Contents lists available at ScienceDirect

Advanced Engineering Informatics

journal homepage: www.elsevier.com/locate/aei

Advancing post-earthquake structural evaluations via sequential regression-based predictive mean matching for enhanced forecasting in the context of missing data

Huan Luo [a,b,*], Stephanie German Paal [b]

[a] College of Civil Engineering & Architecture, China Three Gorges University, Yichang, HuBei 443002, China
Zachry Department of Civil and Environmental Engineering, Texas A&M University, College Station, TX 77843, United States

ARTICLE INFO

Keywords:
Multiple imputation
Bayesian parameter estimation
Predictive mean matching
Sequential regression
Missing data
Post-earthquake structural evaluation

ABSTRACT

After an earthquake, every damaged building needs to be properly evaluated in order to determine its capacity to withstand aftershocks as well as to assess safety for occupants to return. These evaluations are time-sensitive as the quicker they are completed, the less costly the disaster will be in terms of lives and dollars lost. In this direction, there is often not sufficient time or resources to acquire all information regarding the structure to do a high-level structural analysis. The post-earthquake damage survey data may be incomplete and contain missing values, which delays the analytical procedure or even makes structural evaluation impossible. This paper proposes a novel multiple imputation (MI) approach to address the missing data problem by filling in each missing value with multiple realistic, valid candidates, accounting for the uncertainty of 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, conditional 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 data set. Two examples are carried out to validate the usefulness of the SRB-PMM approach based on a database including 262 reinforced concrete (RC) column specimens subjected to earthquake loads. The results from both examples suggest that the proposed SRB-PMM approach is an effective means to handle missing data problems prominent in post-earthquake structural evaluations.

1. Introduction

After an earthquake, every damaged building needs to be properly evaluated in order to determine its capacity to withstand aftershocks as well as to assess safety for occupants to return. These evaluations are time-sensitive as the quicker they are completed, the less costly the disaster will be in terms of lives and dollars lost. Recently, many advanced techniques have been developed to rapidly perform post-earthquake safety and structural assessments. For example, these include a data-driven framework for predicting the safety state of post-earthquake buildings [1] and automated post-earthquake building evaluations [2–7]. However, these methods cannot evaluate the residual load bearing capacity of damaged buildings due to earthquakes, and such evaluations are necessary for some damaged buildings to analyze their seismic performance resisting aftershocks such that the global collapse risk can be identified and rescue teams can take the necessary precautions (e.g., dismantling those damaged buildings having high global collapse risk). The evaluation of residual capacity requires detailed nonlinear structural analyses for each damaged building. However, in the earthquake field, there is often not sufficient time or resources to acquire all design information (e.g., material properties and reinforcement details) to do such a high-level structural analysis. Thus, post-earthquake survey data for some damaged buildings may be incomplete and contain missing values for critical design information. This is where the missing data problem is prevalent in post-earthquake survey data. In turn, this can delay the structural evaluation or even make it impossible. Therefore, it is necessary to develop approaches to address the problems associated with incomplete data.

An incomplete data set involves observations (i.e., data points) with missing values, as shown in Table 1. Table 1 shows an example where

* Corresponding author.
E-mail address: hluo@tamu.edu (H. Luo).

https://doi.org/10.1016/j.aei.2020.101202
Received 6 April 2020; Received in revised form 29 October 2020; Accepted 10 November 2020
Available online 23 January 2021

Table 1
Schematic format of an incomplete data set, where '*NAN*' represents a missing value, and missing values only exist in the partially observed explanatory variables $\mathbf{Z}_{(1)}$, $\mathbf{Z}_{(2)}$, and $\mathbf{Z}_{(3)}$.

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

three explanatory variables (or features/predictors) are partially observed and have missing values (represented by '*NAN*' values), making this data set incomplete. One common strategy to deal with this missing data problem is to bypass the data with missing values and use the available complete data for analysis and inference. The most popular way for this strategy is to simply discard every incomplete observation, transforming the incomplete data set into a reduced, but complete, data set. Nevertheless, considering all observations in the original incomplete data set are from realistic cases, this strategy involves throwing away a potentially large amount of useful information when missing data ratio is high, leading to biased inference, and finally misinterpreted conclusions [8–10]. Another effective strategy is to apply information theory (i.e., mutual information and interaction information) to derive a reliability function for analysis of incomplete data set [11]. However, these kinds of strategies are not always appropriate. In specific, in post-earthquake structural evaluations, bypassing 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. Another effective scheme is to impute the missing values with plausible candidates, resulting in an imputed, complete data set. In this way, this type of imputation approach maintains the size of the original incomplete data set without risking the loss of useful information. By using imputation methods, those damaged building data with missing values will be imputed and further structural analyses can be performed based on the imputed values to inform the rescue teams of the associated global collapse risk.

The most direct imputation method is single imputation, which is performed by filling in a candidate for each missing value, such as imputing each missing value with a fixed value (e.g., mean imputation where any missing values are replaced with the mean of that variable for all other cases, which will not alter the sample mean) or a single value estimated by regression predictions [12] or by nearest neighbor methods [13,14] (where each missing value on some incomplete observations is replaced by a value obtained from related cases in the whole set of observations). 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 [9,15,16]. Thus, analyses of the *imputed*, complete data set by single imputation methods fail to account for the uncertainty of missing data. As an alternative, a multiple imputation (MI) method was developed by Rubin [15] to address this drawback. The method of MI has become a popular means for handling incomplete data sets in statistical analyses. The MI approach involves filling in each missing value with multiple plausible candidates, creating *multiple imputed*, complete data sets for analyses. Each data set is analyzed independently using techniques designed for the complete data set, and then the analyzed results are combined in such a way that the uncertainty of missing data may also be incorporated into the analyses [9,15]. Two popularly used approaches to create multiple candidates for MI include joint modeling (*JM*) of a multivariate imputation model specification [17,18] for all of the partially observed explanatory variables (conditional on any fully observed variables) and fully conditional specifications (*FCS*) of a series of univariate imputation models [19–22] for each partially observed explanatory variable given the other variables.

JM involves specifying a joint distribution for the multivariate data and drawing candidates from the posterior predictive distribution of the missing data [17]. The *JM* methodology is attractive when the specified joint distribution provides a good fit to the multivariate data. The commonly used joint 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 [17]. However, it is often challenging to specify a correct joint distribution [22,23]. 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 [22]. Given starting values, *FCS* draws 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 joint 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 partially observed explanatory variables are fully conditional (e.g., the iterations cannot reach convergence) [23]. Additionally, both JM and FCS produce 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 [9]. This may lead to generated candidate values that cannot be used to do detailed structural analyses for post-earthquake safety and structural assessments (e.g., negative candidate value for reinforcement ratio of a reinforced concrete (RC) column). Therefore, these two methods are not appropriate in such a context.

In Bayesian parameter estimation, a joint distribution can be factored as a product of conditional and marginal distributions [24–26]. 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, we propose a novel MI approach called sequential regression-based predictive mean matching (*SRB-PMM*) to create multiple plausible candidates for imputing each missing value with consideration of the uncertainty due to missing data. The proposed approach simplifies the specification of a suitable multivariate imputation model into a much easier task of specification of a series of univariate models and is able to overcome the possible drawback of *FCS* where the specification of univariate imputation models may be incompatible. Further, the proposed SRB-PMM approach ensures that the plausible candidates are from realistic values instead of simulations, due to the use of PMM, which overcomes meaningless imputations and ensures that the generated candidates will never be outside the observed data range. Therefore, the proposed approach can generate candidate values that can always be used for post-earthquake safety and structural assessments by performing detailed seismic analyses of damaged structures.

For the purpose of this work, a database of 262 RC column specimens subjected to earthquake loads is used to validate the proposed *SRB-PMM approach*, since RC columns are an important lateral load resisting member in an RC frame structure. For an RC frame building damaged by an earthquake, the lateral displacement of damaged RC columns can be measured by a skilled triage team of structural engineers/certified inspectors visually, but the lateral strength at this deformation cannot be acquired directly and its precise magnitude requires a detailed seismic analysis. The lateral load carrying capacity is quantified by the lateral strength of an RC column at its damage state, and is of great importance in determining whether the damaged RC column is safe and functional or that immediate repair is required [27,28]. This paper utilizes the proposed approach to advance post-earthquake safety and structural assessments with an emphasis on strength prediction of RC columns in the context of missing data. The rest of this paper is organized as follows. Section 2 presents the proposed methodology. Section 3 designs and performs two examples to validate the proposed approach based on a database of 262 RC column specimens subjected to earthquake loads. The experimental results are presented and discussed in Section 4. Conclusions are made in Section 5.

