers were utilized for seismic response history prediction of a large-scale 3-bay, 3-story RC frame under cyclic loads as well as of two small-scale 3-bay, 9-story RC frames subjected to four and six consecutive ground motions, respectively. Compared to the experimental data, the results demonstrated that the proposed system-level data-driven framework outperformed the widely used distributed plasticity fiber model in terms of accuracy, prediction capability, and computational efficiency and is an extremely promising tool to achieve good compromise between predictive performance and computational efficiency.
- A novel, robust locally weighted ML model (RLWLS-SVMR) was developed by incorporating an extra weight that is a function of residuals into the reformulation of LWLS-SVMR for eliminating the negative effect induced by outliers. An efficient hybrid algorithm was developed to predict query points adaptively using either LWLS-SVMR or iterative RLWLS-SVMR, depending on whether or not outliers surrounded the query points. The proposed RLWLS-SVMR was compared with three other global robust ML models using synthetic datasets corrupted by non-extreme and extreme

outliers and real-world datasets. The results validated that the proposed RLWLS-SVMR performed more robustly against both non-extreme and extreme outliers than other global, robust ML models.

- A new multiple imputation (MI) method (SRB-PMM) was developed by using sequential regression and predictive mean matching to generate several candidates for imputing (filling in) each missing value, while considering uncertainty of missing data. The proposed SRB-PMM method utilized Bayesian parameter estimation to consecutively infer the model parameters for variables with missing values, conditionally based on the fully observed and imputed variables. Given the model parameters, a hybrid approach integrating PMM with a cross-validation algorithm was developed to obtain the most plausible imputed dataset. The proposed SRB-PMM method was compared with two other MI methods using synthetic, incomplete datasets generated using the developed RC column dataset. The results showed that the proposed SRB-PMM method can generate valid and realistic candidates for the missing values and is an effective means to handle missing data problems prominent in the earthquake engineering field.
- A novel regression-based transfer learning (TL) model (DW-SVTR) was developed by coupling two weight functions with LS-SVMR to reduce the negative effect of small sample bias, where TL is defined as knowledge transfer from a large, relevant dataset (source domain) to a small dataset (target domain). The first weight function used kernel mean matching (KMM) to reweight the source domain data such that the means of the source and target domain data in a reproduced kernel Hilbert space (RKHS) are close. In this way, the source domain data points relevant to the target domain points

have a larger weight than irrelevant source domain points. The second weight is a function of estimated residuals, which aims to further reduce the negative interference of irrelevant source domain points. The proposed DW-SVTR was tested using synthetic datasets and the RC column datasets. The results disclosed that the proposed DW-SVTR can reduce small sample bias and improve prediction performance, even between two irrelevant domains.

7.2 Limitations and Recommendations for Future Work

As machine learning (ML) is a rapidly bourgeoning field and new applications are being developed frequently, the present study is of course subjected to several limitations. First, to achieve the goal of computational efficiency, the present study has proposed a polygonal hysteretic model to establish the modeler in **Chapter III**. This means that the hysteretic curve predicted by the proposed method is the piecewise line not the smooth curve. Though it can reasonably describe the nonlinear behavior of RC columns and frames, it may not be able to represent the hysteretic characteristics of structures that have smoother hysteretic curves. Second, like all ML methods, the novel ML models proposed in **Chapters IV and VI** can accurately predict the response within the input ranges of the training set. Outside of these ranges, it cannot necessarily reliably be used for prediction. In this case, the predicted results must be carefully checked with physical knowledge or experts. Third, the system-level data-driven framework proposed in **Chapter V** has a limitation associated with the shear-building model employed. This limitation restricts the application of the proposed method to RC frames where the beams are stiffer than the columns. When more component-level physical experimental data (e.g., beams) are available, a high-fidelity system-level model can be developed to eliminate the limitation of the shear building model. Though, this will also translate to an increase in the computational time for the high-fidelity model. Lastly, since both component- and system-level data-driven frameworks are developed for the seismic response prediction in the context of two-dimensional structural components and systems, these two frameworks cannot be applied for the three-dimensional problems.

There are several potential domains where the methodology described in this dissertation can be extended. A few of these are the following:

1) The physical relation between structural features (i.e., material strength, component geometry, reinforcement details) and strength or deformation capacity can be investigated and/or could be derived from the experimental dataset. This requires the development of novel hybrid physics-ML-based approaches, where the physics can inform ML how to select meaningful feature interactions that satisfy the physical laws. Therefore, new approaches could potentially identify an optimal set of predictors that correlate with the strength or deformation capacity well. The potential predictors could be formed based on the physical constraints imposed to the formulation of ML methods. In this way, an explainable parametric equation that has generalization performance on par with ML models could be derived.
2) Solutions for engineering and science problems using machine learning are controlled by data without accounting for any human interpretation. But those using physics-based methods are usually subjected to theoretical assumptions. Therefore, ML-based methods are free from human assumptions but need data for computation, while physics-based methods are data-free in computation but involve human assumptions. By inserting the physical constraints into the formulation of ML approaches and utilizing both advantages, it is possible to formulate a novel computational paradigm which would be both data-free and assumption-free.
3) The proposed methodology could be extended to the development of a novel data-driven computing paradigm for flexibility-based beam-column element formulation. Given the material constitutive models, the state determination process of the flexibility-based beam-column element satisfies the equilibrium and compatibility conditions. The equilibrium equation is derived from Newton's laws of motion and the

compatibility relation is the kinematic constraint. Thus, they do not suffer any empiricism or uncertainty. In contrast, the traditional material constitutive models are formulated by a physical model, which is empirical and uncertain. Therefore, the ML techniques can be coupled with flexibility-based beam-column element formulations. In this way, ML is used to model the material constitutive relation based on the material datasets (e.g., concrete, rebar), while the equilibrium and compatibility are still enforced along the element. In this case, the empiricism, error, and uncertainty embedded in the traditional material constitutive models can be minimized without risking the loss of material constitutive information.

4) The results presented in this dissertation have also verified that the proposed methodology achieves a good compromise between predictive performance and computational efficiency. This characteristic demonstrates that the proposed approach is a promising computational tool in quantifying regional seismic risk and for other near-real-time scenarios. Although this dissertation utilized RC frames as illustrative examples to illustrate the performance of the proposed system-level data-driven framework, it is a generalized approach and can be applied to any structural system of interest where appropriate data is available. The application of the proposed framework to other structural systems is straightforward and the application procedure is that outlined in this dissertation, especially in **Chapter V**. For different structural systems, the formulation of the MDOF model may vary from that proposed in **Chapter V**, especially for structural systems where the shear building model is not appropriate.

REFERENCES

ACI (American Concrete Institute). (2002). Building code requirements for structural concrete. *ACI Committee 318. Farmington Hills, MI: ACI.*

ACI (American Concrete Institute). (2014). Building code requirements for structure concrete. *ACI Committee 318. Farmington Hills, MI: ACI.*

Ahmad, M. W., Mourshed, M., & Rezgui, Y. (2017). Trees vs Neurons: comparison between random forest and ANN for high-resolution prediction of building energy consumption. *Energy and Buildings*, 147, 77-89.

Alipour, M., Harris, D. K., Barnes, L. E., Ozbulut, O. E., & Carroll, J. (2017). Load-capacity rating of bridge populations through machine learning: Application of decision trees and random forests. *Journal of Bridge Engineering*, 22(10), 04017076.

Amini, M. A., & Poursha, M. (2018). Adaptive force-based multimode pushover analysis for seismic evaluation of midrise buildings. *Journal of Structural Engineering*, 144(8), 04018093.

Amitsu, S., Shirai, N., Adachi, H., & Ono, A. (1991). Deformation of reinforced concrete column with high or fluctuating axial force. *Transactions of the Japan Concrete Institute*, 13, 355-362.

Antoniou, S., & Pinho, R. (2004). Advantages and limitations of adaptive and non-adaptive force-based pushover procedures. *Journal of Earthquake Engineering*, 8(04), 497-522.

ASCE-ACI Joint Task Committee 426. (1973). Shear strength of reinforced concrete members. *Journal of Structural Engineering*, 99(6), 1091-1187.

Atkeson, C. G., Moore, A. W., & Schaal, S. (1997a). Locally weighted learning. *Artificial Intelligence Review* 11: 11-73.

Atkeson, C. G., Moore, A. W., & Schaal, S. (1997b). Locally weighted learning for control. *Artificial Intelligence Review*, 11(1-5), 75-113.

Batista, G. E., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. *Applied Artificial Intelligence*, 17(5-6), 519-533.

Bergstra, J., & Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(1), 281-305.

Berry, M., Parrish, M., & Eberhard, M. (2004). PEER structural performance database, User's Manual (Version 1.0). University of California, Berkeley.

Bitaraf, M., Hurlebaus, S., & Barroso, L. R. (2012). Active and semi‐active adaptive control for undamaged and damaged building structures under seismic load. *Computer‐Aided Civil and Infrastructure Engineering*, 27(1), 48-64.

Bottou, L., & Vapnik, V. (1992). Local learning algorithms. *Neural Computation*, 4(6), 888-900.

Bracci, J. M., Kunnath, S. K., & Reinhorn, A. M. (1997). Seismic performance and retrofit evaluation of reinforced concrete structures. *Journal of Structural Engineering*, 123(1), 3-10.

Bracci, J. M., Reinhorn, A. M., & Mander, J. B. (1995). Seismic resistance of reinforced concrete frame structures designed for gravity loads: performance of structural system. *Structural Journal*, 92(5), 597-609.

Bracci, J. M., Reinhorn, A. M., & Mander, J. B. (1992). Seismic resistance of reinforced concrete frame structures designed only for gravity loads: part I-design and properties of a one-third scale model structure. Technical Rep. No. NCEER-92, 27.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Buuren, S. V., & Groothuis-Oudshoorn, K. (2010). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, 1-68.

Campbell, N. A., & Mahon, R. J. (1974). A multivariate study of variation in two species of rock crab of the genus Leptograpsus. *Australian Journal of Zoology*, 22(3), 417-425.

Cavadas, F., Smith, I. F., & Figueiras, J. (2013). Damage detection using data-driven methods applied to moving-load responses. *Mechanical Systems and Signal Processing*, 39(1-2), 409-425.

Cecen, H. (1979). Response of ten-story reinforced concrete frames to simulated earthquakes (Doctoral Dissertation, Graduate College, University of Illinois, Urbana).

Cheok, G. S., & Stone, W. C. (1986). Behavior of 1/6-scale model bridge columns subjected to cycle inelastic loading, NBSIR 86-3494. Center for Building Technology, National Engineering Laboratory, National Institute of Standards and Technology, Gaithersburg, Maryland, 20899.

Chopra, A. K. (2007). *Dynamics of structures: theory and applications to earthquake engineering*. Prentice Hall.

Chopra, A. K., & Goel, R. K. (2002). A modal pushover analysis procedure for estimating seismic demands for buildings. *Earthquake Engineering & Structural Dynamics*, 31(3), 561-582.

Cheng, M. Y., & Cao, M. T. (2015). Hybrid intelligent inference model for enhancing prediction accuracy of scour depth around bridge piers. *Structure and Infrastructure Engineering*, 11(9), 1178-1189.

Chou, J. S., Ngo, N. T., & Pham, A. D. (2015). Shear strength prediction in reinforced concrete deep beams using nature-inspired metaheuristic support vector regression. *Journal of Computing in Civil Engineering*, 30(1), 04015002.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, 74(368), 829-836.

Cleveland, W. S., & Devlin, S. J. (1988). Locally weighted regression: an approach to regression analysis by local fitting. *Journal of the American Statistical Association*, 83(403), 596-610.

Cortes, C., & Mohri, M. (2014). Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519, 103-126.

De Brabanter, K., Pelckmans, K., De Brabanter, J., Debruyne, M., Suykens, J. A., Hubert, M., & De Moor, B. (2009). Robustness of kernel based regression: a comparison of iterative weighting schemes. In *International Conference on Artificial Neural Networks* (pp. 100-110). Springer, Berlin, Heidelberg.

De Brabanter, J., Pelckmans, K., Suykens, J. A., & Vandewalle, J. (2002). Robust cross-validation score function for non-linear function estimation. In *International Conference on Artificial Neural Networks* (pp. 713-719). Springer, Berlin, Heidelberg.

Decanini, L., Mollaioli, F., Mura, A., & Saragoni, R. (2004, August). Seismic performance of masonry infilled R/C frames. In *13th World Conference on Earthquake Engineering* (No. 165).

Deierlein, G. G., Reinhorn, A. M., & Willford, M. R. (2010). Nonlinear structural analysis for seismic design. *NEHRP Seismic Design Technical Brief*, 4, 1-36.

Elwood, K. J., & Moehle, J. P. (2005). Drift capacity of reinforced concrete columns with light transverse reinforcement. *Earthquake Spectra*, 21(1), 71-89.

Eom, T., Kang, S., Park, H., Choi, T., and Jin, J. (2014). Cyclic loading test for reinforced concrete columns with continuous rectangular and polygonal hoops. *Engineering Structures*, 67, 39-49.

