Binary Building Attribute Imputation, Evaluation, and Comparison Approaches for Hurricane Damage Data Sets

Carol C. Massarra, Ph.D., A.M.ASCE[1]; Carol J. Friedland, Ph.D., P.E., M.ASCE[2]; Brian D. Marx, Ph.D.[3]; and J. Casey Dietrich, Ph.D., A.M.ASCE[4]

Abstract: Missing building attributes are problematic for development of data-based fragility models. Relative to other disciplines, the application of imputation techniques is limited in the field of engineering. Current imputation techniques to replace missing building attributes lack evaluations of imputation model performance, which ensure accuracy and validity of the imputed data. This paper presents two imputation approaches, along with imputation diagnostic and comparison approaches, for binary building attribute data with missing observations. Predictive mean matching (PMM) and multiple imputation (MI) are used to impute foundation type and number of stories attributes. The diagnostic approach, based on the logistic regression goodness-of-fit test, is used to evaluate the imputation model fit. The comparison approach, based on the percentage of correctly imputed observations, is used to evaluate the imputation model performance. A data set of single-family homes damaged by the 2005 Hurricane Katrina is used to demonstrate implementation of the methodology. Based on the comparison approach, PMM models showed 9% and 2% greater accuracy than MI models in imputing foundation type and number of stories, respectively. **DOI: 10.1061/(ASCE)CF.1943-5509.0001433.** *© 2020 American Society of Civil Engineers.*

Author keywords: Imputation techniques; Binary missing data; Diagnostic approach; Comparison approach.

Introduction

Missing data come hand-in-hand with data themselves and occur in virtually all areas of research, often resulting in a nontrivial effect on data-based conclusions (Ferrari et al. 2011). Yet, relative to other disciplines, the application of techniques to replace missing data is not as common in the field of fragility modeling. Traditionally, fragility models relate hazard intensities (e.g., wind speed or inundation depth) and building attributes (e.g., foundation type or number of stories) to physical damage and are used for either assessment or prediction of damage. The model accuracy is highly sensitive to the quality of the data (e.g., hazard data or building attributes) used to develop the model. Although hazard data can be simulated using advanced modeling such as Simulating Waves Nearshore and Advanced Circulation (SWAN + ADCIRC), building attributes are sometimes difficult to collect due to the lack of detailed tax records, exclusion of old houses from tax recodes, and severity of the storm. With the lack of prestorm and poststorm building attributes, missing building attributes become a very common issue, which may have a significant effect on the conclusion drawn from the data set. Although missing data are a common issue, approaches for accommodating missing data are limited to complete case analysis, where the missing observations are excluded from the data set and only the complete case is used for further analysis.

When complete case analysis is used, fragility models are developed with data that do not consider the full range of buildings. This causes the loss of valuable information, results in reduction of a sample size, and leads to biased estimates of model coefficients. A better and valid approach to accommodate missing data is imputation, which simply put is a process of replacing missing data with substituted values by estimating the association between all of the variables in the model, finally filling in (imputing) reasonable guesses for the missing values. Imputation studies are a relatively more active research area in medical and social science fields than in the engineering field and are more powerful than complete case analysis in imputing missing data. Because imputation techniques preserve sample size, they are generally viewed as the preferred analytical approach over complete case analysis. For this reason, many imputation techniques are developed to make gap end. Imputation techniques such as imputation with mean, median, and mode are simple techniques for imputing numerical and categorical variables with missing observations, but like complete case analysis, the techniques generally introduce bias and ignore relationship with other variables in the data set.

Model-based imputation techniques such as multiple imputation (MI) and predictive mean matching (PMM) are more powerful techniques than complete case analysis and simple imputations techniques. MI is practical and widely applicable technique, and it has a variety of usages for continuous and categorical data with monotone and arbitrary missing patterns. The technique is better suited to highlight the uncertainties about the missing value estimates by creating several different imputed data sets and appropriately combining results obtained from each of them.

[1] Assistant Professor, Dept. of Construction Management, East Carolina Univ., 331 Rawl Building, Greenville, NC 27858 (corresponding author). Email: massarrac19@ecu.edu

[2] Associate Professor, Bert S. Turner Dept. of Construction Management, Louisiana State Univ., Baton Rouge, LA 70803. ORCID: https://orcid.org/0000-0003-0443-5266. Email: friedland@lsu.edu

[3] Professor, Dept. of Experimental Statistics, Louisiana State Univ., Baton Rouge, LA 70803. Email: bmarx@lsu.edu

[4] Associate Professor, Dept. of Civil, Construction, and Environmental Engineering, North Carolina State Univ., Raleigh, NC 27695. ORCID: https://orcid.org/0000-0001-5294-2874. Email: jcdietrich@ncsu.edu

Note. This manuscript was submitted on April 30, 2019; approved on November 6, 2019; published online on March 27, 2020. Discussion period open until August 27, 2020; separate discussions must be submitted for individual papers. This paper is part of the ***Journal of Performance of Constructed Facilities***, © ASCE, ISSN 0887-3828.

Valid inferences are obtained because the technique averages over the distribution of the missing data given the observed data (Little and Rubin 2014). On the other hand, PMM is an easy-to-use imputation technique for continuous and categorical data and is based on other observed observations in the data set, so the imputed values are more realistic. Imputations outside the observed data range will not occur, thereby avoiding problems with meaningless imputations resulting from extrapolation. Because the technique relies on other observed in the data set, there is no need to define the distribution of missing values. Consequently, predictive mean matching is less vulnerable to model misspecification (Little and Rubin 2014).