2. Methodology

This section presents the formulation of the proposed SRB-PMM method, which couples sequential regression and predictive mean matching (PMM) to address missing data problems. First, the proposed SRB-PMM method is used to generate several potential candidates (e.g., *r* candidates) for each missing value in an incomplete data set. In this way, *r* imputed complete data sets will be formed. Second, a machine learning (ML) approach is utilized to select the most plausible data set from the pool of the *r* imputed data sets by minimizing a cost function (e.g., mean squared error (MSE)) using K-fold cross-validation (CV). The most plausible imputed data set will be the one that causes the fitted ML model to have the best performance (e.g., minimum MSE), and the candidates that filled in the most plausible imputed data set are regarded as plausible imputation values. Last, by independently repeating the first two procedures multiple times, multiple plausible candidates for each missing value can be generated to consider the uncertainty of missing data. The detailed procedure for the proposed method is presented below.

2.1. Sequential regression-based predictive mean matching (SRB-PMM)

Assume a data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, where $\boldsymbol{x}_i \in R^{p+q}$, $y_i \in R$ and $n \gg p+q$ is collected from a domain of interest. In this data set, 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 data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ incomplete. For the remainder of this paper, 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 $\boldsymbol{X}^{obs} = (\boldsymbol{X}_1, \cdots, \boldsymbol{X}_p) \in R^{n\times p}$ be a matrix with n observations, and each observation has p fully observed explanatory variables (i.e., there are no missing values for n observations in these p explanatory variables, such as $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_p$ shown in Table 1). Let $\boldsymbol{X}^{miss} = (\boldsymbol{Z}_{(1)}, \cdots, \boldsymbol{Z}_{(q)}) \in R^{n\times q}$ 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 1), and $\boldsymbol{Z}_{(1)}, \cdots, \boldsymbol{Z}_{(q)}$ have been ordered increasingly in terms of the missing data ratios. Let $\boldsymbol{y} \in R^n$ be a vector. Thus, the data set $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$ 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 data set is presented in Table 1. Let $\boldsymbol{O} = (\boldsymbol{O}_1, \cdots, \boldsymbol{O}_q) \in R^{n\times q}$ 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 $j = 1, \cdots, q$, 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) [8,26].

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 [15], where $\boldsymbol{\Theta} = (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_q)$ is the model parameters. The multivariate imputation model $p(\boldsymbol{X}^{miss}|\boldsymbol{X}^{obs}, \boldsymbol{\Theta})$ can be factored as follows [25]:

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

where p_j, $j = 1, \cdots, q$ are the conditional density functions and $\boldsymbol{\theta}_j$ is a vector of parameters in the conditional distribution (e.g., regression coefficients, dispersion parameter). Note that when imputing the missing values for a partially observed variable (e.g., $\boldsymbol{Z}_{(1)}$), only the model parameter (e.g., $\boldsymbol{\theta}_1$) related to this variable (e.g., $\boldsymbol{Z}_{(1)}$) is used, and other model parameters (i.e., $\boldsymbol{\theta}_2, \cdots, \boldsymbol{\theta}_q$) are not required. Therefore, The distribution $p_j(\boldsymbol{Z}_{(j)}|\boldsymbol{Z}_{(1)}, \cdots, \boldsymbol{Z}_{(j-1)}, \boldsymbol{X}^{obs}, \boldsymbol{\theta}_j)$ $(j > 1)$ or $p_j(\boldsymbol{Z}_{(j)}|\boldsymbol{X}^{obs}, \boldsymbol{\theta}_j)$ $(j = 1)$ depends only on parameter $\boldsymbol{\theta}_j$.

Each conditional regression model in Eq. (1) is selected based on the type of variable $\boldsymbol{Z}_{(j)}$. For example, if $\boldsymbol{Z}_{(j)}$ is continuous, a normal linear regression model can be selected; if $\boldsymbol{Z}_{(j)}$ is binary, a logistic regression model can be used; if $\boldsymbol{Z}_{(j)}$ is categorical, a multinomial logistic regression model can be utilized; if $\boldsymbol{Z}_{(j)}$ is count variable, a Poisson loglinear model can be employed. Eq. (1) is initiated by regressing the variable with the fewest number of missing values (i.e., $\boldsymbol{Z}_{(1)}$), $\boldsymbol{Z}_{(1)}$ on $\boldsymbol{X}^{obs}$, where the missing values are imputed by *PMM* based on 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 $p_2(\boldsymbol{Z}_{(2)}|\boldsymbol{Z}_{(1)}, \boldsymbol{X}^{obs}, \boldsymbol{\theta}_2)$. This means, $\boldsymbol{Z}_{(1)}$ is imputed on $\boldsymbol{U}_1 = \boldsymbol{X}^{obs}$, $\boldsymbol{Z}_{(2)}$ is imputed on $\boldsymbol{U}_2 = (\boldsymbol{X}^{obs}, \boldsymbol{Z}_{(1)})$ where $\boldsymbol{Z}_{(1)}$ has imputed values, $\boldsymbol{Z}_{(3)}$ is imputed on $\boldsymbol{U}_3 = (\boldsymbol{X}^{obs}, \boldsymbol{Z}_{(1)}, \boldsymbol{Z}_{(2)})$ where $\boldsymbol{Z}_{(1)}$ and $\boldsymbol{Z}_{(2)}$ have imputed values, and others (i.e., $\boldsymbol{Z}_{(4)}, \cdots, \boldsymbol{Z}_{(q)}$) are imputed in a similarly sequential manner. The detailed imputation procedure for imputing each partially observed explanatory variable using *SRB-PMM* is presented below.

2.1.1. Bayesian inference for sequential regression-based model parameter

Since missing values exist in $\boldsymbol{Z}_{(j)}$, the model for $\boldsymbol{Z}_{(j)}$ cannot be established directly. For the model $p_1(\boldsymbol{Z}_{(1)}|\boldsymbol{X}^{obs}, \boldsymbol{\theta}_1)$, 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:

$$p_1\left(\boldsymbol{Z}_{(1)}^{obs}, \boldsymbol{Z}_{(1)}^{miss}|\boldsymbol{X}^{obs}, \boldsymbol{\theta}_1\right) = p_1\left(\boldsymbol{Z}_{(1)}^{miss}|\boldsymbol{Z}_{(1)}^{obs}, \boldsymbol{X}^{obs}, \boldsymbol{\theta}_1\right) \times p_1\left(\boldsymbol{Z}_{(1)}^{obs}|\boldsymbol{X}^{obs}, \boldsymbol{\theta}_1\right) \tag{2}$$

In this work, the variables in the reinforced concrete (RC) column data set are all continuous. Therefore, 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:

$$\boldsymbol{Z}_{(1)}^{obs} = \boldsymbol{X}^{*obs1}\boldsymbol{\beta}_1 + \boldsymbol{\varepsilon}_1 \tag{3}$$

where $\boldsymbol{\beta}_1 = (\beta_{11}, \cdots\beta_{1p})$ 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), $\boldsymbol{\varepsilon}_1 = (\varepsilon_{11}, \cdots, \varepsilon_{1(n_{ob1})})$ 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., $size(\boldsymbol{X}^{*obs1}, 1) = size(\{\boldsymbol{X}^{obs} : o_{i1} = 1\}, 1) = n_{ob1}$) and $\varepsilon_{11}, \cdots, \varepsilon_{1(n_{ob1})}$ i.i.d.$N(0, \sigma_1^2)$ or $\boldsymbol{\varepsilon}_1$ $N(0, \sigma_1^2\boldsymbol{I})$, 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 $\left(\boldsymbol{X}^{*obs1}\boldsymbol{\beta}_1, \sigma_1^2\boldsymbol{I}\right)$ [26], 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:

$$p\left(\boldsymbol{\beta}_1, \sigma_1^2|\boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right) = p\left(\boldsymbol{\beta}_1|\sigma_1^2, \boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right) \times p\left(\sigma_1^2|\boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right) \tag{4}$$