Fajfar, P., & Gašperšič, P. (1996). The N2 method for the seismic damage analysis of RC buildings. *Earthquake Engineering & Structural Dynamics*, 25(1), 31-46.

Farrokh, M., Dizaji, M. S., & Joghataie, A. (2015). Modeling hysteretic deteriorating behavior using generalized Prandtl neural network. *Journal of Engineering Mechanics*, 141(8), 04015024.

Farrokh, M., & Joghataie, A. (2013). Adaptive modeling of highly nonlinear hysteresis using Preisach neural networks. *Journal of Engineering Mechanics*, 140(4), 06014002.

FEMA, F. (1997). *NEHRP guidelines for the seismic rehabilitation of buildings*. FEMA 273.

Feng, K., & Chaspari, T. (2019, October). Low-resource language identification from speech using transfer learning. In *2019 IEEE 29th International Workshop on Machine Learning for Signal Processing (*MLSP) (pp. 1-6). IEEE.

Filippou, F. C., & Issa, A. (1988). Nonlinear analysis of reinforced concrete frames under cyclic load reversals. Report No. UCB/EERC 88-12. University of California, Berkeley: Earthquake Engineering Research Center.

Friedman, J.H., (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, 1-67.

Gao, Y., & Mosalam, K. M. (2018). Deep transfer learning for image‐based structural damage recognition. *Computer‐Aided Civil and Infrastructure Engineering*, 33(9), 748-768.

Garcke, J., & Vanck, T. (2014, September). Importance weighted inductive transfer learning for regression. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 466-481). Springer, Berlin, Heidelberg.

Ghannoum, W. M., & Moehle, J. P. (2011). Rotation-based shear failure model for lightly confined RC columns. *Journal of Structural Engineering*, 138(10), 1267-1278.

Ghee, A. B., Priestley, M. N., & Paulay, T. (1989). Seismic shear strength of circular reinforced concrete columns. *Structural Journal*, 86(1), 45-59.

Gilberson, M. (1967). The response of non-linear multi-storey structures subjected to earthquake excitations. *EERL Report*, *Earthquake Engineering Research Laboratory*.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., & Schölkopf, B. (2009). Covariate shift by kernel mean matching. *Dataset Shift in Machine Learning*, 3(4), 5.

Grossi, P., & Kunreuther, H. (2005). *Catastrophe Modeling: A New Approach to Managing Risk (Vol. 25)*. Springer Science & Business Media.

Guarize, R., Matos, N. A. F., Sagrilo, L. V. S., & Lima, E. C. P. (2007). Neural networks in the dynamic response analysis of slender marine structures. *Applied Ocean Research*, 29(4), 191-198.

Guler, H. (2014). Prediction of railway track geometry deterioration using artificial neural networks: a case study for Turkish state railways. *Structure and Infrastructure Engineering*, 10(5), 614-626.

Gupta, B., & Kunnath, S. K. (2000). Adaptive spectra-based pushover procedure for seismic evaluation of structures. *Earthquake Spectra*, 16(2), 367-391.

Hajirasouliha, I., & Doostan, A. (2010). A simplified model for seismic response prediction of concentrically braced frames. *Advances in Engineering Software*, 41(3), 497-505.

Hamilton, C. H., Pardoen, G. C., & Kazanjy, R. P. (2002). Experimental testing of bridge columns subjected to reversed-cyclic and pulse-type loading histories. Report 2001-03. Civil Engineering Technical Report Series, University of California, Irvine.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. A. (2011). *Robust statistics: the approach based on influence functions*. John Wiley & Sons.

Hand, D. J., & Vinciotti, V. (2003). Local versus global models for classification problems: Fitting models where it matters. *The American Statistician*, 57(2), 124-131.

Harrison Jr, D., & Rubinfeld, D. L. (1978). Hedonic housing prices and the demand for clean air. *Journal of Environmental Economics and Management*, 5(1), 81-102.

Haselton, C. B., Liel, A. B., & Deierlein, G. G. (2009). Simulating structural collapse due to earthquakes: model idealization, model calibration, and numerical solution algorithms. *Computational Methods in Structural Dynamics and Earthquake Engineering* (COMPDYN).

Hoff, P. D. (2009). *A first course in Bayesian statistical methods* (Vol. 580). New York: Springer.

Huang, J., Gretton, A., Borgwardt, K., Schölkopf, B., & Smola, A. J. (2007). Correcting sample selection bias by unlabeled data. In *Advances in Neural Information Processing Systems*, 601-608.

Huber, P. J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 73-101.

Ibarra, L. F., Medina, R. A., & Krawinkler, H. (2005). Hysteretic models that incorporate strength and stiffness deterioration. *Earthquake Engineering & Structural Dynamics*, 34(12), 1489-1511.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning*. New York: Springer.

Jeng, C. H., & Mo, Y. L. (2004). Quick seismic response estimation of prestressed concrete bridges using artificial neural networks. *Journal of Computing in Civil Engineering*, 18(4), 360-372.

Jeon, J. S., Shafieezadeh, A., & DesRoches, R. (2014). Statistical models for shear strength of RC beam‐column joints using machine‐learning techniques. *Earthquake Engineering & Structural Dynamics*, 43(14), 2075-2095.

Kang, P. (2013). Locally linear reconstruction based missing value imputation for supervised learning. *Neurocomputing*, 118, 65-78.

Krawinkler, H., & Seneviratna, G. D. P. K. (1998). Pros and cons of a pushover analysis of seismic performance evaluation. *Engineering Structures*, 20(4-6), 452-464.

Kvålseth, T. O. (1985). Cautionary note about R2. *The American Statistician*, 39(4), 279-285.

Lagaros, N. D., & Papadrakakis, M. (2012). Neural network based prediction schemes of the non-linear seismic response of 3D buildings. *Advances in Engineering Software*, 44(1), 92-115.

Legeron, F., & Paultre, P. (2000). Behavior of high-strength concrete columns under cyclic flexure and constant axial load. *Structural Journal*, 97(4), 591-601.

Li, Q., & Chaspari, T. (2019, October). Exploring transfer learning between scripted and spontaneous speech for emotion recognition. In *2019 International Conference on Multimodal Interaction* (pp. 435-439).

Little, R. J. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics*, 6(3), 287-296.

Little, R. J, & Rubin, D. B. (1987). *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Liu, C., & Rubin, D. B. (1998). Ellipsoidally symmetric extensions of the general location model for mixed categorical and continuous data. *Biometrika*, 85(3), 673-688.

Luo, H., & Paal, S. G. (2018). Machine learning–based backbone curve model of reinforced concrete columns subjected to cyclic loading reversals. *Journal of Computing in Civil Engineering*, 32(5), 04018042.

Luo, H., & Paal, S. G. (2019). A locally weighted machine learning model for generalized prediction of drift capacity in seismic vulnerability assessments. *Computer‐Aided Civil and Infrastructure Engineering*, 34(11), 935-950.

Luo, H., & Paal, S. G. (2020). Reducing the effect of sample bias for small datasets with double-weighted support vector transfer regression, *Computer-Aided Civil and Infrastructure Engineering*, Wiley. DOI: 10.1111/mice.12617.

Lynn, A. C., Moehle, J. P., Mahin, S. A., & Holmes, W. T. (1996). Seismic evaluation of existing reinforced concrete building columns. *Earthquake Spectra*, 12(4), 715-739.

Mander, J. B., Priestley, M. J., & Park, R. (1988). Theoretical stress-strain model for confined concrete. *Journal of Structural Engineering*, 114(8), 1804-1826.

Marini, A., & Spacone, E. (2006). Analysis of reinforced concrete elements including shear effects. *ACI Structural Journal*, 103(5), 645.

Mazzoni, S., McKenna, F., Scott, M. H., & Fenves, G. L. (2006). OpenSees command language manual. *Pacific Earthquake Engineering Research (PEER) Center*, 264.

McDaniel, C. C. (1997). *Scale effects on the shear strength of circular reinforced concrete columns*. University of California, San Diego.

Menegotto, M. (1973). Method of analysis for cyclically loaded RC plane frames including changes in geometry and non-elastic behavior of elements under combined normal force and bending. In *Proc. of IABSE Symposium on Resistance and Ultimate Deformability of Structures Acted on by Well Defined Repeated Loads* (pp. 15-22).

Menzies, T., Butcher, A., Marcus, A., Zimmermann, T., & Cok, D. (2011). Local vs. global models for effort estimation and defect prediction. In *2011 26th IEEE/ACM International Conference on Automated Software Engineering* (ASE 2011) (pp. 343-351). IEEE.

Miranda, E., & Reyes, C. J. (2002). Approximate lateral drift demands in multistory buildings with nonuniform stiffness. *Journal of Structural Engineering*, 128(7), 840-849.

Mo, Y. L., & Wang, S. J. (2000). Seismic behavior of RC columns with various tie configurations. *Journal of Structural Engineering*, 126(10), 1122-1130.

Moehle, J. (2014). *Seismic design of reinforced concrete buildings*. McGraw Hill Professional.

Moehle, J., & Deierlein, G. G. (2004). A framework methodology for performance-based earthquake engineering. *In 13th World Conference on Earthquake Engineering (Vol. 679)*.

Morris, T. P., White, I. R., & Royston, P. (2014). Tuning multiple imputation by predictive mean matching and local residual draws. *BMC Medical Research Methodology*, 14(1), 75.

Mu, H. Q., & Yuen, K. V. (2015). Novel outlier-resistant extended Kalman filter for robust online structural identification. *Journal of Engineering Mechanics*, 141(1), 04014100.

Nelder, J. A., & Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4), 308-313.

Nelson, W. (1981). Analysis of performance-degradation data from accelerated tests. *IEEE Transactions on Reliability*, 30(2), 149-155.

Ou, Y. C., & Kurniawan, D. P. (2015). Effect of axial compression on shear behavior of high-strength reinforced concrete columns. *ACI Structural Journal*, 112(2), 209-219.

Ozbakkaloglu, T., & Saatcioglu, M. (2004). Rectangular stress block for high-strength concrete. *ACI Structural Journal*, 101(4), 475-483.

Pal, M., & Deswal, S. (2011). Support vector regression based shear strength modelling of deep beams. *Computers & Structures*, 89(13-14), 1430-1439.

Pan, S. J., & Yang, Q. (2009). A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10), 1345-1359.

Pardoe, D., & Stone, P. (2010). Boosting for regression transfer. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*. Omni press.

Park, Y. J., Reinhorn, A. M., and Kunnath, S. K. (1987). IDARC: Inelastic damage analysis of reinforced concrete frame—shear-wall structures. Tech. Rep. NCEER-87-0008, State University of New York at Buffalo, Buffalo, N.Y

Penrose, K., Nelson, A., & Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques. *Medicine & Science in Sports & Exercise*, 17(2).

Poursha, M., Khoshnoudian, F., & Moghadam, A. S. (2009). A consecutive modal pushover procedure for estimating the seismic demands of tall buildings. *Engineering Structures*, 31(2), 591-599.

Priestley, M. J. N., Potangaroa, R. T., & Park, R. (1981). Ductility of spirally-confined concrete columns. *Journal of the Structural Division*, 107(1), 181-202.

Priestley, M. N., Verma, R., & Xiao, Y. (1994). Seismic shear strength of reinforced concrete columns. *Journal of Structural Engineering*, 120(8), 2310-2329.

Pujol, S., Ramfrez, J. A., & Sozen, M. A. (1999). Drift capacity of reinforced concrete columns subjected to cyclic shear reversals. *Special Publication*, 187, 255-274.

Quinlan, J. R. (1993). Combining instance-based and model-based learning. In *Proceedings of the tenth International Conference on Machine Learning* (pp. 236-243).

Quionero-Candela, J., Sugiyama, M., Schwaighofer, A., & Lawrence, N. D. (2009). *Dataset shift in machine learning*. The MIT Press

Raghunathan, T. E., Lepkowski, J. M., Van Hoewyk, J., & Solenberger, P. (2001). A multivariate technique for multiply imputing missing values using a sequence of regression models. *Survey Methodology*, 27(1), 85-96.

Rousseeuw, P. J. (1984). Least median of squares regression. *Journal of the American Statistical Association*, 79(388), 871-880.

Rousseeuw, P. J., & Hubert, M. (2011). Robust statistics for outlier detection. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(1), 73-79.

Rousseeuw, P. J., & Leroy, A. M. (1987). *Robust regression and outlier detection*. New York: Wiley.

Rousseeuw, P., & Yohai, V. (1984). Robust regression by means of S-estimators. In *Robust and Nonlinear Time Series Analysis* (pp. 256-272). Springer, New York, NY.

Roy, M. H., & Larocque, D. (2012). Robustness of random forests for regression. *Journal of Nonparametric Statistics*, 24(4), 993-1006.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581-592.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics*, 4(1), 87-94.

Rubin, D. B. (2004). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91(434), 473-489.