In the field of data-based fragility modeling, implementation of imputation techniques is very limited. Imputation techniques are particularly important for data collected after natural hazard events because building attribute data may not be fully observed due to severity of damage. Pita et al. (2011) used Bayesian belief networks (BBN) and classification and regression trees (CART) machine-learning methods to classify roof shape as a function of wall type, year built, roof cover type, number of stories, and building value. Performance of the BBN and CART were then evaluated using cross validation. Macabuag et al. (2016) used MI techniques to impute building material attributes as a function of footprint area, damage state, building use, and inundation depth. However, performance of the imputation models was not evaluated.

Although results of the previous two studies aided in development of risk and fragility models, building attributes were either imputed based on machine-learning techniques rather than imputation techniques (Pita et al. 2011), or based on imputation techniques without evaluation of the imputation model performance (Macabuag et al. 2016). In other disciplines, it has been shown that evaluation of imputation model performance is essential to ensure validity and accuracy of the imputed data (Bernhardt 2018; Cabras et al. 2011; Gelman et al. 2005; Nguyen et al. 2017) and to provide an approach for model comparison and selection among various imputation models (Fay 1996; Meng 1994).

When an individual imputation model is evaluated, model diagnostics are performed numerically or graphically to determine how well the model estimates missing values. When various imputation models are evaluated, model comparison is performed numerically to determine the model or models with the highest imputation accuracy. Current numerical diagnostic approaches are limited to imputation models for continuous variables (e.g., Abayomi et al. 2008; Farhan and Fwa 2014; Stuart et al. 2009; Van Buuren 2012; White et al. 2011; Zhu et al. 2009), whereas comparison approaches are limited to evaluation of the performance of statistical models fit on the imputed data set rather than imputation models themselves (e.g., Akande et al. 2017; Collins et al. 2001; Raghunathan et al. 2001). An approach for implementing diagnostic and comparison approaches for evaluating the fit and accuracy of imputation models for categorical variables (e.g., building attributes) is essential to improve the quality of postevent data, and therefore, improve the prediction of the data-based fragility model.

This paper presents an imputation approach, along with imputation diagnostic and comparison approaches, for binary building attribute data with missing observations. Binary building attributes with missing observations are imputed by fitting PMM and MI logistic regression–based imputation models on the complete case data set with the selected building attribute as the response variable and observed numerical hazard and environmental attribute variables as explanatory variables. The diagnostic approach is used to numerically diagnose the fit of each individual logistic regression imputation model and to determine the subset of the variables to be used in the analysis. The model choice was based on the following three criteria: (1) the logistic regression goodness-of-fit test, (2) the significance of the observed numerical variables, and (3) the independence of the observed numerical variables.

The comparison approach is used to numerically evaluate the performance of the imputation models themselves. More specifically, observations from the complete data sets of every building attribute are randomly deleted to accomplish a missingness percentage equal to that in the original data set. The deleted observations are then imputed based on the diagnosed PMM and MI imputation models. The percentage of correctly classified observations, expressed as the cross-classification rate (CCR), is calculated by comparing prior and postdeletion values. For every building attribute, the imputation model with the highest CCR and lowest class error (CE) is chosen as the final imputation model.

A data set of single-family homes damaged by 2005 Hurricane Katrina in coastal Mississippi is used to demonstrate application of the methodology. Data for foundation type and number of stories were missing for about 45% of the buildings destroyed by Katrina. The observed numerical hazard and environmental variables, defined as maximum 3-s wind speed, maximum significant wave height, maximum water depth, maximum water speed, and base flood elevation, are used to impute missing foundation type and number of stories data for slab and elevated foundations of 1- and 2-story homes. Hazard intensities and the base flood elevation were obtained from joint SWAN + ADCIRC and FEMA Flood Map (FEMA 2020; Dietrich et al. 2012) Service Center flood insurance rate maps (FIRMs), respectively.

The contributions of this paper are approaches to impute binary building attributes based on imputation techniques rather than statistical techniques, diagnose the fit of individual imputation models, and evaluate the performance of the imputation models rather than the performance of statistical models fit on the imputed data set. A major issue for damage modelers and data collectors is the implementation of techniques to impute building attributes with missing observations because of the lack of knowledge transfer between the disciplines of statistics and engineering. The developed approaches provide damage modelers and data collectors with the knowledge needed to impute binary building attributes and to evaluate the fit and performance of the imputation models themselves. Implementation of imputation techniques improves quality of postevent data because building attribute data may not be fully observed due to the severity of damage or lack of data; meanwhile, implementation of evaluation approaches ensures validity of the imputation models, which is an improvement over current building attribute imputation practices that lack evaluation of imputation model performance.