From Eq. (4), the posterior joint distribution of unknown model parameters $\left(\boldsymbol{\beta}_1, \sigma_1^2\right)$ 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 $p\left(\boldsymbol{\beta}_1|\sigma_1^2, \boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right)$, respectively. Throughout this paper, a *g*-prior [29] is used for these unknown model parameters $\left(\boldsymbol{\beta}_1, \sigma_1^2\right)$. With the use of *g*-prior**,** 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 [26]:

$$\left\{\sigma_1^2\middle|\boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right\}\sim \text{inverse} - \text{gamma}\left(\frac{1+n_{ob1}}{2}, \frac{\widehat{\sigma}_1^2 + SSR}{2}\right) \tag{5}$$

$$\left\{\boldsymbol{\beta}_1\middle|\sigma_1^2, \boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right\}\sim N\left(\frac{g}{g+1}\widehat{\boldsymbol{\beta}}_1, \frac{g}{g+1}\sigma_1^2\left(\left(\boldsymbol{X}^{*obs1}\right)^T\boldsymbol{X}^{*obs1}\right)^{-1}\right) \tag{6}$$

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

Since we can sample from both of these two conditional distributions, a sample value of $\left(\boldsymbol{\beta}_1, \sigma_1^2\right)$ sampled from the posterior joint distribution $p\left(\boldsymbol{\beta}_1, \sigma_1^2|\boldsymbol{X}^{obs1}, \boldsymbol{Z}_{(1)}^{obs}\right)$ can be made by first sampling the $\sigma_1^2$from Eq. (5) and then sampling the $\boldsymbol{\beta}_1$ from Eq. (6) given the sampled 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 $\left\{\left(\boldsymbol{\beta}_1, \sigma_1^2\right)_s\right\}_{s=1}^{S}$ 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 Monte Carlo approximation, where $\overline{\boldsymbol{\beta}}_1 = (1/S)\sum_{s=1}^{S}\left(\boldsymbol{\beta}_1\right)_s$ and $\overline{\sigma}_1^2 = (1/S)\sum_{s=1}^{S}\left(\sigma_1^2\right)_s$. Given the sampled model parameters $\left(\overline{\boldsymbol{\beta}}_1, \overline{\sigma}_1^2\right)$, a regression model can be established by inserting the model parameters into Eq. (3). We now describe 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.

2.1.2. Predictive mean matching (PMM) integrated with a k-fold cross-validation (CV) algorithm

Different from other imputation approaches, the goal of the regression model for *PMM* 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 [30–33]. The similarity is measured by the Euclidean distance between the fitted values for the observed data and the predicted values for the missing data based on the established regression model. For each missing value, the *PMM* first identifies a set of cases with 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. However, in this procedure, the randomly selected case may not be a plausible candidate, since there is no standard method to evaluate whether or not the selected one is plausible.

To solve this problem, we present a hybrid approach to select plausible candidates based on the k-fold cross-validation (CV) algorithm [34]. The purpose of this hybrid method is not to evaluate if a randomly selected single candidate for one missing value in one partially observed explanatory variable is plausible. Instead, it evaluates the *imputed*, complete data set 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 data set. This is because for an incomplete data set, there is an underlying pattern that can be captured by ML methods. Observations (i.e., data points) with and without missing values should follow that pattern. The plausible candidates should be able to make the imputed observations follow the underlying pattern. Conversely, inappropriate imputations may lead to imputed observations that become outliers and deviate from that pattern. Therefore, plausible imputations can result in an imputed complete data set where a fitted ML model should have good generalization performance, while inappropriate imputations may cause a fitted ML model that has poor generalization performance due to the negative effect of potential outliers. In this way, the generated plausible candidates can be reasonably justified by the fitted ML model that has the best performance.

We denote that $\boldsymbol{X}^{obs0} = \left\{\boldsymbol{X}^{obs} : o_{i1} = 0\right\}$, $\boldsymbol{X}^{*obs0}$ 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., $size\left(\boldsymbol{X}^{obs0}, 1\right) = size\left(\left\{\boldsymbol{X}^{obs} : o_{i1} = 0\right\}, 1\right) = n_{ob0}$), and $n_{ob0} + n_{ob1} = n$. The detailed procedure regarding the donor pool (i.e., selected close cases) generation for the missing values in $\boldsymbol{Z}_{(1)}^{miss}$ using the *PMM* algorithm is summarized in **Algorithm 1**:

Algorithm 1. (*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 nearest cases as the candidates for each missing value $Z_{(1),i}^{miss}$ in $\boldsymbol{Z}_{(1)}^{miss}$:
for all $i = 1, \cdots, n_{ob0}$ do
    2.1) Calculate the Euclidian distance vector $\boldsymbol{d}_i = \left\|\widehat{\boldsymbol{Z}}_{(1)}^{obs} - \widehat{Z}_{(1),i}^{miss}\right\|$.
    2.2) Sort $\boldsymbol{d}_i$ increasingly to obtain an increasingly ordered vector $\boldsymbol{d}_i = \left(d_{i(1)}, \cdots, d_{i(n_{ob1})}\right)$.
    2.3) Select r nearest cases from $\boldsymbol{Z}_{(1)}^{obs}$ corresponding to the first r close entries (i.e., $d_{i(1)}, \cdots, d_{i(r)}$) 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 1 above, each missing value in $\boldsymbol{Z}_{(1)}^{miss}$ has *r* candidates to impute. For each missing value, randomly sample one of the *r* candidates to impute the missing value. After all the missing values in $\boldsymbol{Z}_{(1)}^{miss}$ are imputed in the same way, an imputed $\boldsymbol{Z}_{(1)}^{miss}$ 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 $\left(\boldsymbol{Z}_{(1)}^{obs}, \widehat{\boldsymbol{Z}}_{(1),l}^{miss}\right)$, $l = 1, \cdots, r$ 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 $\boldsymbol{Z}_{(2)}$,

$U_1 = X^{obs}$ is updated by $U_{2,l} = \left(U_1, \widehat{Z}_{(1),l}\right), l = 1, \cdots, r$. Then **Algorithm 2** is developed to impute $Z_{(j)}, j = 2, \cdots, q$ in a sequential way.

Algorithm 2. (*Sequentially impute the missing values for* $Z_{(j)}, j = 2, \cdots, q$)

```
Given the $\{U_{2,l}\}_{l=1}^{r}$, where $U_{2,l} = \left(X^{obs}, \widehat{Z}_{(1),l}\right)$.
for all $l = 1, \cdots, r$ do
  for all $j = 2, \cdots, q$ do
    1) Compute the model parameters $\left(\overline{\beta}_j, \overline{\sigma}_j^2\right)$ using Eqs. (2–6) with the replacement
  of variables and parameters for $Z_{(j)}$, i.e., $p_j\left(Z_{(j)} | U_{j,l}, \beta_j, \sigma_j^2\right)$, and $X^{obs}$ is replaced by
  $U_{j,l}$.
    2) Generate $r$ candidates for each missing value in $Z_{(j)}^{miss}$ using algorithm 1 with
  the replacement of variables and parameters for $Z_{(j)}$.
    3) Randomly select one from the $r$ candidates for imputing each missing value in
  $Z_{(j)}^{miss}$.
    4) Denote the finally imputed $Z_{(j)}$ as $\widehat{Z}_{(j),l}$ and update the $U_{j+1,l} = \left(U_{j,l}, \widehat{Z}_{(j),l}\right)$.
end for $j$
    5) Set $\widehat{D}_l = \left(X_l^{impute}, y\right)$, where $\widehat{D}_l$ is an imputed, complete data set and
  $X_l^{impute} = U_{q+1,l} = \left(U_{q,l}, \widehat{Z}_{(q),l}\right)$.
  end for $l$
```

By implementing **Algorithm 2**, one can obtain *r imputed*, complete data sets $\left\{\widehat{D}_l\right\}_{l=1}^{r}$. Next, we use a k-fold cross-validation (CV) algorithm to minimize a cost function and determine which imputed data set is the most plausible based on an ML technique. The following procedure is used to select the most plausible imputed data set, which is defined as the one capable of minimizing the cost function $CF\left(y, f\left(X^{impute}\right)\right)$ by a k-fold cross-validation procedure, where $CF(\bullet)$ represents the cost function and $f(\bullet)$ represents an ML technique:

Algorithm 3. (*Selection of the most plausible imputed data set by K-fold CV procedure*)

```
Given the $r$ imputed data sets $\left\{\widehat{D}_l\right\}_{l=1}^{r}$, where $\widehat{D}_l = \left(X_l^{impute}, y\right)$, cost function $CF(\bullet)$,
  ML technique $f(\bullet)$.
for all $l = 1, \cdots, r$ do
  1) Compute the cost by K-fold CV procedure:
    $CV_{K-fold}\left(\widehat{D}_l\right) = \frac{1}{K}\sum_{k=1}^{K} CF\left(y_{n_k}, f\left(X_{n_k,l}^{impute}\right)\right)$
end for $l$
2) Choose the imputed data set that has the $min\left(\left\{CV_{K-fold}\left(\widehat{D}_l\right)\right\}_{l=1}^{r}\right)$.
```

In **Algorithm 3**, n_k is the size of the kth group (i.e., $n_k = floor(n/K)$); y_{n_k} is the observed response variable for the kth group in terms of the lth *imputed*, complete data set $\widehat{D}_l$; $f\left(X_{n_k,l}^{impute}\right)$ is the predicted response for the kth group by an ML technique $f(\bullet)$ trained on $\left(X_{-n_k,l}^{imputed}, y_{-n_k}\right)$ in terms of $\widehat{D}_l$; $\left(X_{-n_k,l}^{imputed}, y_{-n_k}\right)$ is the complementary set of $\left(X_{n_k,l}^{imputed}, y_{n_k}\right)$ in $\widehat{D}_l$.

2.1.3. Generation of an ensemble of multiple most plausible imputed data sets

Using **Algorithms 1 – 3** above, the most plausible *imputed*, complete data set can be determined. The *m* most plausible *imputed*, complete data sets to constitute an ensemble can be created for MI analyses to account for the uncertainty of missing data by independently repeating **Algorithms 1 – 3** *m* times. Each *imputed*, complete data set 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 Fig. 1 to illustrate this procedure.

3. Illustrative examples

This section presents the details of the numerical experiment design and validation for the performance of the *SRB-PMM* in advancing post-earthquake safety and structural assessment in the context of missing data. Two examples are designed. The first example is to evaluate the capabilities of the proposed *SRB-PMM* in improving the maximum lateral strength prediction performance based on an RC column data set subjected to ten different missing data ratios. The second one intends to illustrate the practical application of the *SRB-PMM* in post-earthquake structural analysis when the target damaged RC column is missing critical structural information. The detailed information is introduced below.

3.1. Lateral strength of RC columns

In structural and earthquake engineering, RC columns are the primary lateral load resisting members in an RC frame building. The lateral load-carrying capacity of RC columns is a critical factor to evaluate if a damaged column is still safe and functional or immediate repair is required. The loss of lateral load-carrying capacity is typically defined by the column's lateral displacement, where the lateral load-carrying capacity drops below 80% of the maximum lateral strength [35]. In the post-earthquake field, for a damaged RC frame building, the lateral displacement and other visual damages (e.g., concrete cracking and spalling) of the damaged columns can be measured and detected by a skilled triage team of structural engineers/certified inspectors visually, but their lateral load-carrying capacity corresponding to different damage states (i.e., lateral displacement) cannot be acquired directly. Further, a seismic analysis requires the damaged columns' structural feature information such as geometry and material properties. Although the magnitudes of geometry can be measured visually, the information regarding the material properties (e.g., concrete compressive strength and reinforcement yield stress) is most likely unknown. The unknown information leads to missing data problems, which can delay the structural evaluation or even make it impossible. This work utilizes the proposed approach to advance post-earthquake safety and structural assessments with the emphasis of strength prediction of RC columns in the context of missing data.

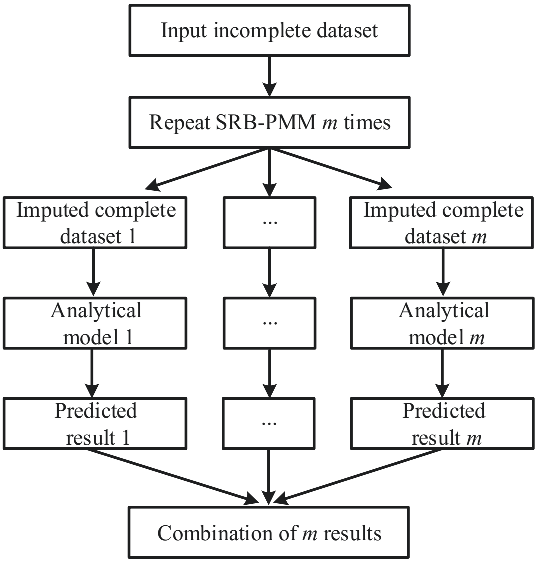


Fig. 1. Schematic flowchart for the prediction based on an ensemble of *m* analytical models.

3.2. RC column data set

The *RC column* data set, which is taken from the authors' previous work [36], is used to perform the two examples. Each column specimen in the data set is subjected to reversed cyclic loads, which simulate the earthquake loads. 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 maximum lateral strength V_m, which is defined as the maximum shear force in the hysteretic force–displacement curve. Table 2 presents the statistical properties of the column features and response variable. Note that some of the predictors are normalized in Tables 2 to maintain commonly used terminologies. More detailed information regarding the data set can be found in [36].

Since the RC column data set does not contain missing values, the two examples are performed on synthetic incomplete data sets. For the example 1, the synthetic incomplete data sets with ten different missing data ratios are generated from the complete column data set 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 the example 2, an RC column randomly sampled from the RC column data set serves as a case study of the target damaged RC column which hypothetically is missing some critical structural information when surveyed in a post-earthquake state. The 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, as introduced in Section 3.1. 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 examples is presented in Section 3.3.

3.3. Designs of two examples

3.3.1. Example 1

As introduced in Section 3.1, the maximum lateral strength of an RC column is a critical factor to evaluate if damaged RC columns have lost the lateral load-carrying capacity. Thus, it is important to accurately predict the maximum lateral strength of the RC columns subjected to earthquake loads. The purpose of this example 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 data set subjected to ten different missing data ratios and thus, to investigate how the missing data ratio affects its performance. Given an incomplete data set, any standard machine learning (ML) approach would fail to directly construct an appropriate data-driven model, as the original analysis procedures are only valid for complete data sets and are not designed to handle missing data [37]. This is because an incomplete data set has no real numbers (e.g., empty or 'NAN') for the missing values, but any standard ML method is essentially performed based on matrix operations, which require a matrix with full real numbers. Thus, the missing values must be addressed (e.g., either removing the observations with missing values or imputing the missing values) before any standard ML methods can be employed. Additionally, the ML model trained on a reduced complete data set serves as the baseline, where the reduced complete data set is formed by removing the observations with missing values in an incomplete data set. The ML model trained on an imputed complete data set (i.e., equal size with the original incomplete data set and maintaining all the original information) is the target model that can improve predictions for the baseline model. This is because the training set (i.e., the reduced data set) for the baseline model is different from the one for the target ML model (i.e., the imputed data set), and the reduced data set may miss useful information when compared to the imputed data set.

Table 2
Statistical properties for the RC column data set.

Parameter	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, $P/A_g f_c$	0	0.9	0.26	0.19
Maximum shear force, V_m (kN)	30	1339	212	182

The synthetic incomplete data sets are generated in the following way. First, for the original complete RC column data set, we use the 10-fold cross-validation procedure 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 the 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. Therefore, for each missing data ratio, there are ten incomplete training sets that are generated from the ten complete training sets introduced above. 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).

Least squares support vector machines for regression (LS-SVMR) [38–40] has achieved great success in structural and earthquake engineering [41–47] and thus is used to construct the data-driven model employed in this work. Other ML approaches may also be valid, but they will not be discussed since this is not the focus of this paper. 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 the Section 2 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) and *FCS* (with a univariate normal model) 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.