Rusiecki, A. (2007). Robust LTS backpropagation learning algorithm. In *International Work-Conference on Artificial Neural Networks* (pp. 102-109). Springer, Berlin, Heidelberg.

Saatcioglu, M., & Grira, M. (1999). Confinement of reinforced concrete columns with welded reinforced grids. *Structural Journal*, 96(1), 29-39.

Saiidi, M., & Sozen, M. A. (1981). Simple nonlinear seismic analysis of R/C structures. *Journal of the Structural Division*, 107(5), 937-953.

Sasani, M. (2004, August). Shear strength and deformation capacity models for RC columns. In *Proceedings of 13th World Conference on Earthquake Engineering*, Vancouver, BC, Canada, Paper (No. 1838).

Schafer, J. L. (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.

Schenker, N., & Taylor, J. M. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, 22(4), 425-446.

Schoettler, M. J., Restrepo, J. I., Guerrini, G., Duck, D. E., & Carrea, F. (2012). A full-scale, single-column bridge bent tested by shake-table excitation. Center for Civil Engineering Earthquake Research, Department of Civil Engineering, University of Nevada

Schultz, A. E. (1986). An experimental and analytical study of the earthquake response of R/C frames with yielding columns (Doctoral dissertation, University of Illinois at Urbana-Champaign).

Scott, B. D., Park, R., & Priestley, M. J. N. (1989). Stress-strain behavior of concrete confined by overlapping hoops at low and high strain ratio Rates (Doctoral dissertation, Doctoral Thesis, Lulea University of Technology, Lulea, Sweden).

Sezen, H. (2008). Shear deformation model for reinforced concrete columns. *Structural Engineering and Mechanics*, 28(1), 39-52.

Sezen, H., & Moehle, J. P. (2002). Seismic behavior of shear-critical reinforced concrete building columns. In *Seventh US National Conference on Earthquake Engineering, Earthquake Engineering Research Institute*, Boston, MA.

Sezen, H., & Moehle, J. P. (2004). Shear strength model for lightly reinforced concrete columns. *Journal of Structural Engineering*, 130(11), 1692-1703.

Shu, J., Zhang, Z., Gonzalez, I., & Karoumi, R. (2013). The application of a damage detection method using Artificial Neural Network and train-induced vibrations on a simplified railway bridge model. *Engineering Structures*, 52, 408-421.

Sivaselvan, M. V., & Reinhorn, A. M. (2000). Hysteretic models for deteriorating inelastic structures. *Journal of Engineering Mechanics*, 126(6), 633-640.

Solomatine, D., See, L. M., & Abrahart, R. J. (2009). Data-driven modelling: concepts, approaches and experiences. In *Practical Hydroinformatics* (pp. 17-30). Springer, Berlin, Heidelberg.

Spacone, E., Camata, G., & Faggella, M. (2008). Nonlinear models and nonlinear procedures for seismic analysis of reinforced concrete frame structures. *Comput Struct Dynam Earthquake Eng: Struct Infrastruct Book Ser*, 2, 323.

Spacone, E., Filippou, F. C., & Taucer, F. F. (1996). Fibre beam–column model for non-linear analysis of R/C frames: Part I. Formulation. *Earthquake Engineering & Structural Dynamics*, 25(7), 711-725.

Spacone, E., Filippou, F. C., & Taucer, F. F. (1996). Fibre beam–column model for non-linear analysis of r/c frames: part ii. applications. *Earthquake Engineering & Structural Dynamics*, 25(7), 727-742.

Sritharan, S., Priestley, M. N., & Seible, F. (1996). Seismic response of column/cap beam tee connections with cap beam prestressing (No. SSRP-96/09).

Stone, W. C., & Cheok, G. S. (1989). Inelastic behavior of full-scale bridge columns subjected to cyclic loading, NIST BSS 166. Building Science Series, Center for Building Technology, National Engineering Laboratory, National Institute of Standards and Technology.

Suykens, J. A., De Brabanter, J., Lukas, L., & Vandewalle, J. (2002). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, 48(1-4), 85-105.

Suykens, J. A. K., Lukas, L., Van Dooren, P., De Moor, B., & Vandewalle, J. (1999). Least squares support vector machine classifiers: a large scale algorithm. In *European Conference on Circuit Theory and Design*, ECCTD (Vol. 99, pp. 839-842). Citeseer.

Suykens, J.A., Van Gestel, T. and De Brabanter, J., 2002. *Least squares support vector machines*. World Scientific

Szu, H. and Hartley, R., 1987. Fast simulated annealing. *Physics Letters A*, 122(3-4), pp.157-162.

Tanaka, H.; and Park, R. (1990). Effect of lateral confining reinforcement on the ductile behavior of reinforced concrete columns, Report 90-2, Department of Civil Engineering, University of Canterbury, 458 pages.

Taucer, F., Spacone, E., & Filippou, F. C. (1991). A fiber beam-column element for seismic response analysis of reinforced concrete structures. Berkekey, California: Earthquake Engineering Research Center, College of Engineering, University of California.

Terzic, V., Schoettler, M. J., Restrepo, J. I., & Mahin, S. A. (2015). Concrete column blind prediction contest 2010: outcomes and observations. PEER Report, 1, 1-145.

Torrey, L., & Shavlik, J. (2010). Transfer learning. In *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques* (pp. 242-264). IGI Global.

Van Laarhoven, P. J., & Aarts, E. H. (1987). Simulated annealing. *In Simulated Annealing: Theory and Applications* (pp. 7-15). Springer, Dordrecht.

Vapnik, V. (1992). Principles of risk minimization for learning theory. In *Advances in Neural Information Processing Systems* (pp. 831-838).

Vapnik, V., 1995. *The nature of statistical learning theory*. NewYork:Springer-Verlag.

Vapnik, V., & Bottou, L. (1993). Local algorithms for pattern recognition and dependencies estimation. *Neural Computation*, 5(6), 893-909.

Verderame, G. M., Fabbrocino, G., & Manfredi, G. (2008). Seismic response of rc columns with smooth reinforcement. Part II: Cyclic tests. *Engineering Structures*, 30(9), 2289-2300.

Vu, D. T., & Hoang, N. D. (2016). Punching shear capacity estimation of FRP-reinforced concrete slabs using a hybrid machine learning approach. *Structure and Infrastructure Engineering*, 12(9), 1153-1161.

Vu, N. H. D., Priestly, M. J., Seible, F., & Benzoni, G. (1999). The seismic response of well-confined circular reinforced concrete columns with low aspect ratios (No. SSRP-97/15).

Watanabe, F., & Ichinose, T. (1991). Strength and ductility design of RC members subjected to combined bending and shear. In *Proc. Workshop on Concrete Shear in Earthquake*, University of Houston, Texas (pp. 429-438).

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big data*, 3(1), 9.

Whitney, C. S. (1937, March). Design of reinforced concrete members under flexure or combined flexure and direct compression. In *Journal Proceedings* (Vol. 33, No. 3, pp. 483-498).

Wu, R. T., & Jahanshahi, M. R. (2018). Deep convolutional neural network for structural dynamic response estimation and system identification. *Journal of Engineering Mechanics*, 145(1), 04018125.

Xavier-de-Souza, S., Suykens, J.A., Vandewalle, J. and Bollé, D., 2010. Coupled simulated annealing. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(2), pp.320-335.

Xie, L., Lu, X., Guan, H., & Lu, X. (2015). Experimental study and numerical model calibration for earthquake-induced collapse of RC frames with emphasis on key columns, joints, and the overall structure. *Journal of Earthquake Engineering*, 19(8), 1320-1344.

Xie, Y., Ebad Sichani, M., Padgett, J. E., & DesRoches, R. (2020). The promise of implementing machine learning in earthquake engineering: A state-of-the-art review. *Earthquake Spectra*, 8755293020919419.

Yeh, I. C. (2007). Modeling slump flow of concrete using second-order regressions and artificial neural networks. *Cement and Concrete Composites*, 29(6), 474-480.

Ying, Z. and Keong, K.C., 2004, August. Fast leave-one-out evaluation and improvement on inference for LS-SVMs. In *Proceedings of the 17th International Conference on Pattern Recognition*, 2004. ICPR 2004. (Vol. 3, pp. 494-497). IEEE.

Yuen, K. V., & Mu, H. Q. (2012). A novel probabilistic method for robust parametric identification and outlier detection. *Probabilistic Engineering Mechanics*, 30, 48-59.

Yuen, K. V., & Ortiz, G. A. (2017). Outlier detection and robust regression for correlated data. *Computer Methods in Applied Mechanics and Engineering*, 313, 632-646.

Yun, G. J., Ghaboussi, J., & Elnashai, A. S. (2008). A new neural network ‐ based model for hysteretic behavior of materials. *International Journal for Numerical Methods in Engineering*, 73(4), 447-469.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. In: *Bayesian inference and decision techniques, Stud. Bayesian Econometrics Statist.*, vol 6, North-Holland, Amsterdam, pp 233–243

Zhang, R., Chen, Z., Chen, S., Zheng, J., Büyüköztürk, O., & Sun, H. (2019). Deep long short-term memory networks for nonlinear structural seismic response prediction. *Computers & Structure*s, 220, 55-68.

Zhang, Q., Gong, J., & Ma, Y. (2014). Seismic shear strength and deformation of RC columns failed in flexural shear. *Magazine of Concrete Research*, 66(5), 234-248.

APPENDIX A

DATABASE OF RECTANGULAR REINFORCED CONCRETE COLUMNS

No.	Reference	a/d	f_c (MPa)	f_{yl} (MPa)	f_{yt} (MPa)	p_l	p_t	$P/A_g f_c$	V_y (kN)	V_m (kN)	V_u (kN)	δ_y (%)	δ_m (%)	δ_u (%)	α	β	γ
1	Berry et al. (2004)	2.35	23.1	375	297	0.0179	0.0070	0.26	515.1	656.88	613.2	0.51	1.53	2.83	2.80	0.06	0.95
2		2.34	41.4	375	316	0.0179	0.0107	0.21	591.49	764.21	723.02	0.41	1.11	2.14	1.81	0.00	0.96
3		2.35	21.4	375	297	0.0179	0.0075	0.42	510.09	641.56	586.18	0.32	0.81	1.76	2.81	0.03	0.93
4		2.34	23.5	375	294	0.0179	0.0134	0.60	570.16	696.51	696.51	0.30	1.31	1.31	4.04	0.07	0.98
5		4.26	23.6	427	320	0.0151	0.0111	0.38	160.52	192.05	153.64	0.69	1.27	3.10	32.47	0.18	0.99
6		4.24	25	427	280	0.0151	0.0087	0.21	143.89	169.23	146.42	0.65	1.67	3.65	7.28	0.13	1.00
7		4.13	46.5	446	364	0.0151	0.0045	0.10	161.39	199.59	167.12	0.55	1.83	6.12	8.61	0.20	0.98
8		4.13	44	446	360	0.0151	0.0065	0.30	235.05	279.25	226.2	0.54	1.01	2.91	11.34	0.17	0.94
9		4.13	44	446	364	0.0151	0.0041	0.30	221.34	276.98	221.58	0.44	0.91	2.39	4.36	0.34	1.00
10		4.13	40	446	255	0.0151	0.0030	0.30	210.27	264.56	211.65	0.51	1.06	1.99	2.63	0.27	0.98
11		4.13	28.3	440	466	0.0151	0.0068	0.22	169.39	213.3	170.64	0.72	2.09	4.41	7.47	0.18	0.97
12		4.13	40.1	440	466	0.0151	0.0085	0.39	225.82	268.88	245.14	0.60	1.93	3.15	4.83	0.02	0.98
13		4.13	41	474	372	0.0151	0.0062	0.50	243.61	292.02	233.62	0.52	1.16	2.13	9.46	0.13	0.96
14		4.13	40	474	388	0.0151	0.0029	0.50	221.43	295.02	271.22	0.41	1.06	1.55	3.11	0.45	0.98
15		4.13	42	474	308	0.0151	0.0117	0.70	253.81	295.55	236.44	0.32	0.49	0.90	48.40	0.35	0.99
16		4.13	39	474	372	0.0151	0.0065	0.70	249.4	295.37	240.3	0.35	0.65	1.04	54.82	0.32	0.99
17		4.13	40	474	308	0.0151	0.0225	0.70	282.87	309.5	247.6	0.44	0.88	2.35	51.26	0.00	0.98
18		4.44	25.6	474	333	0.0157	0.0107	0.20	146.01	166.83	133.46	0.72	1.27	4.38	39.97	0.07	0.95
19		4.44	25.6	474	333	0.0157	0.0107	0.20	135.88	167.76	134.21	0.63	1.14	3.62	49.67	0.05	0.93
20		4.44	25.6	474	333	0.0157	0.0107	0.20	139.2	175.29	140.23	0.53	1.48	3.49	13.06	0.27	0.94
21		4.44	25.6	474	333	0.0157	0.0107	0.20	138.03	170.42	136.34	0.57	1.16	4.28	14.00	0.05	0.98
22		3.24	32	511	325	0.0125	0.0073	0.10	327.73	385.62	374.98	0.74	1.71	4.50	5.57	0.15	0.99
23		3.24	32	511	325	0.0125	0.0073	0.10	348.09	409.2	327.36	0.75	2.39	6.31	6.75	0.10	0.99