Missing Data Imputation: Notation, Models, and Details

For a data set with sample size N, G binary (i.e., Levels 0 and 1) explanatory variables $(X_1, X_2, \ldots, X_G)$ with missing observations, and F continuous explanatory variables $(Z_1, Z_2, \ldots, Z_F)$, two model-based imputation techniques $T = (T_{\mathrm{PMM}}, T_{\mathrm{MI}})$, defined as PMM and MI, are applied to impute each X_g with missing observation. For each variable X_g with missingness, sample sizes for the complete and missing case subdata sets are $N_{g,cc}$ and $N_{g,\mathrm{miss}}$, respectively, where $N_g = N_{g,cc} + N_{g,\mathrm{miss}}$. Variables of these subdata sets are defined as $X_{g,cc}$, $Z_{f,cc}$, $X_{g,\mathrm{miss}}$, and $Z_{f,\mathrm{miss}}$, where $g = 1, 2, \ldots, G$, $f = 1, 2, \ldots, F$, the subscript cc denotes observations of fully observed X_g, and the subscript miss denotes observations with missing X_g values. The missing mechanism of the data set is assumed to be missing at random (MAR), meaning that the probability of missing X_g values depends only on the observed variables in the data set. Based on T, a set of two logistic regression-based imputation models $LR_g = (LR_{g,PMM}, LR_{g,MI})$ is fitted on the $N_{g,cc}$ complete cases. The response variable of LR_g is X_g, and the explanatory variables are the continuous explanatory variables $(Z_1, Z_2, \ldots, Z_F)$.

Predictive Mean-Matching Imputation Techniques

For the predictive mean-matching imputation technique, T_{PMM}, the imputation procedure imputes a missing value by matching its estimated predictive probability to the nearest complete case estimated predictive probability. A logistic regression model $LR_{g,PMM}$ is fitted on the complete cases for each X_G with missing observations

$$\ln\left(\frac{P(X_{g,cc}=1)}{1-P(X_{g,cc}=1)}\right) = \alpha_{g0} + \sum_{f=1}^{F} \alpha_{gf} Z_{f,cc} \quad \text{for } g = 1, 2, \ldots, G \tag{1}$$

where $P(X_{g,cc} = 1)$ = probability of $X_{g,cc}$ being in Level 1; α_{g0} = model intercept; and α_{gf} = model coefficients.

For each complete case observation $x_{g,cc}$ of variable X_G, the estimated predictive probability that $x_{g,cc}$ with explanatory variables $Z_{f,cc}$ belongs to Level 1 is estimated as follows:

$$P(x_{g,cc}=1) = \frac{\exp(\alpha_{g0} + \sum_{f=1}^{F} \alpha_{gf} Z_{f,cc})}{1+\exp(\alpha_{g0} + \sum_{f=1}^{F} \alpha_{gf} Z_{f,cc})} \tag{2}$$

For each missing observation $x_{g,\mathrm{miss}}$ of variable X_G, the estimated predictive probability that $x_{g,\mathrm{miss}}$ with explanatory variable $Z_{f,cc}$ belongs to Level 1 is estimated as follows:

$$P(x_{g,\mathrm{miss}}=1) = \frac{\exp(\alpha_{g0} + \sum_{f=1}^{F} \alpha_{gf} Z_{f,cc})}{1+\exp(\alpha_{g0} + \sum_{f=1}^{F} \alpha_{gf} Z_{f,cc})} \tag{3}$$

The absolute difference $|D_g|$ between $P(x_{g,\mathrm{miss}} = 1)$ and every $P(x_{g,cc} = 1)$ is calculated and used to construct a distance matrix $\mathbf{Q}$ with number of rows representing the number of complete cases and number of columns representing the number of missing cases. For every column in $\mathbf{Q}$, $x_{g,\mathrm{miss}}$ is set equal to the $x_{g,cc}$ value corresponding to the row with the smallest $|D_g|$ value. For rows with equal $|D_g|$ values, $x_{g,\mathrm{miss}}$ is selected as the mode of the corresponding $x_{g,cc}$ values unless the variable levels are equally represented, in which case, $x_{g,\mathrm{miss}}$ is selected at random.

Multiple Imputation Techniques

The application of multiple imputation techniques, T_{MI}, is dependent on the missingness pattern, i.e., arbitrary or monotone. For the arbitrary missingness pattern, MI using fully conditional specification (FCS) is used, whereas MI using logistic regression is used for the monotone missingness pattern. A data set with variables $X_1, X_2, \ldots, X_g$ has a monotone missingness pattern if variable X_j and all previous variables X_k, $k < j$, are observed and X_{j+1} and all subsequent variables X_m, $m > j$, are missing for observation i; otherwise, the data set has an arbitrary missingness pattern. The imputation procedure for T_{MI} generates M_g imputations by performing draws from the predictive posterior distribution(s) of $X_{g,cc}$ conditioned on $Z_{f,cc}$. The number of required imputations M_g depends on the fraction of missing data λ_g (Rubin 1978) and is determined by the relative efficiency index $RE = [1 + (\lambda_g/M_g)]^{-1}$, where Rubin (1978) precalculated values of RE based on fractions of missing data (λ) and number of imputations (m) to determine the number of required imputations. A value of M_g is chosen so that RE_g is greater than 90%. The imputation algorithm sequentially iterates through the variables to impute the missing values using $LR_{g,MI}$ fitted on the complete cases Eq. (1).

For each X_g, model coefficients are randomly drawn from a multivariate normal distribution with mean and variance equal to the model coefficients of Eq. (1). This procedure results in M_g logistic regression models, M_g model intercepts, and $M_g \times F$ model coefficients, and hence, M_g imputed data sets. Intercepts and coefficients for each model are different and collectively defined as α_g. Statistical analysis is then performed on the M_g imputed data sets resulting in M_g statistical model coefficients. Values of the M_g coefficients are pooled into a final result by averaging the coefficient values. Although generating M_g imputed data sets ensures variability in the imputed data set without biasing estimates, pooling the statistical model coefficients results in one set of final statistical model coefficients rather than one set of final imputation model coefficients.