3.3.2. Example 2

In the second example, 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 RC column data set is also used in this example. Specifically, we first randomly sample an RC column from the 262 column specimens, and this column then serves as the target damaged column with missing critical feature information. The critical feature information considered in this example 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 missed in 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 data sets are generated based on the remaining 261 column specimens in a similar way as in the *Example 1* but with two differences. The first difference is that this example only has one incomplete data set 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 example, the partially observed explanatory variables are restricted to the mentioned five features.

In this example, we limit the missing data ratio to 5% and 10%. Therefore, in total, there are two synthetic incomplete data sets. The sampled column missing the information pertaining to the five critical features is then added to these two synthetic, incomplete data sets. Then, the *SRB-PMM* method is used to impute the missing values in the synthetic, incomplete data sets. 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 hysteretic force–displacement relation.

3.4. Implementations

Both examples 1 and 2 require the implementation of the proposed *SRB-PMM* method. To implement the proposed approach, some parameters and the conditional density functions (CDFs) of partially observed variables introduced in Section 2 need to be established. Since the variables in the RC column data set are all continuous, we specify a normal linear model for each CDF. Therefore, the dispersion parameter $\sigma_j^2$and regression coefficient $\boldsymbol{\beta}_j$ can be drawn from inverse-gamma and multivariate normal distributions respectively, as introduced in Section 2 and Algorithm 2. The number of close cases *r* presented in Algorithms 1–3 is set to five. The cost function (i.e, $CF(\cdot)$) presented in Algorithm 3 is mean squared error (MSE), which is evaluated by LS-SVMR based on the 10-fold cross-validation procedure. The number of plausible candidates, *m* presented in Section 3.1.3 is set to three. After these parameters are determined, Algorithms 1–3 can be performed to implement the proposed *SRB-PMM* method. The detailed implementation of the *JM* and *FCS* methods can be found in [17,22]. The *m* candidates to account for the uncertainty of missing data for the *JM* and *FCS* methods are also set to three. All codes regarding the Example 1 are implemented in Matlab. To illustrate the post-earthquake structural evaluation, the OpenSees [48] is used to perform the seismic analysis of the sampled column with the imputed feature information for Example 2.

3.5. Performance quantification criteria

The predictive performance in example 1 is quantified comprehensively by the coefficient of determination (R^2), mean absolute error (MAE), and root mean squared error (RMSE) metrics. Given a response variable $\boldsymbol{y} = \{y_i\}_{i=1}^n$ and predicted response $\widehat{\boldsymbol{y}} = \left\{\widehat{y}_i\right\}_{i=1}^n$, R^2, MAE, RMSE are calculated as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{\sum_{i=1}^n (y_i - \overline{y})^2} \tag{7}$$

$$MAE = \frac{1}{n}\sum_{i=1}^n |y_i - \widehat{y}_i| \tag{8}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^n (y_i - \widehat{y}_i)^2}{n}} \tag{9}$$

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

4. Results and discussion

In this section, the experimental results of two examples are presented to validate the performance of the proposed *SRB-PMM* approach. For the first example, 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 example, 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 in advancing the post-earthquake safety and structural evaluation in the context of missing data is presented.

4.1. Results for example 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 Fig. 2. Note that the results for *Complete-LS-SVMR* do not vary 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 Fig. 2, 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 is inversely related to the missing data ratio, which is to be expected. Additionally, compared to the results of *Delete-LS-SVMR* that serve as the baseline, 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*

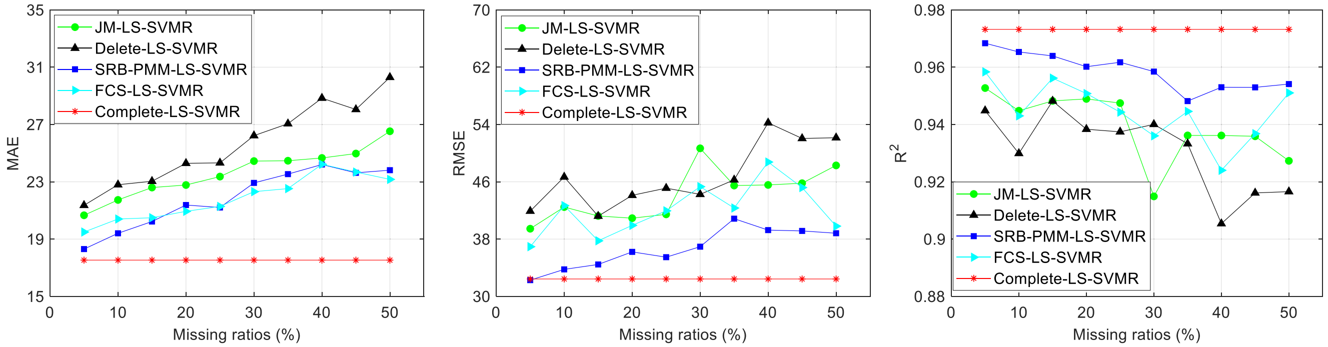


Fig. 2. The performance comparison of *SRB-PMM-LSSVMR, FCS-LS-SVMR, JM-LS-SVMR, Delete-LS-SVMR,* and *Complete-LS-SVMR* in terms of the average R^2, RMSE, and MAE metrics versus ten missing data ratios.

suggests that directly deleting the observations with missing values is not an effective way to handle the missing data since it degrades the prediction performance of the data-driven modeling procedure substantially.

To further investigate these findings, the following criteria [49] 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:

$$\mathrm{R}^2\ \text{improvement}\ (\%) = 100 \times \left(\frac{R^2\ \text{with imputation}}{R^2\ \text{without imputation}} - 1\right) \tag{10}$$

$$\text{RMSE improvement}\ (\%) = 100 \times \left(1 - \frac{\text{RMSE with imputation}}{\text{RMSE without imputation}}\right) \tag{11}$$

$$\text{MAE improvement}\ (\%) = 100 \times \left(1 - \frac{\text{MAE with imputation}}{\text{MAE without imputation}}\right) \tag{12}$$

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 3. We then use the average improvements 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 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 3. Both *JM* and *FCS* occasionally degrade the prediction performance, which is illustrated by the appearance of some negative values in Table 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.

4.2. Results for example 2

An RC column (specimen No. 6 in Tanaka and Park [50]) is randomly sampled from the column data set, which hypothetically serves as the target damaged column collected in the earthquake field. The column's critical feature information introduced in *Section 3.3.2* 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 3.3.2*. After the synthetic, incomplete data sets are generated, we independently run the *SRB-PMM* three times for each missing data ratio to account for uncertainty due to the missing data. For each run, a group of 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. Fig. 3(a,b,c) and 4(a,b,c) present the imputed values and the seismic analysis results of the sampled column generated from the synthetic incomplete column data sets with 5% and 10% missing data ratios, respectively. Fig. 3(d) and 4 (d) show the average of the simulated results to account for the uncertainty due to the missing data.

By observation of Fig. 3(a,b,c) and 4(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, Fig. 3(a,c) show that the seismic analysis results underestimate the actual seismic performance of the sampled column, while Fig. 3(b) overestimates the true seismic performance; and for the 10% missing data ratio, Fig. 4(a,c) overestimate the actual seismic performance in spite of Fig. 4(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