24	3.24	32.1	511	325	0.0125	0.0090	0.30	504.4	588.11	470.49	0.70	1.72	4.86	2.00	0.05	0.99
25	3.24	32.1	511	325	0.0125	0.0090	0.30	485.52	618.67	494.94	0.56	1.44	4.82	1.96	0.14	0.99
26	3.10	26.9	432	305	0.0188	0.0070	0.10	319.35	393.1	355.57	0.65	2.15	5.89	20.56	0.04	0.99
27	1.50	20.6	392.8	323	0.0068	0.0088	0.33	125.09	158.92	149.34	0.66	1.55	3.37	0.30	0.03	0.56
28	4.34	24.8	362	325	0.0142	0.0032	0.03	101.39	115.75	115.75	0.66	5.09	5.09	31.05	0.00	1.00
29	4.34	24.8	362	325	0.0142	0.0032	0.03	94.76	108.7	104.41	0.55	3.33	5.37	24.01	0.00	1.00
30	4.34	24.8	362	325	0.0142	0.0032	0.03	88.79	101.46	81.17	0.53	3.10	3.88	7.03	0.00	1.00
31	2.17	21.1	341	559	0.0222	0.0062	0.80	51.05	63.78	51.02	0.62	1.05	1.69	54.56	0.32	1.00
32	3.49	27.9	374	506	0.0162	0.0038	0.11	69.23	76.38	61.1	0.76	1.73	4.26	7.53	0.02	0.91
33	3.49	27.9	374	506	0.0162	0.0038	0.11	70.85	80.03	64.02	0.67	1.60	4.16	5.20	0.05	0.87
34	3.49	27.9	374	506	0.0162	0.0038	0.11	69.23	76.38	61.1	0.76	1.73	4.26	6.51	0.06	0.92
35	3.49	24.8	374	352	0.0162	0.0038	0.12	69.71	84.58	82.2	0.65	1.75	2.28	1.09	0.00	0.95
36	3.49	27.9	374	506	0.0162	0.0038	0.11	62.42	75.16	68.71	0.54	1.85	2.51	1.06	0.00	0.98
37	3.49	27.9	374	506	0.0162	0.0038	0.11	63.69	74.68	72.81	0.60	2.03	2.86	1.12	0.00	0.99
38	2.62	85.7	399.6	328.4	0.0380	0.0164	0.40	195.41	238.99	227.9	0.59	1.82	2.40	2.02	0.02	0.99
39	2.62	85.7	399.6	792.3	0.0380	0.0164	0.40	200.38	244.13	214.17	0.61	2.01	6.98	2.18	0.07	0.99
40	2.62	85.7	399.6	328.4	0.0380	0.0164	0.63	196.65	242.14	193.71	0.60	1.52	1.91	16.32	0.00	0.98
41	2.62	85.7	399.6	792.3	0.0380	0.0164	0.63	194.35	246.68	213.94	0.57	1.51	4.50	3.59	0.00	0.99
42	2.62	115.8	399.6	328.4	0.0380	0.0164	0.25	198.55	240.89	195.71	0.59	1.70	6.36	2.04	0.01	0.93
43	2.62	115.8	399.6	792.3	0.0380	0.0164	0.25	208.43	245.77	214.12	0.61	1.02	6.38	1.62	0.03	0.92
44	2.62	115.8	399.6	328.4	0.0380	0.0164	0.42	245.56	283.43	226.74	0.65	1.92	4.06	4.07	0.26	0.99
45	2.62	115.8	399.6	792.3	0.0380	0.0164	0.42	250.16	287.97	230.38	0.62	1.74	4.74	1.70	0.06	0.99
46	2.21	99.5	379	774	0.0243	0.0053	0.35	305.19	392.85	360.08	0.44	0.81	2.03	11.08	0.40	0.98
47	2.21	99.5	379	774	0.0243	0.0077	0.35	345.97	406.63	325.3	0.57	2.02	3.81	4.47	0.04	0.95
48	2.21	99.5	379	344	0.0243	0.0065	0.35	357.89	428.33	342.66	0.63	1.01	1.67	53.43	0.93	0.65
49	2.21	99.5	379	1126	0.0243	0.0053	0.35	310.37	390.01	312.01	0.48	0.71	2.83	65.19	0.14	0.98
50	2.21	99.5	379	774	0.0243	0.0053	0.35	368.56	405.59	385.47	0.61	0.83	1.00	56.86	0.14	1.00
51	2.21	99.5	379	857	0.0243	0.0052	0.35	353.67	420.07	375.06	0.56	0.90	1.07	12.52	0.13	0.98
52	2.28	99.5	339	774	0.0181	0.0051	0.35	292.15	363.01	290.41	0.35	0.53	1.05	83.12	0.35	0.91
53	6.14	29.1	367	363	0.0163	0.0061	0.10	52.3	61.72	60.24	0.80	1.02	2.50	4.59	0.01	0.80
54	6.14	30.7	367	363	0.0163	0.0036	0.09	47.42	61.2	58.8	0.76	1.06	3.23	7.34	0.04	0.95

55	6.14	29.2	367	363	0.0163	0.0061	0.10	48.52	57.43	52.46	0.79	1.52	3.05	5.96	0.01	0.97
56	6.14	27.6	429	363	0.0163	0.0036	0.10	42.23	49.08	44.44	0.76	1.12	2.40	2.58	0.00	1.00
57	6.14	29.4	429	392	0.0163	0.0061	0.19	64.98	74.13	61.69	0.97	1.45	3.04	9.71	0.01	0.93
58	6.14	31.8	429	392	0.0163	0.0036	0.18	63.24	74.91	64.61	0.94	1.82	3.08	2.68	0.04	0.89
59	6.14	33.3	363	392	0.0163	0.0061	0.26	68.22	78.86	63.09	0.90	1.21	1.93	8.74	0.05	0.99
60	6.14	32.4	363	392	0.0163	0.0036	0.27	57.77	77.97	62.38	0.72	1.42	2.16	45.27	0.06	0.99
61	6.14	31	363	373	0.0163	0.0061	0.28	66.65	76.96	61.57	0.72	1.02	2.20	24.29	0.05	1.00
62	6.14	31.8	363	373	0.0163	0.0036	0.27	65.44	78.48	65.78	0.91	1.56	2.48	47.36	0.05	0.88
63	3.28	39.3	439	454	0.0194	0.0094	0.21	411.08	466.77	373.42	0.95	1.83	4.39	7.66	0.08	0.99
64	3.30	39.8	439	616	0.0194	0.0051	0.31	436.89	483.1	386.48	0.75	0.97	2.82	2.18	0.10	1.00
65	3.05	43.6	430	470	0.0321	0.0030	0.00	228.4	276.2	216.96	1.73	4.53	6.44	47.76	0.20	0.80
66	3.05	34.8	430	470	0.0321	0.0060	0.14	234.76	267	213.6	1.66	2.71	4.31	78.84	0.32	1.00
67	3.05	32	438	470	0.0321	0.0091	0.15	272.63	325.9	260.72	1.32	6.82	8.58	50.52	0.06	0.98
68	3.09	37.3	437	425	0.0321	0.0085	0.13	282.9	342.8	306.6	1.51	6.43	8.97	52.78	0.01	0.99
69	3.09	39	437	425	0.0321	0.0085	0.13	292.35	341.8	304.7	1.50	6.55	8.79	51.48	0.03	0.99
70	5.18	80	430	430	0.0151	0.0054	0.30	109.62	130.15	104.12	0.78	1.21	1.65	4.55	0.00	0.99
71	5.18	80	430	430	0.0151	0.0054	0.30	101.62	120.55	99.44	0.78	1.29	1.44	2.76	0.04	1.00
72	5.18	80	430	430	0.0151	0.0054	0.20	77.83	95.31	76.24	0.75	1.38	2.14	16.48	0.02	0.97
73	5.18	80	430	430	0.0151	0.0054	0.20	112.75	137.58	110.06	0.69	1.22	1.58	56.27	0.67	0.78
74	5.18	80	430	430	0.0151	0.0081	0.20	119.2	141.24	112.99	0.72	1.30	1.97	60.17	0.36	0.68
75	5.18	80	430	430	0.0151	0.0081	0.30	105.9	125.57	100.45	0.79	1.17	2.15	60.75	0.22	0.79
76	5.18	80	430	430	0.0151	0.0081	0.30	112.29	131.09	104.87	0.83	1.21	1.69	56.00	0.78	1.00
77	5.18	80	430	430	0.0151	0.0081	0.20	92.96	110.31	88.25	0.88	1.27	2.08	49.50	0.23	0.88
78	5.18	80	430	430	0.0151	0.0159	0.20	82.76	101.02	80.82	0.78	1.40	2.39	42.36	0.15	0.95
79	5.18	80	430	430	0.0151	0.0159	0.30	104.91	126.45	101.16	0.77	1.47	2.50	11.32	0.12	0.99
80	5.18	80	430	430	0.0151	0.0159	0.20	106.79	131.62	105.29	0.73	1.38	2.29	8.10	0.17	0.91
81	5.18	80	430	430	0.0151	0.0159	0.30	112.13	134.8	107.84	0.85	1.44	2.50	21.85	0.16	1.00
82	5.18	80	430	430	0.0603	0.0054	0.20	149.98	175.29	144.34	1.29	1.81	4.27	11.83	0.00	1.00
83	5.18	80	430	430	0.0603	0.0054	0.30	141.26	164.68	147.35	0.98	1.43	3.03	6.10	0.00	0.99
84	5.18	80	430	430	0.0603	0.0054	0.30	139.1	165.49	132.6	1.11	2.33	3.74	11.72	0.01	0.99
85	5.18	80	430	430	0.0603	0.0054	0.20	172.37	204.77	163.81	1.30	1.87	3.45	55.80	0.00	0.99

86	5.18	80	430	430	0.0603	0.0081	0.20	137.98	158.04	126.44	1.29	2.24	3.98	24.65	0.00	1.00
87	5.18	80	430	430	0.0603	0.0081	0.20	165.5	195.1	156.08	1.07	2.06	4.11	5.03	0.03	1.00
88	5.18	80	430	430	0.0603	0.0081	0.30	148.65	175.01	140.01	1.08	1.51	3.45	5.75	0.00	0.99
89	5.18	80	430	430	0.0603	0.0081	0.30	144.29	170.52	136.42	1.06	1.58	3.13	5.91	0.02	1.00
90	5.18	80	430	430	0.0603	0.0159	0.20	145.73	171.53	137.22	1.27	2.20	5.83	51.87	0.02	1.00
91	5.18	80	430	430	0.0603	0.0159	0.20	144.18	166.88	133.5	1.24	1.77	4.62	12.26	0.00	0.99
92	5.18	80	430	430	0.0603	0.0159	0.30	143.08	170.4	136.32	1.27	2.62	5.31	50.24	0.03	0.99
93	5.18	80	430	430	0.0603	0.0159	0.30	148.23	172.28	137.82	1.15	2.51	4.35	29.46	0.02	0.99
94	4.01	27.2	448	428	0.0222	0.0017	0.10	307.81	368.43	294.75	1.40	5.16	6.33	42.80	0.02	0.77
95	4.01	27.2	448	428	0.0222	0.0017	0.24	342.59	400.28	320.22	1.12	3.44	4.27	4.35	0.08	1.00
96	3.99	28.1	448	428	0.0222	0.0023	0.09	300.83	379.95	303.96	1.55	5.88	6.68	19.50	0.11	1.00
97	3.99	28.1	448	428	0.0222	0.0023	0.23	378.81	424.53	339.63	1.45	4.62	5.47	4.75	0.04	0.99
98	2.11	76	510	510	0.0355	0.0158	0.10	271.16	324.13	282.05	1.09	6.89	9.39	8.15	0.10	0.99
99	2.11	76	510	510	0.0355	0.0158	0.20	334.05	378.35	310.72	1.06	2.91	8.05	4.27	0.09	1.00
100	2.11	86	510	510	0.0246	0.0158	0.10	236.12	275.46	228.63	0.98	2.64	7.29	6.56	0.07	0.97
101	2.11	86	510	510	0.0246	0.0158	0.19	282.44	318.88	255.1	1.12	3.41	6.88	4.03	0.07	1.00
102	2.10	118	393	1415	0.0186	0.0081	0.60	281.06	334.25	308	0.35	0.68	0.91	42.97	0.48	1.00
103	2.10	118	393	1424	0.0186	0.0127	0.60	319.02	365	287	0.39	0.86	1.81	4.00	0.13	0.98
104	2.10	118	393	1424	0.0186	0.0167	0.60	325.56	391.75	313.4	0.52	1.32	2.47	7.16	0.06	0.99
105	2.10	118	393	1424	0.0186	0.0127	0.35	316.43	362.5	290	0.54	1.32	2.97	1.36	0.24	1.00
106	2.10	118	393	1424	0.0186	0.0167	0.35	302.8	370	296	0.45	1.31	3.46	0.94	0.11	0.98
107	8.40	40.6	407	351	0.0101	0.0010	0.34	44.28	54.54	43.63	0.68	1.36	1.61	3.61	0.37	1.00
108	6.28	72.1	454	463	0.0258	0.0140	0.50	114.32	135.92	110.74	0.24	0.86	1.38	2.03	0.32	0.99
109	6.33	71.7	454	542	0.0258	0.0124	0.36	125.29	148.84	121.07	0.34	0.54	2.40	3.62	0.14	0.98
110	6.33	71.8	454	542	0.0258	0.0124	0.50	118.67	143.59	112.87	0.24	0.36	1.75	5.07	0.08	0.95
111	6.28	71.9	454	463	0.0258	0.0224	0.50	125.67	138.85	113.08	0.41	0.51	2.26	2.65	0.15	1.00
112	6.28	101.8	454	463	0.0258	0.0246	0.45	152.88	181.97	145.57	0.21	0.31	0.72	2.57	0.24	0.97
113	6.28	101.9	454	463	0.0258	0.0294	0.46	154.31	170.12	136.09	0.44	0.53	1.34	2.18	0.07	0.99
114	6.33	102	454	542	0.0258	0.0120	0.45	147.96	159.18	127.34	0.44	0.50	1.06	2.19	0.07	0.99
115	6.28	102.2	454	463	0.0258	0.0187	0.47	156.29	177.97	145.38	0.31	0.40	1.12	4.76	0.36	1.00
116	5.12	34	455.6	570	0.0195	0.0040	0.43	161.49	194.64	155.72	0.74	1.58	2.88	51.33	0.09	0.99