For this study, pooling is performed on the M_g imputation model coefficients rather than on the M_g statistical model coefficients. This procedure results in one final imputed data set while maintaining variability in the imputed data set. For each X_g, imputation models are fitted M_g times based on Eq. (1). The intercepts and coefficients obtained from Eq. (1) are pooled by averaging over the M_g imputations to calculate the average estimated intercepts and model coefficients $\overline{\alpha_g}$, as $\overline{\alpha_g} = (1/M_g)\sum_{m_g=1}^{M_g} \alpha_{mg}$ where α_{mg} is the estimate of α_g in the $m_{g\mathrm{th}}$ model. Thus, the imputation model $LR_{g,MI}$ is redefined as follows:

$$\ln\left(\frac{P(X_{g,cc}=1)}{1-P(X_{g,cc}=1)}\right) = \overline{\alpha_{g0}} + \sum_{f=1}^{F} \overline{\alpha_{gf}}\, Z_{f,cc} \quad \text{for } g = 1, 2, \ldots, G \tag{4}$$

The model defined in Eq. (4) is used as the MI imputation model rather than that defined in Eq. (1). The estimated predictive probability that an individual missing observation $x_{g,\mathrm{miss}}$ belongs to Level 1 is calculated

$$P(x_{g,\mathrm{miss}}=1) = \frac{\exp(\overline{\alpha_{g0}} + \sum_{f=1}^{F} \overline{\alpha_{gf}} Z_{f,cc})}{1+\exp(\overline{\alpha_{g0}} + \sum_{f=1}^{F} \overline{\alpha_{gf}} Z_{f,cc})} \tag{5}$$

If $P(x_{g,\mathrm{miss}} = 1)$ is greater than 0.5, the missing observation is imputed as Level 1; otherwise, it is imputed as Level 0.

Imputation Model Diagnostic and Comparison Approaches

As with any model-based procedure, the fit of the model should be checked using goodness-of-fit tests (Abayomi et al. 2008), and the subset of the variables to be used in the analysis should be determined (Collins et al. 2001). The model diagnostic approach is used to evaluate the fit of the imputation models and to determine the subset of the variables, based on three criteria defined as follows:

- Satisfaction of variable inflation factor (*VIF*): the *VIF* for Z_f is given as $VIF_f = 1/(1 - R_f^2)$, where R_f^2 is the coefficient of determination for a multiple regression model, considering Z_f is the dependent variable and the remaining Z_f variables are the independent variables. A VIF_f greater than 10 indicates that Z_f

is almost a perfect linear combination of other explanatory variables (i.e., multicollinearity); therefore, the standard errors of the model coefficients will be inflated. Any correlated variables ($Z_{f,\mathrm{corr}}$) are not included simultaneously within PMM and MI imputation models.

- Satisfaction of model requirements (goodness of fit): the Hosmer and Lemeshow test is used to assess goodness of fit based on the chi-square test. Any imputation model with chi-square p-value <0.05 is rejected.
- Statistical significance of model parameters: at least one explanatory variable must be significant or the imputation model is rejected.

For each X_g and each imputation technique, logistic regression models are fitted on every combination of Z_f, resulting in two sets of $(2^{Z_f} - 1) - D = K$ imputation models [e.g., $LR_{gMI} = (LR_{gMI,1}, \ldots, LR_{gMI,K})$] where each LR_{gMI} has M_g imputations. D is the number of models with $Z_{f,\mathrm{corr}}$. The three criteria defined previously are used to evaluate the $2K$ models. Models satisfying the three criteria are further evaluated based on the following comparison approach.

From the $N_{g,cc}$ complete cases, missing values are randomly generated by deleting observations of $X_{g,cc}$ so that the percentage of missingness for $X_{g,cc}$ equals that of X_g in the original data set. The deletion procedure results in G sample data sets with sizes N_{g_s}, and variables X_{g_s} and Z_{f_s} with complete and deleted cases defined as $X_{g_s,cc}$ and $X_{g_s,\mathrm{mis}}$, respectively. Deleted observations $x_{g_s,\mathrm{miss}}$ are imputed using the imputation models defined for T_{PMM} [Eq. (1)] and T_{MI} [Eq. (4)], respectively. For every LR_g that satisfies the three criteria, an error matrix is constructed, where the sum of all frequencies in the matrix is N_{g_s}. Rows (j) of the matrix represent the frequency of the observed class levels for $X_{g_s,\mathrm{miss}}$ prior to deletion, and columns (e) represent the frequency of the imputed class levels. The percentage of correctly imputed values, expressed as the cross-classification rate (CCR_{LR_g}), is calculated as follows:

$$CCR_{LR_g} = \frac{\sum_{j=1}^{2} w_{\mathrm{LR}_g,jj}}{\sum_{j=1}^{2}\sum_{e=1}^{2} w_{\mathrm{LR}_g,je}} \tag{6}$$

where $w_{\mathrm{LR}_g,jj}$ = number of correctly classified observations found along the diagonal of the error matrix; and the denominator is equal to N_{gs}. The percentage of each misclassified class, expressed as CE, is calculated as $CE_{LR_g} = 1 - w_{\mathrm{LR}_g,jj}/\sum_{e=1}^{2} w_{\mathrm{LR}_g,je}$. Balance between CCR and CE is used to choose the final imputation model, where models with high CCR values but high CE values are considered less reasonable models for imputing binary variables with missing observations.