Table 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

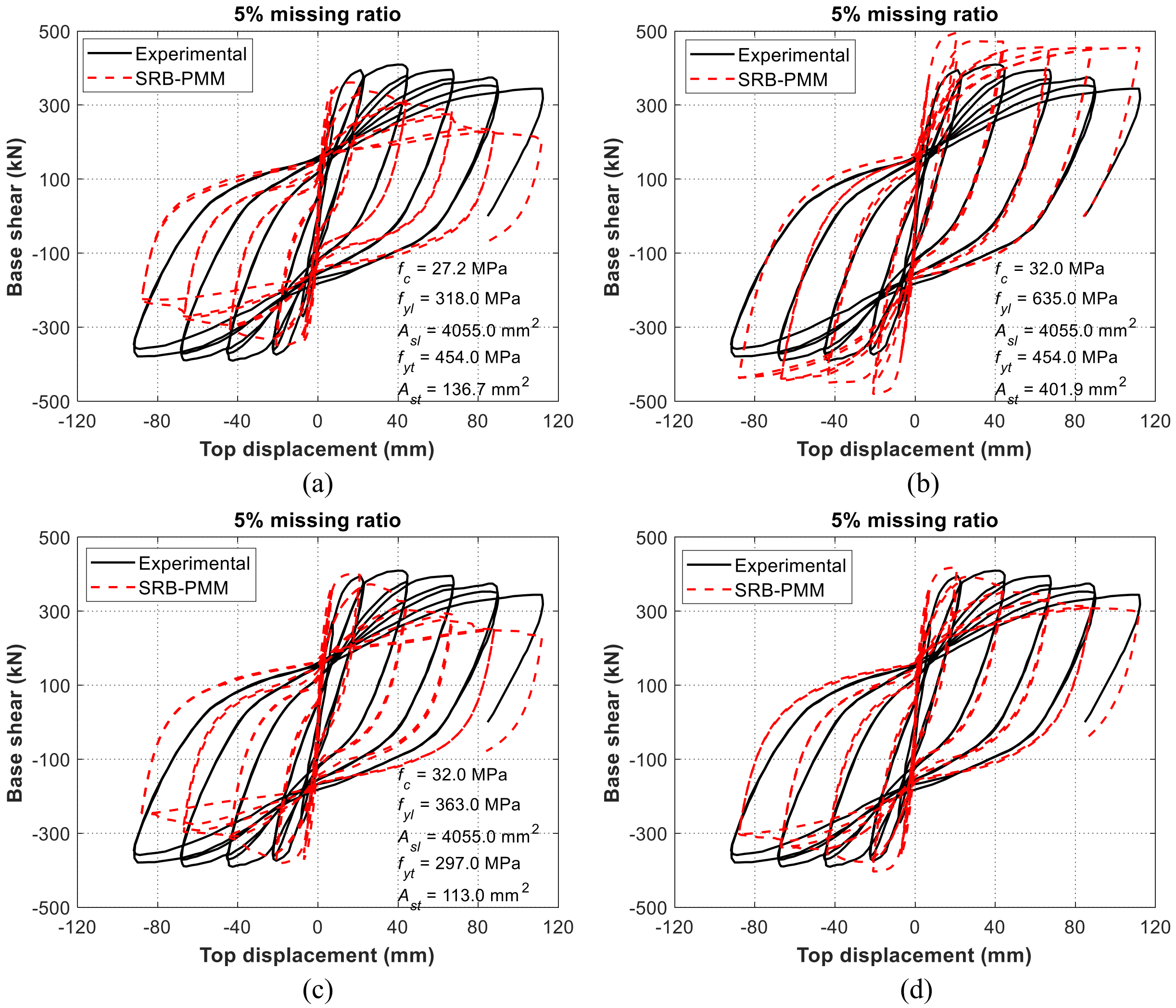


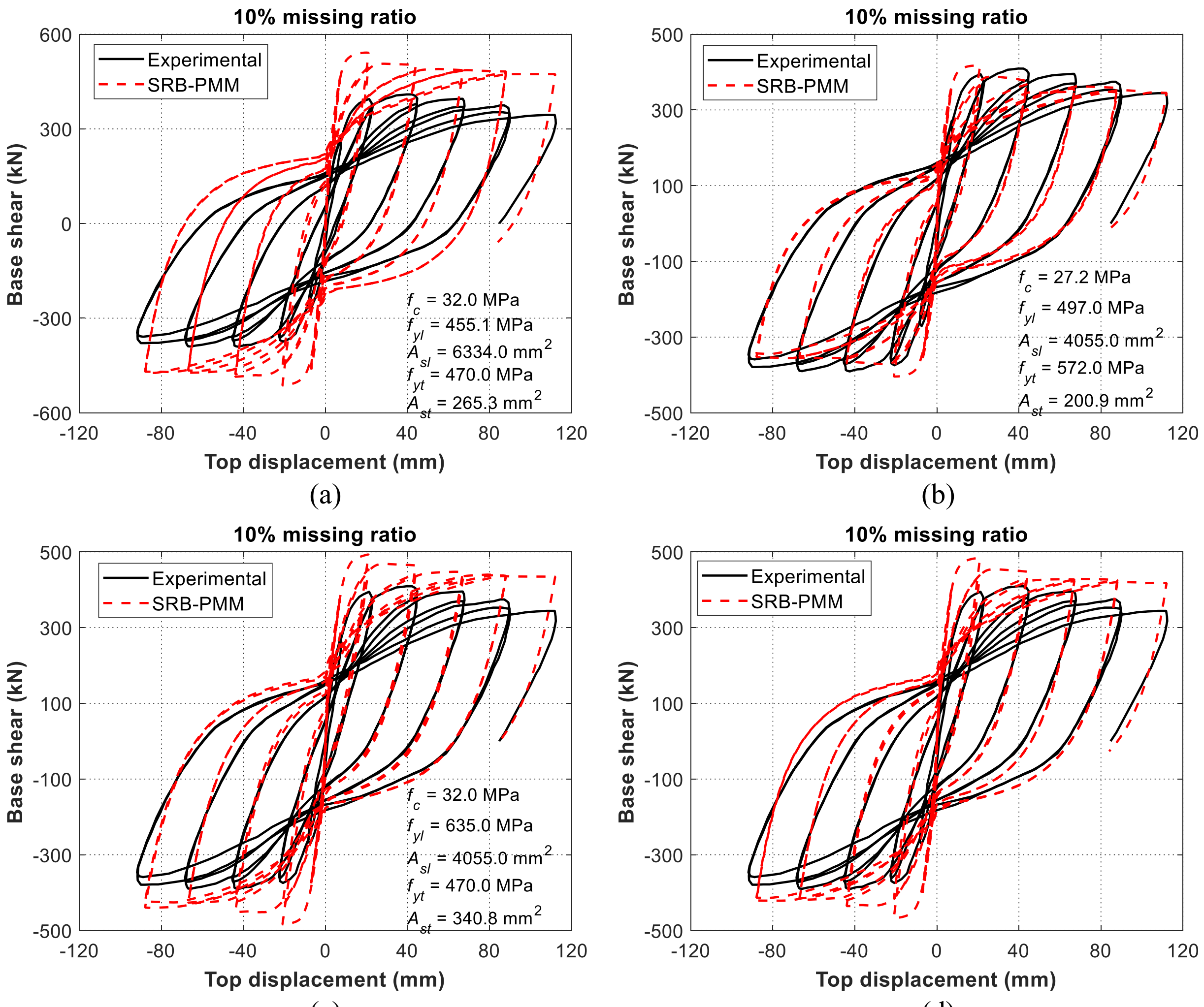
Fig. 3. Seismic analysis result for the sampled RC column missing critical feature information. (a), (b), and (c) are the three results comparison between the experimental and simulated results, and the three simulated results are obtained from the three imputed information presented on the figures using the *SRB-PMM* based on the column data set 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.

estimation can be justified even if the actual seismic performance is unknown. Both Fig. 3(d) and 4(d) account for the uncertainty of missing data, and these results show reasonable estimations. Therefore, this example demonstrates that the proposed *SRB-PMM* method performs well for the incomplete column data sets with 5% and 10% missing data ratios, which in turn illustrates its practical application in post-earthquake structural evaluation subjected to missing data problems.

4.3. *Discussion of results*

Results from these two examples demonstrate that the proposed *SRB-PMM* method is able to generate realistic, valid candidates for imputing the missing values, without risking meaningless imputations as is characteristic of existing, popular imputation approaches. The first example further illustrates that the proposed *SRB-PMM* enhances the generalization performance of the data-driven model for the maximum lateral strength prediction of the RC columns subjected to earthquake loads 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 plausible candidates, which yields the *SRB-PMM-LS-SVMR model,* trained on the imputed data set, having comparable performance to the model formed using the original complete training set (i.e., *Complete-LS-SVMR*).