117	5.12	34	455.6	570	0.0195	0.0079	0.43	151.94	186.39	149.12	0.61	1.52	3.78	8.39	0.01	0.99
118	5.12	34	455.6	570	0.0195	0.0079	0.20	132.41	164.29	135.35	0.87	1.92	7.06	8.70	0.03	0.99
119	5.12	34	455.6	570	0.0293	0.0054	0.46	159.2	203.94	163.15	0.57	1.73	3.52	57.26	0.03	0.91
120	5.12	34	455.6	570	0.0293	0.0105	0.46	168.34	204.46	171.57	0.87	1.86	6.09	8.56	0.03	0.94
121	5.12	34	477.8	570	0.0229	0.0105	0.46	180.29	221.46	181.16	0.83	3.01	6.03	8.70	0.03	0.94
122	5.12	34	455.6	580	0.0293	0.0051	0.46	175.65	209.17	182.34	0.87	2.01	6.06	22.89	0.03	0.96
123	5.12	34	455.6	580	0.0293	0.0051	0.23	161.03	198.49	158.79	1.22	1.95	6.65	7.47	0.03	0.96
124	5.12	34	427.8	580	0.0328	0.0051	0.46	178.26	219.33	175.46	0.83	1.95	5.48	5.54	0.02	0.94
125	5.12	34	427.8	570	0.0328	0.0105	0.46	174.08	209.03	167.23	0.83	2.77	6.14	10.69	0.00	1.00
126	3.74	69.6	586.1	406.8	0.0193	0.0091	0.05	55.92	70.28	56.23	1.21	2.07	4.51	5.34	0.05	0.87
127	3.74	69.6	586.1	406.8	0.0193	0.0091	0.05	52.85	68.06	54.45	1.11	1.76	5.14	6.09	0.05	0.79
128	3.44	67.8	572.3	513.7	0.0193	0.0092	0.10	78.73	95.64	76.51	1.24	2.83	6.38	4.68	0.03	0.94
129	3.40	67.8	573.3	514.7	0.0193	0.0091	0.10	77.98	92.97	74.37	1.21	2.79	6.34	4.99	0.04	0.96
130	3.37	65.5	572.3	513.7	0.0193	0.0092	0.21	85.2	107.65	86.12	1.05	2.82	5.00	1.20	0.05	1.00
131	3.24	65.5	573.3	514.7	0.0193	0.0090	0.21	81.59	101.86	81.49	1.04	2.81	4.93	1.39	0.03	0.98
132	3.41	37.9	572.3	513.7	0.0193	0.0091	0.00	50.88	59.16	47.33	1.51	3.06	6.19	10.34	0.05	0.73
133	3.48	37.9	573.3	514.7	0.0193	0.0091	0.00	50.14	58.3	46.64	1.48	2.97	6.13	10.94	0.10	0.81
134	3.70	48.3	586.1	406.8	0.0193	0.0092	0.14	61.94	71.2	56.96	1.18	1.97	3.65	2.10	0.06	0.76
135	3.72	48.3	587.1	407.8	0.0193	0.0091	0.14	54.86	69.4	55.52	1.02	1.93	3.71	2.35	0.06	0.70
136	3.35	38.1	572.3	513.7	0.0193	0.0091	0.36	70.9	84.5	69.4	1.05	2.10	4.21	2.06	0.45	0.99
137	3.35	38.1	573.3	514.7	0.0193	0.0091	0.36	69.87	84.5	69.4	1.06	2.10	4.16	2.41	0.31	0.99
138	3.83	24.9	497	459.5	0.0214	0.0061	0.11	200.81	250	200	1.15	2.42	5.87	5.87	0.02	0.78
139	3.83	26.7	497	459.5	0.0214	0.0061	0.16	247.02	267.58	214.06	1.37	2.56	6.37	7.54	0.01	0.84
140	3.83	26.1	497	459.5	0.0214	0.0061	0.22	242.78	305.3	244.24	1.28	2.47	4.78	3.36	0.01	0.84
141	3.83	25.3	497	459.5	0.0214	0.0061	0.11	219.48	248.05	203.94	1.40	2.62	6.92	17.00	0.01	0.78
142	3.83	27.1	497	459.5	0.0214	0.0061	0.16	245.52	260.5	214.6	1.50	2.46	6.84	14.43	0.02	0.92
143	3.83	26.8	497	459.5	0.0214	0.0061	0.21	274.66	309.81	250.25	1.35	2.74	5.18	3.29	0.02	0.76
144	3.83	26.4	497	459.5	0.0214	0.0057	0.11	209.72	234.62	192.7	1.32	2.37	6.72	5.55	0.00	1.00
145	3.83	27.5	497	459.5	0.0214	0.0057	0.15	245.31	259.52	216.62	1.87	2.66	7.46	49.07	0.02	0.92
146	3.83	26.9	497	459.5	0.0214	0.0057	0.21	267.74	299.56	241.65	1.59	2.70	5.62	3.82	0.01	1.00
147	4.22	102.7	517.1	793	0.0245	0.0062	0.00	36.15	44.3	35.44	1.65	5.08	5.25	57.92	0.10	0.99

148	4.22	86.3	517.1	793	0.0245	0.0062	0.20	59.32	73.16	58.52	0.69	1.56	3.39	29.51	0.05	0.83
149	4.22	87.5	455.1	793	0.0245	0.0071	0.00	25.77	32.16	27.63	1.00	3.05	4.38	51.84	0.16	1.00
150	4.22	83.4	455.1	793	0.0245	0.0071	0.10	43.52	51.27	41.02	0.85	1.76	3.01	59.99	0.04	0.99
151	4.22	90	455.1	793	0.0245	0.0071	0.20	56.03	62.92	50.34	0.71	1.22	2.77	56.04	0.06	0.99
152	4.22	67.5	475.8	1262	0.0245	0.0071	0.00	28.14	37.48	36.53	1.47	6.07	6.10	58.01	0.06	0.99
153	4.22	74.6	475.8	1262	0.0245	0.0071	0.10	37.59	46.98	37.59	0.68	1.70	4.67	50.10	0.02	0.98
154	4.22	81.8	475.8	1262	0.0245	0.0071	0.20	46.68	55.45	44.36	0.61	1.36	2.85	37.07	0.21	1.00
155	4.22	75.8	475.8	1262	0.0245	0.0055	0.20	44.36	56.21	44.97	0.52	1.37	2.54	53.97	0.06	1.00
156	4.22	87	475.8	1262	0.0245	0.0047	0.20	51.63	59.15	47.32	0.74	1.28	2.66	48.88	0.11	1.00
157	4.22	71.2	475.8	1262	0.0245	0.0041	0.20	43.24	53.34	42.67	0.65	1.51	2.87	45.22	0.14	0.98
158	6.99	92.4	451	391	0.0215	0.0186	0.14	99.72	114.85	92.88	1.05	1.62	7.57	1.96	0.03	0.98
159	6.99	93.3	430	391	0.0215	0.0186	0.28	129.68	159.77	127.82	0.90	1.56	4.26	3.06	0.12	0.97
160	6.99	98.2	451	418	0.0215	0.0186	0.39	151.8	162	132.6	1.14	1.37	3.54	3.37	0.19	1.00
161	6.99	94.8	451	391	0.0215	0.0087	0.14	80.06	104.87	85.5	0.80	1.52	4.55	3.33	0.24	1.00
162	6.99	97.7	430	391	0.0215	0.0087	0.26	141.98	167.76	134.21	0.94	1.44	2.52	2.16	0.41	0.99
163	6.99	104.3	451	418	0.0215	0.0087	0.37	166	185	148	0.89	1.24	1.85	2.30	0.63	0.96
164	6.99	78.7	446	438	0.0215	0.0186	0.40	142.4	167.23	136.78	0.81	1.28	8.69	3.24	0.01	1.00
165	6.99	109.2	446	438	0.0215	0.0186	0.41	189.7	212.88	170.3	0.96	2.70	5.67	2.07	0.01	0.78
166	6.99	109.5	446	825	0.0215	0.0206	0.41	174.18	198.22	158.58	1.01	1.32	4.82	2.56	0.01	0.88
167	6.99	104.2	446	825	0.0215	0.0140	0.37	162.93	185.99	148.79	1.04	1.40	3.15	4.01	0.02	0.92
168	6.99	104.5	446	744	0.0215	0.0206	0.53	183.91	201.54	161.23	0.85	1.00	3.32	4.47	0.11	0.97
169	6.99	109.4	446	492	0.0215	0.0186	0.51	191.4	208.39	166.71	0.96	1.13	3.30	2.21	0.07	1.00
170	2.45	33.7	453	410.9	0.0245	0.0027	0.08	98.67	112.95	90.36	0.84	1.82	3.00	50.25	0.00	0.40
171	2.45	33.7	453	410.9	0.0245	0.0027	0.08	97.55	112.95	103.61	0.89	1.99	3.02	21.57	0.00	0.39
172	2.45	32.1	453	410.9	0.0245	0.0052	0.09	100.24	112.39	89.91	0.92	2.94	3.87	43.40	0.00	0.38
173	2.45	32.1	453	410.9	0.0245	0.0052	0.09	97.29	112.39	90.06	0.83	2.88	4.01	55.27	0.00	0.31
174	2.45	29.9	453	410.9	0.0245	0.0027	0.10	96.52	112.3	91.09	0.88	2.43	3.12	45.08	0.00	0.33
175	2.45	29.9	453	410.9	0.0245	0.0027	0.10	93.65	112.3	99.59	0.77	2.40	2.99	7.59	0.00	0.32
176	2.45	27.4	453	410.9	0.0245	0.0037	0.10	101.47	114.08	91.26	0.97	2.87	3.02	52.86	0.00	0.31
177	2.45	27.4	453	410.9	0.0245	0.0037	0.10	99.75	114.08	97.63	0.93	2.85	2.96	47.44	0.00	0.33
178	2.45	36.4	453	410.9	0.0245	0.0027	0.16	110.33	130.12	114.92	0.75	1.80	2.95	42.76	0.00	0.56