Case Study: Hurricane Katrina

A data set containing observations describing hazard intensities and building attributes for $N = 866$ single-family homes in the three counties of coastal Mississippi (Hancock, Harrison, and Jackson) that border the Gulf of Mexico is used for the application of the methodology. These homes ranged in damage from no damage/very minor damage to collapse (Massarra et al. 2019). The continuous variables (Z_F) are maximum 3-s gust wind speed ($U_{3,\max}$), maximum significant wave height ($H_{S,\max}$), maximum water depth ($D_{\max}$), maximum water speed ($U_{\max}$), and base flood elevations (X_{BFE}). All hazard intensities represent the maximum values of the time series obtained from the tightly coupled SWAN + ADCIRC models (Dietrich et al. 2012) after 2005 Hurricane Katrina. SWAN represents the wave field as a phase-averaged spectrum (Booij et al. 1999). SWAN was extended in its functionality (Zijlema 2010) and then coupled tightly with ADCIRC, which solves modified forms of the shallow-water equations for the evolution of the total water depth $H = h + \zeta$, where h is the local bathymetry and ζ is the free-surface elevation relative to the geoid, and the depth-averaged current velocities U and V (Luettich and Westerink 2004; Westerink et al. 2008). These models are coupled tightly so information is passed through local memory, efficient on high-performance computing systems, and validated for hurricane wave and flooding applications along the US Atlantic and Gulf coastlines (Dietrich et al. 2012). SWAN+ADCIRC uses unstructured mesh with triangular finite elements, which can vary in size, ranging from kilometers in open water, to hundreds of meters near the coastline and through the floodplains, and to tens of meters in the small-scale natural and artificial channels that convey surge into the inland region.

The model results in this study were computed on the SL16 mesh, which was developed and validated for the devastating Gulf hurricanes of 2005 and 2008 (Dietrich et al. 2012). Data for variable X_{BFE} were obtained from the FEMA Flood Map Service Center flood insurance rate maps (FIRM$_S$) for Hancock (1983, 1987, and 1992), Harrison, (1980, 1983, 1984, 1988, and 2002), and Jackson (1983, 1987, and 1992) counties, respectively. The flood maps were georeferenced in ArcGIS (version 10.1), and X_{BFE} values were recorded at building footprint locations. The binary variables with missing data (X_G) are foundation type (FT) and number of stories (NS), where the two levels are defined as (slab, elevated) and (1-story, 2-story). Because of the differences in hazard intensities across counties, it is not optimal to use a single county's imputation model to impute FT and NS in other counties because this may result in a low accuracy of the imputation model. Therefore, imputation models were fitted on a county scale. Table 1 describes the explanatory variables Z_f and the response variable X_g of the imputation models.

Frequency and percentage of observed (η) or missingness (λ) for X_{FT} and X_{NS} are given in Table 2. The data were evaluated to determine the missingness patterns, which were found to be arbitrary

Table 1. Explanatory and response variables used to construct the imputation models in Hancock, Harrison, and Jackson Counties

			Range (continuous) or levels (binary)		
Variable	Symbol	Description	Hancock	Harrison	Jackson
Z_f	$U_{3,\max}$	Maximum 3-s gust wind speed (m/s)	48.57–67.99	55.32–67.42	47.62–62.54
	$H_{S,\max}$	Maximum significant wave height (m)	0.3–3.22	0–2.22	0–1.84
	$D_{\max}$	Maximum water depth above local ground level (m)	0.94–7.94	0–5.96	0–5.18
	$U_{\max}$	Maximum water speed (m/s)	0.18–2.8	0–1.06	0–1.45
	X_{BFE}	Base flood elevation (m)	0.35–5.23	0.34–4	0.32–4.31
X_g	X_{FT}	Foundation type	Slab (0), elevated (1)	Slab (0), elevated (1)	Slab (0), elevated (1)
	X_{NS}	Number of stories	1-story (1), 2-story (0)	1-story (1), 2-story (0)	1-story (1), 2-story (0)

Table 2. Frequency and percentages of observed (η) or missingness (λ) for X_{FT} and X_{NS} in Hancock, Harrison, and Jackson Counties

	Frequency (η or λ)			
$X_{\mathrm{FT}}/X_{\mathrm{NS}}$	Hancock	Harrison	Jackson	Total
Slab	39 (21%)	86 (23%)	151 (50%)	276
Elevated floor	56 (30%)	130 (34%)	40 (13%)	266
Missing	89 (48%)	163 (43%)	112 (37%)	364
1-story	26 (14%)	160 (42%)	187 (62%)	373
2-story	15 (8%)	39 (10%)	45 (15%)	99
Missing	143 (78%)	180 (47%)	71 (23%)	394

for the three data sets in the three counties of the study area. Using a relative efficiency index RE > 90%, the number of required imputations M_g for T_{MI} for X_{FT} and X_{NS} was calculated as 10 in Hancock and Harrison and as 5 in Jackson. Correlation between $H_{S,\max}$ and $D_{\max}$ was found to be high, which results in $VIF_{H_{S,\max}}$ and $VIF_{D_{\max}} > 10$. Therefore, $H_{S,\max}$ and $D_{\max}$ were not included simultaneously in the same imputation models.