This validation demonstrates the great potential to obtain the missing feature information (e.g., material properties) for damaged RC columns in the post-earthquake investigation field, which can be used for seismic analyses to assess the safety of damaged RC columns. It is routine that the material strength and reinforcement details can be acquired only when the original column's design information is known. However, it is rather hard to obtain this type of information in the field, and thus, a seismic analysis of the damaged RC column is conventionally impossible with current procedures. However, with the proposed *SRB-PMM* method, it is possible to impute the missing information for the damaged RC columns with an established RC column data set and use those imputed candidates to perform the seismic analysis and assess the safety of the damaged RC column. The results from Example 1 have demonstrated that the imputed candidates are plausible and realistic even if the established RC column data set has a missing data ratio of less than 10%. This fact has been validated by the R^2, RMSE, and MAE values, which are close to those of the *Complete-LS-SVMR* (ground truth) when the missing data ratio is less than 10% (Fig. 2). The R^2, RMSE, and MAE values reflect the generalization performance of the proposed *SRB-PMM-LS-SVMR* developed using the imputed data set, as introduced in Section 3.5. A high R^2 value that approximates 1 and low RMSE and



(a) (b)

(c) (d)

Fig. 4. 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 data set 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.

MAE values that approximate 0 for the proposed *SRB-PMM-LS-SVMR* can demonstrate that the imputed candidates are plausible, and the seismic analysis results obtained based on the imputed candidates are reliable. Nevertheless, this strategy should be employed with caution when the RC column data set has a missing data ratio greater than 10%, where the performance difference between the proposed *SRB-PMM-LS-SVMR* and *Complete-LS-SVMR* is increasingly large as demonstrated by the apparent discrepancy for the R^2, RMSE, and MAE values between them in Fig. 2. The deteriorated performance may be due to the situation where the imputed candidates are far away from the actual missing values, making the imputed observations become outliers. In this case, the seismic analysis result using the imputed candidates should be checked with physical knowledge or expert opinions.

In the post-earthquake investigation field, although the lateral displacement and other visual damages (e.g., concrete cracking and spalling) of damaged RC columns can be measured and detected manually or by advanced computer vision techniques [2–7], the lateral load-carrying capacity corresponding to different damage states (i.e., lateral displacement) cannot be acquired directly. Further, the information regarding the material properties and reinforcement details for damaged RC columns is most likely unavailable or difficult to retrieve in a timely manner, and thus a detailed seismic analysis is not feasible immediately after the earthquake. The second example is designed to address this problem. Since the results from Example 1 have demonstrated that the proposed *SRB-PMM* method can result in the *SRB-PMM-LS-SVMR* model with comparable performance to the *Complete-LS-SVMR* model when the column data set has a missing data ratio less than 10%, the missing data ratio for the second example is restricted to 5% and 10%. Figs. 3 and 4 showed the relation of the estimated lateral strength versus different lateral displacement levels for a damaged RC column, where the estimated lateral strength reflects the lateral load-carrying capacity, and the different lateral displacement levels represent different damage states. Thus, the estimated lateral load-carrying capacity for the damaged RC column can be obtained once the lateral displacement is measured. Both Figures illustrated that the uncertainty of missing data must be incorporated, since a single seismic analysis result may over- or under-estimate the actual lateral load-carrying capacity of damaged RC columns as presented in Fig. 3(a,b,c) and 4(a,c) for the column data sets having 5% and 10% missing data ratios, respectively. The triage team of structural engineers will most likely make an incorrect decision to dismantle lightly damaged RC columns if the actual lateral load-carrying capacity is under-estimated, leading to a waste of resources and money. When the actural load-carrying capacity is over-estimated, an incorrect decision could also be made to keep or retrofit seriously damaged RC columns, posing a substantial threat. However, with the consideration of the uncertainty due to missing data as in the proposed *SRB-PMM* method, the seismic analysis results can reasonably estimate the actual lateral load-carrying capacity of damaged RC columns (Fig. 3(d) and 4(d)). Therefore, the triage team can make reasonable and accurate decisions to either dismantle or retrofit damaged RC columns according to the estimated results that duly incorporate the uncertainty associated with the missing data.

4.4. Discussion of limitations

In the context of post-earthquake safety and structural evaluations for damaged buildings, missing data (e.g., information pertaining to the material properties) are unobserved values that would be meaningful for analysis if observed. This can be especially demonstrated by Example 2, where the seismic analysis for the damaged RC column cannot be performed if the missing values related to material properties (e.g., concrete compressive strength and reinforcement area) are not imputed (filled in). However, it should be noted that the imputed or filled-in values (i.e., added data) generated by the proposed SRB-PMM method should not be considered as the true values but rather as values that are statistically plausible given other observed information. In this sense, the proposed approach should not be regarded as an imputation procedure for recovering the missing values. Instead, the proposed approach can generate the added data that should result in physical predictions that statistically reasonably approximate the actual results or the predictions that would have been obtained by the case of no missing data, as demonstrated in Examples 1 and 2.

Although the proposed SRB-PMM method is used to advance the post-earthquake safety and structural assessment in the context of missing data, it is a general approach and can be applied in any discipline when missing data problems occur. The use of the proposed method requires the reasonable specification of a probability density function (PDF) for each partially observed variable, as introduced in Section 2. The PDF depends on the choice of the conditional regression models, which in turn, rely on the type of target variable. If the target variable is continuous, a normal linear regression model can be selected; if the target variable is binary, a logistic regression model can be used; if the target variable is categorical, a multinomial logistic regression model can be utilized; and finally, if the target variable is a count variable, a Poisson loglinear model can be employed. Once the conditional regression model is chosen according to the target variable type, the PDF can be determined. For example, since the variables in the RC column data set are all continuous, a normal linear regression model is selected for each partially observed variable, and the associated PDFs are thus multivariate normal distributions with different mean vector and covariance matrix as illustrated in Section 2. Therefore, for different types of target variables, the PDFs should be carefully specified when using the proposed approach.

Additionally, the proposed SRB-PMM method requires the incomplete data set that has both fully and partially observed predictors; thus, the missing values will be restricted in the partially observed predictors. The proposed approach will work well with large data sets and provide imputations that possess many characteristics of the complete data set. This is because, when the data set is very large, the number of complete observations related to missing cases will increase. In this sense, the proposed approach can borrow the observed data from the cases that have the fitted values closest to those predicted values for missing cases to fill in the missing data. However, this will also increase the computational cost and, in that sense, may cause difficulty in scaling the approach up. The proposed SRB-PMM method may not work well in situations where the data set is small or the proportion of incomplete observations is high such that no or only few related complete observations could be found. Further, the proposed approach is not appropriate for incomplete data sets where only a small number of predictors are available.

5. Conclusions

This paper proposed a novel multiple imputation (MI) method called sequential regression-based predictive mean matching (*SRB-PMM*) to address missing data problems. The *SRB-PMM* 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 most number of missing values. To validate the usefulness of *SRB-PMM* in advancing post-earthquake safety and structural assessment, two examples are designed and performed based on an RC column data set. The results from the two examples demonstrate the wide-scale capabilities of the proposed approach towards expediting post-earthquake structural evaluations, where all critical structural properties may not be known in the field. As the proposed SRB-PMM method is a multiple imputation (MI) method, the uncertainty due to missing data is also incorporated into the final structural analyses. On the basis of these two examples, the results show that by independently running the method three times, it is sufficient to cover the variation induced by the uncertainty of missing data. Therefore, based on the two examples, it can be concluded that the proposed SRB-PMM method is a useful and effective tool to handle missing data problems in post-earthquake safety and structural assessment.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This material is based in part upon work supported by the National Science Foundation under the grant CMMI #1944301. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] J.A. Goulet, C. Michel, A.D. Kiureghian, Data-driven post-earthquake rapid structural safety assessment, Earthquake Eng. Struct. Dyn. 44 (4) (2015) 549–562.
[2] S. German, I. Brilakis, R. DesRoches, Rapid entropy-based detection and properties measurement of concrete spalling with machine vision for post-earthquake safety assessments, Adv. Eng. Inf. 26 (4) (2012) 846–858.
[3] Z. Zhu, S. German, I. Brilakis, Visual retrieval of concrete crack properties for automated post-earthquake structural safety evaluation, Autom. Constr. 20 (7) (2011) 874–883.
[4] S. German, J.S. Jeon, Z. Zhu, C. Bearman, I. Brilakis, R. DesRoches, L. Lowes, Machine vision-enhanced postearthquake inspection, J. Comput. Civil Eng. 27 (6) (2013) 622–634.
[5] S.G. Paal, J.S. Jeon, I. Brilakis, R. DesRoches, Automated damage index estimation of reinforced concrete columns for post-earthquake evaluations, J. Struct. Eng. 141 (9) (2015) 04014228.
[6] D. Lattanzi, G.R. Miller, M.O. Eberhard, O.S. Haraldsson, Bridge column maximum drift estimation via computer vision, J. Comput. Civil Eng. 30 (4) (2016) 04015051.
[7] C. Koch, S.G. Paal, A. Rashidi, Z. Zhu, M. König, I. Brilakis, Achievements and challenges in machine vision-based inspection of large concrete structures, Adv. Struct. Eng. 17 (3) (2014) 303–318.
[8] D.B. Rubin, Inference and missing data, Biometrika 63 (3) (1976) 581–592.
[9] R.J. Little, D.B. Rubin, Statistical Analysis with Missing Data, John Wiley & Sons, 2019.
[10] J.L. Schafer, M.K. Olsen, Multiple imputation for multivariate missing-data problems: a data analyst's perspective, Multivar. Behav. Res. 33 (4) (1998) 545–571.
[11] L. Uechi, D.J. Galas, N.A. Sakhanenko, Multivariate analysis of data sets with missing values: an information theory-based reliability function, J. Comput. Biol. 26 (2) (2019) 152–171.
[12] G.E. Batista, M.C. Monard, An analysis of four missing data treatment methods for supervised learning, Appl. Artif. Intell. 17 (5–6) (2003) 519–533.
[13] L. Beretta, A. Santaniello, Nearest neighbor imputation algorithms: a critical evaluation, BMC Med. Inf. Decis. Making 16 (3) (2016) 74.
[14] G. Tutz, S. Ramzan, Improved methods for the imputation of missing data by nearest neighbor methods, Comput. Stat. Data Anal. 90 (2015) 84–99.