179		2.45	36.4	453	410.9	0.0245	0.0027	0.16	104.13	130.12	104.09	0.71	1.92	3.35	45.09	0.00	0.54
180		2.45	34.9	453	410.9	0.0245	0.0037	0.08	99.35	115.81	92.65	0.83	1.98	2.96	52.19	0.00	0.35
181		2.45	34.9	453	410.9	0.0245	0.0037	0.08	98.87	115.81	98.18	0.80	1.98	3.17	34.15	0.00	0.37
182		2.45	36.5	453	410.9	0.0245	0.0037	0.08	101.06	116.51	106.96	0.86	2.91	3.04	52.16	0.00	0.31
183		2.45	36.5	453	410.9	0.0245	0.0037	0.08	98.89	116.51	93.21	0.85	2.94	3.05	54.39	0.00	0.34
184		2.70	37.6	461	485	0.0243	0.0051	0.30	158.88	201.04	160.83	0.60	2.94	3.73	7.35	0.10	1.00
185		2.70	37.6	461	485	0.0243	0.0051	0.60	159.9	185.66	148.53	0.50	0.87	1.86	53.02	0.00	1.00
186		2.16	39.2	388	524	0.0169	0.0084	0.57	1059.5	1239	991.2	0.49	0.75	2.54	51.60	0.13	0.99
187		2.16	39.2	388	524	0.0169	0.0084	0.57	1071.01	1338.8	1217.01	0.34	0.71	1.99	10.00	0.05	1.00
188		2.24	32.2	388	524	0.0194	0.0089	0.59	998.79	1201.3	1109.41	0.42	1.92	2.01	11.24	0.01	0.98
189		3.34	35.9	363	368	0.0158	0.0020	0.03	134.42	151.8	123.94	0.64	1.38	3.51	7.26	0.05	0.98
190		3.34	35.7	363	368	0.0158	0.0020	0.03	129.57	147.96	124.47	0.80	1.36	3.41	55.26	0.30	0.99
191		3.34	34.3	363	368	0.0158	0.0020	0.03	132.91	153.11	147.09	0.71	2.88	5.96	70.03	0.40	1.00
192		3.34	33.2	363	368	0.0158	0.0020	0.03	131.32	157.06	143.41	0.61	1.83	8.15	106.82	0.69	0.72
193		3.34	36.8	363	368	0.0158	0.0020	0.03	131.47	159.46	127.57	0.61	2.65	7.08	50.85	0.14	0.97
194		3.34	35.9	363	368	0.0158	0.0020	0.03	145.68	170.59	159.81	0.80	7.94	8.50	119.42	0.74	0.64
195	Xie et al. (2015)	4.17	31.1	582	441	0.0101	0.0060	0.11	25.84	38	31	0.59	1.78	3.73	52.24	0.03	0.75
196		4.17	34.5	582	441	0.0101	0.0060	0.11	26.14	37.83	31	0.56	1.70	3.73	47.45	0.05	0.86
197		4.17	32.5	481	441	0.0127	0.0060	0.21	34.3	42.06	33	0.57	1.54	4.13	48.66	0.07	0.99
198		4.17	30.1	582	441	0.0127	0.0060	0.21	33.73	39.87	32	0.56	1.82	4.13	48.93	0.01	0.90
199	Berry et al. (2004)	1.60	21.6	371	344	0.0127	0.0066	0.17	66.45	86.91	69.53	0.38	1.29	1.72	51.82	0.39	0.84
200		1.78	27.1	318	336	0.0266	0.0025	0.07	407.65	471.31	377.05	0.42	0.80	2.11	1.24	0.51	0.95
201		1.08	19.8	341	559	0.0222	0.0062	0.80	75.58	82.71	66.16	0.61	1.00	1.04	1.27	0.26	0.42
202		1.08	19.8	341	559	0.0222	0.0062	0.80	85.56	91.31	71.55	0.86	1.06	2.18	2.46	0.11	0.66
203		1.08	19.8	341	559	0.0222	0.0105	0.80	94.98	114.95	101.87	1.13	1.85	3.71	17.55	0.58	0.36
204		1.32	31.8	340	249	0.0313	0.0022	0.18	115.4	130.58	104.46	0.50	0.82	1.16	2.45	0.39	0.42
205		1.32	33	340	249	0.0313	0.0022	0.45	113.32	133.96	107.17	0.24	0.34	0.84	16.07	0.50	0.51
206		3.10	33.6	496	345	0.0245	0.0016	0.07	72.66	87.89	70.31	1.34	3.52	3.60	3.58	0.92	0.34
207		1.18	34.9	441	414	0.0301	0.0016	0.16	258.45	322.65	258.12	0.64	1.10	1.10	17.02	0.76	0.46
208		2.22	34.9	441	414	0.0301	0.0028	0.16	213.35	263.17	210.54	0.79	1.37	1.83	11.70	0.68	0.46

209	1.18	42	441	414	0.0301	0.0031	0.27	337.21	409.38	327.5	0.63	0.92	1.01	4.52	0.21	0.77
210	1.63	29.9	462	414	0.0244	0.0009	0.10	175.12	213.61	199.84	0.68	0.93	2.01	3.65	0.56	0.47
211	3.52	26.9	331	399.9	0.0303	0.0007	0.09	236	277	221.6	0.66	1.04	1.05	20.00	0.17	0.33
212	3.52	27.6	331	399.9	0.0303	0.0007	0.26	287.4	328	262.4	0.60	0.92	1.23	12.94	0.88	0.99
213	3.52	27.6	331	399.9	0.0303	0.0017	0.26	281.8	355	284	0.60	1.06	1.74	5.16	0.09	0.67
214	3.52	26.9	331	399.9	0.0303	0.0007	0.09	254	270	216	0.62	0.76	1.13	2.95	0.18	0.98
215	2.91	21.9	434	400	0.0188	0.0019	0.00	353.89	406.99	325.59	0.70	1.94	2.25	3.76	0.21	0.73
216	1.39	16	434	400	0.0188	0.0004	0.00	577.36	604.55	483.64	0.62	0.71	1.08	21.38	0.88	0.99
217	1.60	21	371	344	0.0127	0.0115	0.35	85.81	110.7	88.56	0.38	0.93	1.55	52.77	0.49	0.83
218	2.12	32	369	316	0.0201	0.0048	0.14	82.88	101.38	81.1	0.90	1.60	3.85	62.32	0.68	1.00
219	2.13	29.9	370	316	0.0265	0.0047	0.15	87.35	110.59	88.47	0.76	1.28	1.98	52.32	0.36	1.00
220	1.14	32.3	336	341	0.0177	0.0039	0.60	24.9	31.61	29.39	2.09	3.74	4.07	13.15	0.18	0.92
221	1.14	34	336	341	0.0177	0.0039	0.70	30.44	36.74	29.39	2.95	4.34	5.01	16.91	0.18	0.99
222	1.14	32.8	336	341	0.0177	0.0039	0.90	22.28	29.56	25.65	2.07	3.30	3.50	45.64	0.74	1.00
223	2.17	21.1	341	559	0.0222	0.0062	0.80	53.82	66.55	59.85	0.30	0.47	1.02	52.49	0.53	0.99
224	2.17	21.1	341	559	0.0222	0.0105	0.90	51.49	67.37	53.9	0.51	1.92	3.77	2.43	0.11	0.54
225	3.25	28.8	341	559	0.0222	0.0062	0.70	41.6	51.22	40.98	0.41	0.67	1.51	4.27	0.11	0.89
226	3.25	28.8	341	559	0.0222	0.0062	0.70	43.57	54.91	43.92	0.36	0.68	1.37	13.85	0.26	1.00
227	3.25	28.8	341	559	0.0222	0.0105	0.70	43.25	51.78	42.28	0.56	0.77	2.86	5.81	0.89	1.00
228	1.66	25.8	361	426	0.0213	0.0080	0.26	101.26	129.98	103.98	0.52	1.50	2.52	8.73	0.72	1.00
229	1.66	25.8	361	426	0.0213	0.0080	0.62	105.23	133.78	107.02	0.32	0.75	1.03	5.06	0.05	0.68
230	1.29	46.3	441	414	0.0412	0.0077	0.74	451	505.6	404.48	0.40	0.76	0.86	4.93	0.24	1.00
231	3.10	34.7	496	345	0.0245	0.0016	0.12	78.52	98.76	94.99	0.79	3.61	3.63	15.22	0.18	0.82
232	3.10	34.7	496	345	0.0245	0.0016	0.12	84.85	101.31	83.29	0.89	2.64	3.57	15.86	0.28	0.95
233	3.10	26.1	496	345	0.0245	0.0023	0.15	85.54	104.59	99.48	1.17	2.78	4.88	3.92	0.04	0.77
234	3.10	26.1	496	345	0.0245	0.0023	0.15	80.48	98.48	98.48	1.07	5.43	5.49	6.16	0.14	0.99
235	3.10	33.6	496	345	0.0245	0.0016	0.11	79	94.23	75.38	1.44	2.58	4.81	2.19	0.03	0.98
236	3.10	33.6	496	345	0.0245	0.0016	0.11	89.89	104.9	99.78	1.46	2.60	5.59	49.66	0.04	1.00
237	3.10	33.6	496	345	0.0245	0.0016	0.07	76.34	93.27	80.42	1.22	3.42	3.42	6.48	0.02	0.98
238	3.10	33.4	496	345	0.0245	0.0031	0.12	83.21	93.07	92.32	1.47	6.61	6.70	6.23	0.12	0.97
239	3.10	33.4	496	345	0.0245	0.0031	0.12	81.38	99.37	88.46	1.14	6.69	6.72	9.27	0.12	0.96

240		3.06	33.5	496	317	0.0245	0.0074	0.12	100.77	119.77	116.41	1.69	5.87	5.91	4.31	0.08	0.95
241		3.06	33.5	496	317	0.0245	0.0074	0.12	99.42	114.67	101.38	1.44	6.18	6.24	3.29	0.08	0.99
242		3.06	33.5	496	317	0.0245	0.0045	0.12	95.33	115.89	115.89	1.25	5.96	5.96	1.98	0.03	0.64
243		3.06	33.5	496	317	0.0245	0.0045	0.12	99.65	121.04	106.35	1.43	5.71	5.75	3.26	0.45	0.99
244		3.52	33.1	331	399.9	0.0194	0.0007	0.07	182.46	240.77	207.62	0.48	1.92	2.59	8.84	0.16	0.98
245		3.52	25.5	331	399.9	0.0194	0.0007	0.28	268.19	306	279	0.64	0.96	1.03	1.46	0.15	0.99
246		3.52	33.1	331	399.9	0.0194	0.0007	0.07	191.78	229	183.2	0.53	1.08	1.67	8.48	0.52	0.99
247		3.52	25.5	331	399.9	0.0303	0.0017	0.28	310.88	367	293.6	0.73	1.57	1.69	5.14	0.03	1.00
248		2.11	86	510	449	0.0246	0.0075	0.10	224	267.57	220.47	1.12	1.92	6.40	59.86	0.07	0.96
249		2.11	86	510	449	0.0246	0.0075	0.19	266.19	324.13	287.83	0.81	1.73	4.25	2.24	0.03	0.75
250		3.76	21.1	434.4	476	0.0247	0.0017	0.15	247.92	302.52	242.02	0.90	1.81	2.34	7.45	0.03	0.86
251		3.76	21.1	434.4	476	0.0247	0.0017	0.60	243.75	300.99	261.79	0.49	0.82	0.92	8.66	0.03	0.99
252		3.76	21.8	434.4	476	0.0247	0.0017	0.15	240.23	294.58	235.66	0.93	2.05	2.68	4.93	0.03	0.90
253	Cecen (1979)	2.56	32	480	745	0.0328	0.0089	0.18	4.13	4.74	4.27	2.73	4.48	6.72	26.03	0.00	0.62
254		2.56	32	480	745	0.0328	0.0089	0.16	4.04	4.59	4.13	2.48	3.97	5.96	26.09	0.00	0.63
255		2.56	32	480	745	0.0328	0.0089	0.14	3.63	4.16	3.75	2.33	3.66	5.49	25.89	0.00	0.61
256		2.56	32	480	745	0.0328	0.0089	0.13	3.02	3.46	3.11	2.19	3.42	5.13	26.04	0.00	0.61
257		2.56	32	480	745	0.0328	0.0089	0.11	2.46	2.82	2.54	1.99	3.10	4.65	26.02	0.00	0.61
258		2.56	32	480	745	0.0187	0.0089	0.09	2.23	2.56	2.3	2.06	3.26	4.89	25.93	0.00	0.64
259		2.56	32	480	745	0.0187	0.0089	0.07	2.01	2.3	2.07	1.83	2.89	4.33	25.98	0.00	0.58
260		2.56	32	480	745	0.0187	0.0089	0.05	1.84	2.1	1.89	1.95	3.03	4.54	26.02	0.00	0.56
261		2.56	32	480	745	0.0187	0.0089	0.04	1.71	1.96	1.76	1.88	2.88	4.31	26.18	0.00	0.56
262		2.56	32	480	745	0.0187	0.0089	0.02	1.6	1.84	1.65	1.69	2.69	4.02	26.14	0.00	0.62
263		2.56	32	480	745	0.0328	0.0089	0.18	4.13	4.74	4.27	2.73	4.48	6.72	25.93	0.00	0.61
264		2.56	32	480	745	0.0328	0.0089	0.16	4.04	4.59	4.13	2.48	3.97	5.96	26.15	0.00	0.56
265		2.56	32	480	745	0.0328	0.0089	0.14	3.63	4.16	3.75	2.33	3.66	5.49	26.04	0.00	0.59
266		2.56	32	480	745	0.0328	0.0089	0.13	3.02	3.46	3.11	2.19	3.42	5.13	26.00	0.00	0.57
267		2.56	32	480	745	0.0328	0.0089	0.11	2.46	2.82	2.54	1.99	3.10	4.65	26.04	0.00	0.61
268		2.56	32	480	745	0.0187	0.0089	0.09	2.23	2.56	2.3	2.06	3.26	4.89	25.99	0.00	0.60
269		2.56	32	480	745	0.0187	0.0089	0.07	2.01	2.3	2.07	1.83	2.89	4.33	25.99	0.00	0.62