Case Study Imputation Model Diagnostics

For each imputation technique and each county, combinations of explanatory variables were used to fit the logistic regression models (LR), resulting in 138 models. These models are described with $H_{S,\max}$, $D_{\max}$, $U_{\max}$, and X_{BFE} variables and are tested to satisfy the three diagnostic criteria. Among the 138 models, 33 LR models described in Table 3 satisfied the three diagnostic criteria and were used to impute X_{FT} and X_{NS}. Cells with a letter X indicate the variables that are included in each LR model, where K represents the model number.

From the complete case subdata sets of X_{FT_s} and X_{NS_s} in each county with sample sizes $N_{\mathrm{FT,cc}}$ equal to (95, 216, 191) and $N_{\mathrm{NS,cc}}$ equal to (41, 199, 232), three sample data sets with sample sizes N_{g_s} were constructed by randomly deleting observations. The observations are deleted so that N_{FT_s} was equal to (46, 93, 71) and N_{NS_s} was equal to (32, 94, 53), which represent equivalent missingness percentages (λ) in each county defined in Table 2. The procedure was repeated 1,000 times, and the mean error was reported.

The frequency and percentage of observed and missing (λ) cases for X_{FT_s} and X_{NS_s} in the three subdata sets are given in Table 4. The X_{g_s} deleted observations for X_{FT_s} and X_{NS_s} were imputed based on the LR models defined in Table 3.

Table 3. Model (K) with variables that satisfied the three criteria for X_{FT} and X_{NS} for both T_{PMM} and T_{MI}

K	$U_{3,\max}$	$H_{S,\max}$	$D_{\max}$	$U_{\max}$	X_{BFE}
1	X	—	X	X	X
2	X	X	—	X	X
3	X	—	X	X	—
4	X	X	—	X	—
5	X	—	—	X	X
6	—	—	X	X	X
7	—	X	—	X	X
8	X	—	X	—	X
9	X	X	—	—	X
10	X	—	X	—	—
11	X	X	—	—	—
12	X	—	v	—	X
13	—	—	X	X	—
14	—	X	—	X	—
15	—	—	X	—	X
16	—	X	—	—	X
17	—	—	—	X	X
18	—	—	X	—	—
19	—	X	—	—	—
20	—	—	—	—	X
21	—	—	X	X	X
22	—	X	—	X	X
23	X	—	—	X	—
24	—	—	X	X	—
25	—	—	X	—	X
26	—	—	—	X	—
27	—	—	X	X	X
28	—	X	—	X	X
29	—	—	X	—	X
30	X	—	X	—	X
31	—	—	X	X	—
32	—	—	X	—	X
33	X	X	—	—	—

Table 4. Percentage of observed η or missingness λ of X_{FT_s} and X_{NS_s} in Hancock, Harrison, and Jackson Counties after observation deletion

		Frequency (η or λ)		
Symbol	Description	Hancock	Harrison	Jackson
X_{FT_s}	Observed	49 (52%)	123 (57%)	120 (63%)
	Missing	46 (48%)	93 (43%)	71 (37%)
X_{NS_s}	Observed	9 (22%)	105 (53%)	179 (77%)
	Missing	32 (78%)	94 (47%)	53 (23%)

Case Study Imputation Model Comparison

Among the models defined in Table 3 and for each T and each X_g, one imputation model LR with the highest CCR and the lowest CE was chosen. Error matrix values for LR obtained from T_{PMM} and T_{MI} are provided for X_{FT} and X_{NS} in Table 5. Rows of the matrices represent the frequency of slab, elevated foundation, and one and two stories prior to deletion, and columns represent the frequency of the imputed binary variable levels after deletion.

The results show that in general the performance of PMM imputation models is higher than that of the MI imputation models, with CCR ranging from 60% to 93%. For X_{FT}, T_{PMM} Models 13 and 24 have higher $CCR_{LR_{FT}}$ and lower $CE_{LR_{FT}}$ than T_{MI} Models 2 and 24 in Hancock and Jackson counties, whereas T_{MI} Model 25 has higher $CCR_{LR_{FT}}$ and lower $CE_{LR_{FT}}$ than T_{PMM} Model 22 Harrison county. For X_{NS}, T_{PMM} Models 11, 31, and 10 have higher $CCR_{LR_{NS}}$ and lower $CE_{LR_{NS}}$ than T_{MI} Models 15, 32, and 28 in the three counties of the study area. Based on the CCR and CE values, final imputation models, bold in Table 5, are defined for X_{FT} and X_{NS} in each county of the study area.

Eqs. (7) and (9) define final T_{PMM} imputation models (Models 13 and 24) for foundation type in Hancock and Jackson Counties, respectively, and Eq. (8) defines the final T_{MI} imputation model (Model 25) for foundation type in Harrison County. These models estimate the probability that buildings with missing foundation data have elevated foundations

$$\ln\left(\frac{P(X_{\mathrm{FT}} = \mathrm{Elevated})}{1 - P(X_{\mathrm{FT}} = \mathrm{Elevated})}\right) = -6.14 + 1.83D_{\max} - 2.84U_{\max} \tag{7}$$

$$\ln\left(\frac{P(X_{\mathrm{FT}} = \mathrm{Elevated})}{1 - P(X_{\mathrm{FT}} = \mathrm{Elevated})}\right) = 0.37 - 0.26D_{\max} + 0.6X_{\mathrm{BFE}} \tag{8}$$

Table 5. Observed versus imputed X_{g_s} error matrices, mean CE, and CCR for X_{FT_s} and X_{NS_s} in Hancock, Harrison, and Jackson Counties