[15] D.B. Rubin, Multiple Imputation for Nonresponse in Surveys, Vol. 81, John Wiley & Sons, 2004.
[16] D.B. Rubin, Multiple imputation after 18+ years, J. Am. Stat. Assoc. 91 (434) (1996) 473–489.
[17] J.L. Schafer, Analysis of Incomplete Multivariate Data, Chapman and Hall/CRC, 1997.
[18] J.L. Schafer, R.M. Yucel, Computational strategies for multivariate linear mixed-effects models with missing values, J. Comput. Graphical Stat. 11 (2) (2002) 437–457.
[19] S. Van Buuren, Multiple imputation of discrete and continuous data by fully conditional specification, Stat. Methods Med. Res. 16 (3) (2007) 219–242.
[20] D. Heckerman, D.M. Chickering, C. Meek, R. Rounthwaite, C. Kadie, Dependency networks for inference, collaborative filtering, and data visualization, J. Mach. Learn. Res. 1 (Oct) (2000) 49–75.
[21] S. Van Buuren, Flexible Imputation of Missing Data, CRC Press, 2018.
[22] S.V. Buuren, K. Groothuis-Oudshoorn, mice: Multivariate imputation by chained equations in R, J. Stat. Softw. (2010) 1–68.
[23] J.W. Bartlett, S.R. Seaman, I.R. White, J.R. Carpenter, Alzheimer's Disease Neuroimaging Initiative, Multiple imputation of covariates by fully conditional specification: accommodating the substantive model, Stat. Methods Med. Res., 24 (4) (2015) 462–487.
[24] A. Gelman, Parameterization and Bayesian modeling, J. Am. Stat. Assoc. 99 (466) (2004) 537–545.
[25] T.E. Raghunathan, J.M. Lepkowski, J. Van Hoewyk, P. Solenberger, A multivariate technique for multiply imputing missing values using a sequence of regression models, Survey Methodol. 27 (1) (2001) 85–96.
[26] P.D. Hoff, A First Course in Bayesian Statistical Methods, Vol. 580, Springer, New York, 2009.
[27] Applied Technology Council, Seismic evaluation and retrofit of concrete buildings. (ATC-40). ATC (Applied Technology Council), Redwood City, 1996.
[28] FEMA, FEMA 306: Evaluation of Earthquake Damaged Concrete And Masonry Wall Buildings – Basic procedures manual. Federal Emergency Management Agency, Washington D.C., 1998.
[29] A. Zellner, 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, 1986, pp. 233–243.
[30] N. Schenker, J.M. Taylor, Partially parametric techniques for multiple imputation, Comput. Stat. Data Anal. 22 (4) (1996) 425–446.
[31] R.J. Little, Missing-data adjustments in large surveys, J. Bus. Econ. Stat. 6 (3) (1988) 287–296.
[32] T.P. Morris, I.R. White, P. Royston, Tuning multiple imputation by predictive mean matching and local residual draws, BMC Med. Res. Method. 14 (1) (2014) 75.
[33] D.B. Rubin, Statistical matching using file concatenation with adjusted weights and multiple imputations, J. Bus. Econ. Stat. 4 (1) (1986) 87–94.
[34] G. James, D. Witten, T. Hastie, R. Tibshirani, An Introduction to Statistical Learning, springer, New York, 2013.
[35] K.J. Elwood, J.P. Moehle, Drift capacity of reinforced concrete columns with light transverse reinforcement, Earthquake Spectra 21 (1) (2005) 71–89.
[36] H. Luo, S.G. Paal, Machine learning–based backbone curve model of reinforced concrete columns subjected to cyclic loading reversals, J. Comput. Civil Eng. 32 (5) (2018) 04018042.
[37] C.A. Leke, T. Marwala, Deep Learning and Missing Data in Engineering Systems, Springer, London, 2019.
[38] J. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, J. Vandewalle, Least Squares Support Vector Machines, World Scientific, 2002.
[39] J.A.K. Suykens, L. Lukas, P. Van Dooren, B. De Moor, J. Vandewalle, Least squares support vector machine classifiers: a large scale algorithm, in: European Conference on Circuit Theory and Design, ECCTD, Vol. 99, Citeseer, 1999, pp. 839–842.
[40] K. De Brabanter, J. Suykens, B. De Moor, Nonparametric regression via StatLSSVM, J. Stat. Softw. 55 (2) (2013) 1–22.
[41] H. Luo, S.G. Paal, A locally weighted machine learning model for generalized prediction of drift capacity in seismic vulnerability assessments, Comput.-Aided Civ. Infrastruct. Eng. 34 (11) (2019) 935–950.
[42] H. Luo, S.G. Paal, Reducing the effect of sample bias for small datasets with double-weighted support vector transfer regression, Computer-Aided Civil Infrastruct. Eng., Wiley. (2020), https://doi.org/10.1111/mice.12617.
[43] D.T. Vu, N.D. Hoang, Punching shear capacity estimation of FRP-reinforced concrete slabs using a hybrid machine learning approach, Struct. Infrastruct. Eng. 12 (9) (2016) 1153–1161.
[44] M.Y. Cheng, N.D. Hoang, Estimating construction duration of diaphragm wall using firefly-tuned least squares support vector machine, Neural Comput. Appl. 30 (8) (2018) 2489–2497.
[45] N.D. Hoang, Image processing based automatic recognition of asphalt pavement patch using a metaheuristic optimized machine learning approach, Adv. Eng. Inf. 40 (2019) 110–120.
[46] N.D. Hoang, X.L. Tran, H. Nguyen, Predicting ultimate bond strength of corroded reinforcement and surrounding concrete using a metaheuristic optimized least squares support vector regression model, Neural Comput. Appl. (2019) 1–21.
[47] J.S. Chou, N.T. Ngo, A.D. Pham, Shear strength prediction in reinforced concrete deep beams using nature-inspired metaheuristic support vector regression, J. Comput. Civil Eng. 30 (1) (2015) 04015002.
[48] S. Mazzoni, F. McKenna, M.H. Scott, G.L. Fenves, OpenSees command language manual, Pacific Earthquake Engineering Research (PEER) Center 264 (2006).
[49] P. Kang, Locally linear reconstruction based missing value imputation for supervised learning, Neurocomputing 118 (2013) 65–78.
[50] H. Tanaka, R. Park, Effect of Lateral Confining Reinforcement on the Ductile Behavior of Reinforced Concrete Columns, Report 90-2, Department of Civil Engineering, University of Canterbury, June 1990, p. 458.