270	2.56	32	480	745	0.0187	0.0089	0.05	1.84	2.1	1.89	1.95	3.03	4.54	26.07	0.00	0.59
271	2.56	32	480	745	0.0187	0.0089	0.04	1.71	1.96	1.76	1.88	2.88	4.31	26.07	0.00	0.58
272	2.56	32	480	745	0.0187	0.0089	0.02	1.6	1.84	1.64	1.69	2.69	4.02	26.07	0.00	0.60

APPENDIX B

DATABASE OF CIRCULAR REINFORCED CONCRETE COLUMNS

No.	Reference	a/d	f_c (MPa)	f_{yl} (MPa)	f_{yt} (MPa)	p_l	p_t	$P/A_g f_c$	V_y (kN)	V_m (kN)	V_u (kN)	δ_y (%)	δ_m (%)	δ_u (%)	α	β	γ
1	Berry et al. (2004)	5.73	33.2	373	312	0.0257	0.0044	0.06	162.22	192.00	163.40	0.80	2.18	3.03	51.97	0.02	1.00
2		3.65	34.8	371	312	0.0257	0.0044	0.06	286.08	343.00	319.69	1.00	2.70	5.20	22.47	0.00	0.93
3		6.77	33.8	373	342	0.0257	0.0044	0.06	128.33	149.00	119.28	0.66	1.23	3.44	52.25	0.03	1.00
4		5.69	40.0	305	389	0.0257	0.0126	0.00	110.00	134.00	130.82	1.03	4.13	5.55	55.58	0.00	1.00
5		5.60	35.1	305	263	0.0256	0.0187	0.01	30.17	37.00	30.59	2.35	4.75	7.01	56.34	0.61	1.00
6		3.88	33.0	294	207	0.0218	0.0248	0.34	62.18	77.00	77.06	0.73	4.23	4.38	2.75	0.02	1.00
7		4.17	26.0	308	308	0.0243	0.0076	0.21	113.03	139.00	128.72	0.50	1.24	3.58	19.49	0.01	1.00
8		4.19	28.5	308	280	0.0243	0.0153	0.59	132.90	163.00	130.76	0.48	1.20	2.00	52.37	0.13	1.00
9		2.09	28.4	303	300	0.0243	0.0075	0.24	540.06	687.00	669.00	0.42	1.69	3.67	12.49	0.03	1.00
10		2.09	32.9	303	423	0.0243	0.0080	0.41	632.40	781.00	711.23	0.47	1.39	2.73	45.56	0.06	1.00
11		2.10	32.5	307	280	0.0243	0.0261	0.37	687.16	812.00	801.70	0.41	0.97	2.29	5.37	0.13	1.00
12		2.10	32.5	307	280	0.0243	0.0261	0.74	735.00	937.00	822.50	0.59	1.65	2.59	65.29	0.00	1.00
13		2.62	29.9	448	372	0.0320	0.0102	0.20	326.62	364.00	353.50	1.72	4.19	6.56	51.43	0.02	0.83
14		4.19	32.3	337	466	0.0243	0.0062	0.14	121.69	142.00	113.86	0.59	1.18	4.54	45.82	0.23	1.00
15		4.18	40.0	474	372	0.0182	0.0064	0.53	177.09	212.00	189.30	0.56	1.18	2.02	50.87	0.21	0.99
16		4.19	39.0	474	338	0.0182	0.0147	0.74	180.00	206.00	165.17	0.58	0.97	1.59	59.57	0.72	1.00
17		2.11	38.0	423	300	0.0320	0.0142	0.19	399.84	461.00	410.86	0.92	3.51	5.18	7.33	0.02	0.89
18		2.11	37.0	475	300	0.0320	0.0142	0.39	449.49	579.00	462.90	0.84	2.50	3.76	3.00	0.08	1.00
19		7.05	38.8	240	240	0.0183	0.0063	0.05	23.98	31.00	24.51	0.53	1.29	2.99	50.14	0.05	1.00
20		7.05	36.2	240	240	0.0183	0.0063	0.09	25.80	33.00	26.10	0.45	1.00	2.13	47.40	0.08	1.00
21		4.02	34.5	448	620	0.0558	0.0145	0.24	32.00	41.00	37.71	1.40	5.13	8.01	58.51	0.00	1.00

22	6.25	35.8	475	493	0.0200	0.0063	0.07	1121	1289	1030.96	1.20	3.59	6.00	50.99	0.01	1.00
23	3.13	34.3	475	435	0.0200	0.0149	0.07	2443.9	2968	2558.4	0.99	4.67	7.78	16.13	0.03	1.00
24	3.12	24.1	446	441	0.0196	0.0141	0.10	52.85	59.00	47.59	1.03	5.13	9.43	43.72	0.28	1.00
25	3.12	23.1	446	441	0.0196	0.0141	0.21	60.55	73.00	58.74	0.81	1.82	8.21	17.18	0.19	1.00
26	6.24	25.4	446	476	0.0196	0.0068	0.10	26.07	32.00	30.30	1.07	3.61	7.37	57.42	0.11	0.88
27	3.12	24.4	446	441	0.0196	0.0141	0.10	48.54	63.00	50.16	0.56	2.39	7.57	18.70	0.11	1.00
28	6.24	23.3	446	476	0.0196	0.0068	0.10	25.94	30.00	24.17	1.09	2.01	4.75	48.69	0.02	0.78
29	1.79	26.5	399	355	0.0046	0.0284	0.14	90.54	117.00	93.40	0.35	3.10	3.31	1.60	0.01	0.81
30	2.36	31.6	375	366	0.0091	0.0101	0.21	68.02	102.00	81.62	0.42	1.15	4.53	0.86	0.02	0.91
31	2.38	31.6	345	335	0.0254	0.0341	0.21	115.77	146.00	116.80	0.64	1.80	5.72	7.10	0.01	0.58
32	1.76	31.3	363	381	0.0385	0.0134	0.00	150.23	176.00	170.91	0.75	1.52	3.42	11.75	0.03	0.85
33	1.76	29.3	363	381	0.0385	0.0134	0.12	175.13	212.00	206.23	0.71	1.56	3.44	5.27	0.00	0.78
34	2.35	30.5	363	381	0.0385	0.0063	0.12	121.40	154.00	148.19	0.65	1.48	3.65	3.76	0.08	0.92
35	2.35	30.9	363	381	0.0385	0.0063	0.23	142.85	174.00	160.98	0.70	1.66	3.65	2.94	0.00	0.88
36	4.72	29.0	448	434	0.0204	0.0094	0.09	63.88	74.00	72.64	1.19	3.64	5.46	50.61	0.01	1.00
37	4.72	35.5	448	434	0.0204	0.0094	0.09	61.82	72.00	71.13	1.18	4.17	4.22	53.89	0.00	0.88
38	4.72	35.5	448	434	0.0204	0.0094	0.09	66.77	77.00	77.19	1.51	5.40	5.56	51.74	0.01	1.00
39	4.72	35.5	448	434	0.0204	0.0094	0.09	64.91	77.00	71.87	1.14	3.67	6.87	55.50	0.32	1.00
40	4.72	32.8	448	434	0.0204	0.0094	0.09	67.14	79.00	71.67	1.04	2.43	5.91	9.51	0.00	1.00
41	4.72	32.8	448	434	0.0204	0.0094	0.09	60.58	68.00	62.74	1.15	2.00	5.86	11.15	0.00	1.00
42	4.72	32.5	448	434	0.0204	0.0094	0.09	59.28	75.00	68.54	0.95	2.72	6.60	11.12	0.00	1.00
43	4.72	27.0	448	434	0.0204	0.0094	0.10	64.51	74.00	69.47	1.12	2.37	6.61	15.03	0.00	0.93
44	4.72	27.0	448	434	0.0204	0.0094	0.10	57.00	68.00	55.28	1.02	3.58	5.45	11.66	0.00	1.00
45	4.72	27.0	448	434	0.0204	0.0094	0.10	60.55	72.00	57.85	0.90	3.18	5.07	11.24	0.00	1.00
46	6.29	41.1	455	414	0.0266	0.0089	0.15	307.98	357	357.50	1.45	8.72	8.74	8.23	0.03	1.00
47	2.11	38.3	428	430	0.0241	0.0114	0.31	467.00	582	482.80	0.83	3.63	5.03	3.75	0.02	1.00
48	2.11	35.0	486	434	0.0521	0.0304	0.33	849.00	1100	1099	1.85	7.53	9.47	51.20	0.05	1.00
49	8.57	36.6	477	445	0.0362	0.0092	0.30	130.70	149.00	135.10	1.87	3.41	9.09	6.32	0.00	1.00
50	8.57	40.0	477	437	0.0362	0.0060	0.27	152.84	175.00	168.44	1.92	3.72	5.75	9.99	0.15	1.00
51	8.57	38.6	477	445	0.0362	0.0092	0.28	144.20	167.00	153.40	1.86	7.17	9.31	7.11	0.00	1.00

52	4.15	31.0	462	607	0.0149	0.0070	0.07	248.00	285.00	262.00	0.89	5.09	7.32	7.74	0.13	1.00
53	8.30	31.0	462	607	0.0149	0.0070	0.07	134.00	151.00	150.00	1.80	7.34	9.14	23.42	0.06	0.96
54	10.38	31.0	462	607	0.0149	0.0070	0.07	83.77	98.00	90.03	1.68	3.01	10.46	43.75	0.00	0.89
55	4.15	31.0	462	607	0.0075	0.0070	0.07	152.00	180.00	170.00	0.63	1.45	5.21	15.00	0.00	1.00
56	4.15	31.0	462	607	0.0298	0.0070	0.07	415.00	480.00	479.00	1.42	7.14	7.30	29.21	0.15	1.00
57	3.15	34.5	441	607	0.0273	0.0090	0.09	467.00	555.00	548.00	0.95	6.72	7.22	12.05	0.02	1.00
58	8.39	34.5	441	607	0.0273	0.0090	0.09	174.00	203.00	193.00	1.78	9.00	9.53	43.40	0.00	1.00
59	10.49	34.5	441	607	0.0273	0.0090	0.09	162.35	190.00	188.60	2.06	14.04	14.66	15.00	0.00	1.00
60	3.16	31.4	448	431	0.0192	0.0054	0.05	341.17	410.00	327.73	1.39	5.70	7.24	53.18	0.10	0.98
61	3.16	34.6	448	431	0.0192	0.0054	0.04	342.16	431.00	416.76	1.14	4.84	6.63	31.69	0.02	0.81
62	3.16	33.0	461	434	0.0192	0.0081	0.04	371.20	453.00	409.10	1.06	4.92	8.24	12.24	0.05	1.00
63	6.96	65.0	419	1000	0.0328	0.0154	0.31	56.25	71.00	62.80	0.94	2.35	9.13	16.68	0.00	0.96
64	7.02	65.0	419	420	0.0328	0.0349	0.31	54.54	68.00	60.96	0.91	2.25	7.44	5.49	0.01	1.00
65	6.97	90.0	419	580	0.0328	0.0175	0.42	66.00	85.00	81.26	0.85	2.89	3.29	57.16	0.07	1.00
66	7.02	90.0	419	420	0.0328	0.0174	0.42	65.50	78.00	71.39	0.67	2.74	4.16	12.92	0.02	1.00
67	6.96	90.0	419	1000	0.0328	0.0154	0.21	60.72	74.00	62.53	1.10	3.43	8.62	6.45	0.01	1.00
68	6.96	90.0	419	420	0.0328	0.0343	0.42	79.08	94.00	90.67	0.86	1.97	6.60	4.19	0.00	1.00
69	3.13	56.2	455	455	0.0099	0.0013	0.13	244.95	308.00	260.70	0.74	2.81	3.45	1.92	0.20	1.00
70	3.13	56.3	455	455	0.0099	0.0013	0.11	234.00	293.00	234.40	0.64	1.64	3.86	4.90	0.03	0.69
71	3.13	57.0	455	455	0.0099	0.0013	0.10	225.88	272.00	217.90	0.59	1.58	3.34	0.70	0.00	0.88
72	3.13	52.7	455	455	0.0099	0.0013	0.11	226.82	265.00	212.24	0.68	2.12	2.97	7.50	0.01	0.47
73	4.15	37.2	462	607	0.0149	0.0070	0.12	275.27	330.00	258.60	1.23	5.22	7.32	9.24	0.14	1.00
74	4.15	37.2	462	607	0.0149	0.0035	0.06	240.26	288.00	224.30	1.16	5.25	5.28	51.19	0.07	0.81
75	6.20	32.6	315	352	0.0254	0.0017	0.19	190.42	237.00	214.30	0.75	3.01	3.76	2.62	0.08	1.00
76	5.42	60.6	430	414	0.0213	0.0176	0.00	97.18	117.00	111.74	1.41	4.23	10.11	53.08	0.00	0.77
77	5.42	62.6	430	414	0.0213	0.0176	0.00	99.25	111.00	106.70	1.30	4.21	11.36	5.91	0.02	0.92
78	5.42	69.6	430	414	0.0213	0.0192	0.00	107.18	137.00	120.70	0.95	2.84	10.23	7.47	0.00	0.50
79	5.42	69.6	430	414	0.0213	0.0192	0.10	138.58	167.00	133.85	0.66	1.35	2.81	45.99	0.02	0.59