	Imputed X_{g_s}												
County	T	FT	K	Slab	Elevated	$CE_{LR_{FT}}$	$CCR_{LR_{FT}}$	NS	K	1-story	2-story	$CE_{LR_{NS}}$	$CCR_{LR_{NS}}$
Observed X_{g_s}													
Hancock	T_{PMM}	Slab	**13**	**21**	**2**	**9%**	**93%**	1-story	**11**	**19**	**1**	**5%**	**81%**
		Elevated		**1**	**22**	**4%**		2-story		**5**	**7**	**42%**	
	T_{MI}	Slab	2	8	15	65%	61%	1-story	15	15	5	25%	78%
		Elevated		3	20	13%		2-story		2	10	17%	
Harrison	T_{PMM}	Slab	22	25	13	34%	60%	1-story	**31**	**69**	**4**	**5%**	**80%**
		Elevated		24	31	44%		2-story		**15**	**6**	**71%**	
	T_{MI}	Slab	**25**	**20**	**18**	**47%**	**72%**	1-story	32	70	3	4%	77%
		Elevated		**8**	**47**	**15%**		2-story		19	2	90%	
Jackson	T_{PMM}	Slab	**24**	**49**	**9**	**16%**	**85%**	1-story	**10**	**36**	**9**	**20%**	**77%**
		Elevated		**2**	**11**	**15%**		2-story		**3**	**5**	**38%**	
	T_{MI}	Slab	24	48	10	17%	79%	1-story	28	34	11	24%	70%
		Elevated		5	8	38%		2-story		5	3	63%	

Note: Bold = final imputation models. Rows of the matrices represent the frequency of slab, elevated foundation, and one and two stories prior to deletion; columns represent the frequency of the imputed binary variable levels after deletion.

$$\ln\left(\frac{P(X_{\mathrm{FT}}=\text{Elevated})}{1-P(X_{\mathrm{FT}}=\text{Elevated})}\right)=-2.86+0.29D_{\max}+4.18U_{\max} \tag{9}$$

Eqs. (10)–(12) define the final T_{PMM} imputation models (Models 11, 31, and 10) for number of stories in Hancock, Harrison, and Jackson Counties, respectively. These models estimate the probability that buildings with missing number of stories data are 1-story buildings

$$\ln\left(\frac{P(X_{\mathrm{NS}}=\text{1-story})}{1-P(X_{\mathrm{NS}}=\text{1-story})}\right)=24.28-0.32U_{3,\max}-5.62H_{S,\max} \tag{10}$$

$$\ln\left(\frac{P(X_{\mathrm{NS}}=\text{1-story})}{1-P(X_{\mathrm{NS}}=\text{1-story})}\right)=2.66+0.2D_{\max}-6.08U_{\max} \tag{11}$$

$$\ln\left(\frac{P(X_{\mathrm{NS}}=\text{1-story})}{1-P(X_{\mathrm{NS}}=\text{1-story})}\right)=5.67-0.06U_{3,\max}-0.33D_{\max} \tag{12}$$

Given the binary missing explanatory variables, LR_{PMM} imputation models performed better than LR_{MI} imputation models in imputing foundation type and number of stories for five of the six models in the three counties of the study area. The missing observations have been imputed as a function of only continuous observed variables (i.e., hazard intensities and base flood elevation) without considering other building attributes; therefore, more observed building attributes in regions that share similar building construction patterns would allow the extension of the methodology to more comprehensive imputation models that impute building attributes with missing observations as a function of hazard and other known building attributes. Application of the developed approaches using comprehensive data sets may lead to stronger evaluation of the performance of imputation models.

Figs. 1 and 2 show comparisons between observed and imputed foundation types and number of stories, respectively, for the selected models [Eqs. (7)–(12)]. The relative frequency of imputed and observed foundation types are similar in Hancock County, whereas elevated foundations were imputed at a higher rate than observed in Harrison and Jackson Counties. One-story homes were imputed at a higher rate than observed in Hancock and Harrison Counties but at a lower rate than observed in Jackson Country. The alignment of imputed and observed frequencies reflect the model CCR (Table 5), where models with CCR greater than 90% result in imputed observations very close to the observed observations (Model 13 for foundation in Hancock), whereas models with CCR less than 90% resulted in imputed observations with greater variation from the observed observations.

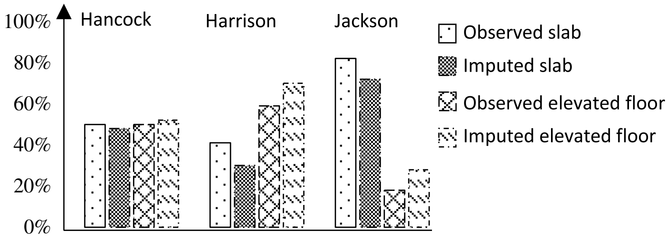

Fig. 1. Observed versus imputed foundation types in Hancock, Harrison, and Jackson Counties.

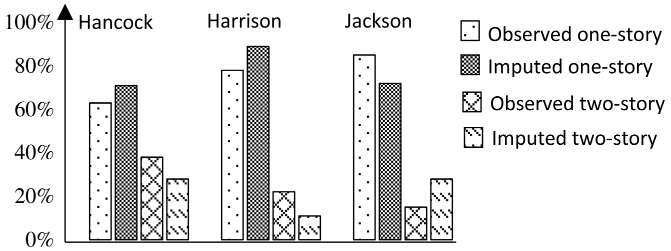

Fig. 2. Observed versus imputed number of stories in Hancock, Harrison, and Jackson Counties.