80	5.42	69.6	492	414	0.0213	0.0192	0.10	148.69	177.00	141.79	0.77	1.30	3.13	2.39	0.05	1.00
81	5.42	69.6	506	414	0.0213	0.0192	0.10	144.10	176.00	140.65	0.70	1.28	3.16	4.67	0.08	0.71
82	5.42	69.6	506	414	0.0213	0.0192	0.10	139.19	171.00	136.52	0.75	2.09	3.00	11.67	0.05	0.72
83	5.42	69.6	492	414	0.0213	0.0192	0.10	149.95	182.00	120.60	0.77	1.83	3.79	4.31	0.02	0.70
84	5.48	32.7	565	434	0.0198	0.0092	0.04	131.26	153.00	132.30	1.59	4.44	7.55	52.91	0.00	0.99
85	5.48	34.2	565	434	0.0198	0.0092	0.04	136.76	159.00	144.30	1.85	4.44	10.70	58.37	0.00	1.00
86	5.48	33.9	565	434	0.0198	0.0092	0.04	139.91	157.00	152.60	1.81	8.66	13.15	57.86	0.00	0.62
87	6.83	22.0	379	379	0.0156	0.0024	0.19	87.17	106.00	99.85	0.85	1.90	3.83	2.65	0.81	0.76
88	4.74	36.5	459	692	0.0117	0.0053	0.00	59.26	70.00	62.32	1.01	2.99	6.17	14.55	0.00	0.95
89	4.74	36.5	459	692	0.0117	0.0053	0.00	61.52	74.00	66.43	1.09	2.53	5.13	11.84	0.07	1.00
90	4.74	35.6	459	692	0.0117	0.0053	0.00	82.80	106.00	84.80	1.62	8.14	11.05	1.00	0.00	1.00
91	4.19	27.0	337	466	0.0243	0.0112	0.61	142.63	175.00	140.10	0.58	0.86	2.53	66.01	0.20	1.00
92	8.04	34.5	448	620	0.0558	0.0145	0.24	18.00	19.00	15.25	2.50	3.77	7.86	15.89	0.09	1.00
93	4.02	34.5	448	620	0.0558	0.0145	0.35	38.50	42.00	37.35	1.74	3.81	7.99	15.75	0.00	1.00
94	6.96	90.0	419	1000	0.0328	0.0154	0.42	67.50	81.00	73.23	0.95	2.77	4.61	51.48	0.00	1.00
95	6.96	90.0	419	1000	0.0328	0.0154	0.42	63.50	78.00	62.75	1.22	1.81	5.92	57.50	0.00	1.00
96	5.48	31.7	565	434	0.0198	0.0092	0.04	172.00	192.00	185.75	2.05	9.67	10.50	110.94	0.00	0.68
97	2.11	30.6	436	316	0.0320	0.0051	0.00	255.00	289.17	280.70	1.21	1.67	2.26	12.08	0.15	0.22
98	1.57	30.1	436	328	0.0320	0.0051	0.00	376.20	391.65	319.50	1.53	1.70	1.85	5.90	0.42	0.88
99	2.09	29.5	448	372	0.0320	0.0038	0.00	269.10	280.66	230.80	1.47	1.50	3.56	35.53	0.53	0.61
100	2.09	33.4	436	326	0.0320	0.0051	0.10	287.20	352.28	295.90	1.05	1.51	2.01	7.14	0.21	0.72
101	1.57	35.0	436	326	0.0320	0.0051	0.10	419.18	504.83	394.60	1.31	2.36	2.58	5.31	0.36	0.95
102	1.83	36.7	482	326	0.0320	0.0038	0.18	398.42	486.64	392.30	1.15	1.79	1.94	53.55	0.70	0.46
103	2.09	33.2	436	326	0.0320	0.0038	0.00	233.45	270.46	216.37	0.84	1.33	3.58	71.13	0.72	0.41
104	2.11	30.9	436	310	0.0320	0.0039	0.00	226.53	284.83	222.80	0.86	1.44	2.22	10.19	0.11	0.35
105	1.50	32.8	296	0	0.0320	0.0000	0.00	204.46	239.26	194.30	0.52	0.91	1.22	7.38	0.12	0.67
106	1.18	28.8	366	368	0.0385	0.0047	0.00	120.87	176.36	141.09	0.34	0.98	2.30	10.78	0.20	0.27
107	1.18	29.3	366	368	0.0385	0.0094	0.00	152.73	203.82	166.50	0.49	1.28	2.64	13.18	0.17	0.32
108	1.19	28.6	366	0	0.0385	0.0000	0.13	140.65	158.23	128.00	0.22	0.31	0.64	1.40	0.39	0.64

109	1.18	29.8	366	368	0.0385	0.0047	0.12	143.82	192.49	153.99	0.20	0.63	1.71	2.06	0.62	0.68
110	1.18	28.6	366	368	0.0385	0.0094	0.13	178.28	225.30	180.24	0.33	0.99	1.99	2.83	0.27	0.43
111	1.18	31.4	366	368	0.0385	0.0134	0.12	163.70	213.14	170.51	0.34	0.95	2.44	1.25	0.49	0.27
112	1.18	30.5	366	368	0.0513	0.0094	0.12	178.27	228.03	193.00	0.32	0.96	1.97	2.29	0.42	0.78
113	1.19	28.7	366	0	0.0385	0.0000	0.25	165.37	188.44	150.75	0.18	0.26	0.43	2.24	0.51	0.66
114	1.18	27.8	366	368	0.0385	0.0047	0.26	160.22	191.95	141.10	0.19	0.62	1.21	1.52	0.68	0.69
115	1.18	30.5	366	368	0.0385	0.0094	0.24	192.49	238.42	190.74	0.21	0.95	1.79	2.73	0.36	0.49
116	1.18	31.3	366	368	0.0385	0.0134	0.23	215.36	279.09	218.10	0.31	1.28	2.33	2.71	0.32	0.54
117	1.18	31.3	363	381	0.0385	0.0063	0.12	186.70	246.58	197.26	0.37	0.96	2.08	1.99	0.69	0.77
118	1.79	31.1	363	0	0.0385	0.0000	0.12	114.57	132.03	105.62	0.28	0.42	0.79	1.88	0.80	0.58
119	1.76	31.2	363	381	0.0385	0.0063	0.12	150.39	186.48	149.18	0.55	1.10	2.38	2.15	0.29	0.55
120	1.76	20.5	363	381	0.0385	0.0063	0.18	137.25	171.15	136.92	0.56	1.28	2.38	4.99	0.24	0.30
121	1.18	31.1	363	381	0.0385	0.0063	0.23	194.13	234.05	187.24	0.28	0.96	1.67	3.28	0.32	0.41
122	1.76	29.7	363	381	0.0385	0.0063	0.24	158.58	201.24	160.99	0.44	1.13	2.29	2.41	0.34	0.42
123	1.76	18.9	363	381	0.0385	0.0063	0.38	145.46	176.08	140.86	0.45	1.11	1.96	2.70	0.32	0.53
124	1.76	41.3	363	381	0.0385	0.0063	0.18	187.03	228.59	182.87	0.50	1.04	2.15	2.03	0.33	0.50
125	2.06	26.8	454	200	0.0136	0.0013	0.00	290.60	331.10	264.88	0.48	0.63	0.90	15.16	0.21	0.66
126	2.06	31.2	438	200	0.0136	0.0013	0.00	276.00	326.30	261.04	0.45	0.67	0.88	17.18	0.12	0.61
127	2.09	26.6	303	300	0.0243	0.0112	0.57	615.28	729.06	705.98	0.38	1.13	1.42	44.67	0.12	1.00
128	2.09	37.5	436	328	0.0320	0.0051	0.00	310.91	321.38	290.70	2.31	2.83	4.13	23.09	0.53	1.00
129	2.09	37.2	296	328	0.0320	0.0051	0.00	199.64	220.69	215.70	0.72	0.96	3.90	50.64	0.01	0.61
130	2.62	36.0	436	328	0.0320	0.0051	0.00	237.92	276.18	268.50	3.23	4.00	4.02	40.31	0.82	1.00
131	2.09	31.1	436	328	0.0320	0.0076	0.00	325.80	330.92	320.30	2.51	2.74	3.88	33.97	0.09	0.22
132	2.09	28.7	448	372	0.0320	0.0102	0.20	403.40	445.13	378.60	1.88	3.66	5.49	56.07	0.15	1.00
133	2.11	31.2	448	332	0.0320	0.0102	0.20	415.92	437.35	384.75	2.07	3.26	5.21	54.93	0.17	1.00
134	2.09	29.9	448	372	0.0320	0.0051	0.20	380.10	407.10	325.68	1.89	2.14	2.22	10.71	0.67	0.89
135	1.57	28.6	436	328	0.0320	0.0102	0.10	510.80	525.82	420.66	2.23	2.59	4.09	30.10	0.11	0.71
136	2.09	33.7	424	326	0.0324	0.0051	0.00	270.62	316.38	273.80	1.11	2.04	4.37	5.63	0.35	1.00
137	2.09	34.8	436	326	0.0192	0.0051	0.00	188.32	230.34	204.80	0.85	3.31	4.96	9.08	0.12	1.00
138	2.62	34.3	436	326	0.0320	0.0051	0.10	258.82	312.36	249.89	1.04	1.87	2.92	6.73	0.09	0.53
139	2.11	32.3	436	332	0.0320	0.0076	0.00	283.90	332.54	315.00	1.20	1.89	4.09	11.62	0.16	0.53

140	2.11	33.1	436	310	0.0320	0.0077	0.00	285.54	340.48	191.30	1.05	3.96	6.02	4.88	0.06	0.40
141	2.09	37.0	475	340	0.0320	0.0047	0.39	372.82	489.30	391.44	0.48	1.40	2.12	7.19	0.00	0.46
142	3.32	35.9	240	240	0.0183	0.0063	0.05	64.23	74.85	67.51	0.57	1.05	3.50	55.33	0.04	1.00
143	3.30	34.4	240	240	0.0183	0.0063	0.10	66.58	79.69	63.75	0.49	0.98	3.04	49.45	0.01	1.00
144	1.19	26.5	375	335	0.0091	0.0427	0.25	138.83	174.82	139.86	0.50	2.10	3.83	2.35	0.01	0.37
145	1.19	26.5	382	335	0.0162	0.0312	0.12	137.61	181.65	145.32	0.67	1.83	3.68	5.71	0.02	0.43
146	2.35	31.6	382	387	0.0162	0.0075	0.10	80.29	101.73	97.67	0.64	1.73	4.69	8.22	0.01	0.64
147	1.18	30.2	366	368	0.0257	0.0094	0.12	190.96	254.24	203.39	0.46	1.32	2.81	6.40	0.12	0.30
148	1.76	32.0	363	381	0.0385	0.0063	0.00	144.24	168.20	134.56	0.83	1.52	2.80	46.51	0.17	0.36
149	1.76	42.2	363	381	0.0385	0.0063	0.09	170.62	212.18	169.74	0.57	1.33	2.68	2.46	0.27	0.41
150	1.54	30.0	462	361	0.0052	0.0028	0.06	326.31	399.52	391.40	0.35	1.26	2.61	2.45	0.04	0.67
151	1.54	30.0	462	361	0.0104	0.0017	0.06	462.81	587.36	557.60	0.48	1.68	1.78	1.90	0.18	0.83
152	2.11	35.0	468.2	434.4	0.0521	0.0270	0.15	774.00	985.10	979.90	1.69	9.25	9.53	48.03	0.13	1.00
153	2.65	34.7	458.5	691.5	0.0137	0.0010	0.00	123.22	143.26	114.61	0.45	0.68	1.14	52.17	0.79	1.00
154	2.65	34.7	458.5	691.5	0.0137	0.0010	0.00	150.98	164.42	131.54	0.80	1.03	1.32	50.24	0.00	0.57
155	3.12	24.3	446	441	0.0196	0.0141	0.20	59.84	77.00	61.24	0.78	2.71	7.15	19.31	0.13	1.00
156	2.11	39.4	427.5	430.2	0.0241	0.0114	0.15	442.40	550.00	542.00	0.96	4.12	5.53	8.55	0.04	0.94
157	1.57	34.4	436	326	0.0320	0.0038	0.10	357.28	437.45	359.70	1.02	1.46	1.69	14.45	0.22	0.58
158	2.06	29.8	454	200	0.0136	0.0013	0.00	332.58	405.48	387.90	0.55	1.94	2.44	14.70	0.10	0.72
159	2.09	36.2	436	326	0.0320	0.0102	0.10	378.86	436.30	349.04	1.34	4.37	4.90	31.49	0.10	0.95
160	2.65	35.4	458.5	691.5	0.0117	0.0026	0.00	143.58	169.79	163.20	0.79	2.42	4.06	7.95	0.04	0.76