Study Limitations

The study is limited to 1- and 2-story wood-framed single-family homes with slab and elevated foundations subjected to hurricane hazards. Although the application of the developed diagnostic and comparison approaches is for a binary variable with missing observations, these methodologies may be expanded to multinomial variables and other imputation techniques [e.g., iterative robust model-based imputation (IRMI), multiple imputations of incomplete multivariate data (AMELIA), and sequential imputation for missing values (IMPSEQ)]. The results have been derived as a function of the range of hazard values experienced during Hurricane Katrina in coastal Mississippi and are therefore limited to this study.

Summary and Conclusions

This paper presented two imputation approaches along with imputation diagnostic and comparison approaches for binary building attribute data with missing observations. PMM and MI logistic regression–based imputation models were used to impute building attributes with missing observations. A diagnostic approach, based on the logistic regression goodness-of-fit test and the significance and independence of the observed numerical variables, was presented to evaluate the fit of PMM and MI imputation models. A comparison approach based on the percentage of correctly imputed observations and expressed in terms of CCR was presented to evaluate the performance of PMM and MI imputation models. A case study based on a data set of single-family homes damaged by 2005 Hurricane Katrina in coastal Mississippi was presented to demonstrate the application of the methodology. Missing foundation type and number of stories for single-family homes were imputed as a function of maximum 3-s wind speed, maximum significant wave height, maximum surge depth, maximum water speed, and base flood elevation. The contributions and findings of this paper are as follows:

- Approaches were developed to (1) impute binary building attributes with missing observations using imputation models rather than statistical models, (2) diagnose the fit of individual imputation models for binary categorical variables with missing observations rather than for continuous variables with missing observations, (3) determine the variables to be used in the imputation models, (4) evaluate the performance of the imputation models themselves rather than the statistical models fit on the imputed data set, and (5) determine the final imputation models to impute building attributes by comparing the performance of several imputation models.
- In the case study, PMM imputation models performed better performance than MI imputation models, where the average performance accuracy for PMM imputation models was 9% greater for foundation type and 4% greater for number of stories than for MI imputation models.

The developed approaches provide data collectors and damage modelers with the knowledge and guidance needed to impute building attributes with missing observations, particularly for model developers who rely on field data for risk assessment and building fragility models. Implementation of imputation techniques improves the quality of postevent data with incomplete building attributes, and implementation of diagnostic and comparison approaches ensure accuracy and validity of the imputation models themselves. Improving the quality of postevent data sets improves the development of data-based fragility and damage models, thus improving building damage prediction. Future work will extend the logistic regression–based imputation models to multinomial regression–based imputation models. Also, data with more observed building attributes in regions that share similar common building construction patterns will be collected so that missing observations can be imputed as a function of hazard parameters and other building attributes.

Data Availability Statement

Some or all data, models, or code generated or used during the study are available from the corresponding author by request, including data and code that are used to develop the imputation models.

Acknowledgments

The first author gratefully acknowledges funding from the Louisiana Board of Regents Graduate Fellowship in Engineering Grant No. LEQSF(2008-13)GF-01, the Donald W. Clayton Graduate Ph.D. Assistantship in Engineering at Louisiana State University, and the Chevron Engineering Graduate Student Fellowship at Louisiana State University. Hurricane Katrina reconnaissance videos were provided by MCEER.

References

Abayomi, K., A. Gelman, and M. Levy. 2008. “Diagnostics for multivariate imputations.” *J. Royal Stat. Soc. Series C (Appl. Stat.)* 57 (3): 273–291. https://doi.org/10.1111/j.1467-9876.2007.00613.x.

Akande, O., F. Li, and J. Reiter. 2017. “An empirical comparison of multiple imputation methods for categorical data.” *Am. Statistician* 71 (2): 162–170. https://doi.org/10.1080/00031305.2016.1277158.

Bernhardt, P. W. 2018. “Model validation and influence diagnostics for regression models with missing covariates.” *Stat. Med.* 37 (8): 1325–1342. https://doi.org/10.1002/sim.7584.

Booij, N., R. Ris, and L. H. Holthuijsen. 1999. “A third-generation wave model for coastal regions: 1. Model description and validation.” *J. Geophys. Res. Oceans* 104 (C4): 7649–7666. https://doi.org/10.1029/98JC02622.

Cabras, S., M. E. Castellanos, and A. Quirós. 2011. “Goodness-of-fit of conditional regression models for multiple imputation.” *Bayesian Anal.* 6 (3): 429–455. https://doi.org/10.1214/11-BA617.

Collins, L. M., J. L. Schafer, and C.-M. Kam. 2001. “A comparison of inclusive and restrictive strategies in modern missing data procedures.” *Psychol. Methods* 6 (4): 330. https://doi.org/10.1037/1082-989X.6.4.330.

Dietrich, J. C., S. Tanaka, J. J. Westerink, C. Dawson, R. Luettich Jr., M. Zijlema, L. H. Holthuijsen, J. Smith, L. Westerink, and H. Westerink. 2012. “Performance of the unstructured-mesh, SWAN+ ADCIRC model in computing hurricane waves and surge.” *J. Sci. Comput.* 52 (2): 468–497. https://doi.org/10.1007/s10915-011-9555-6.

Farhan, J., and T. Fwa. 2014. “Improved imputation of missing pavement performance data using auxiliary variables.” *J. Transp. Eng.* 141 (1): 04014–04065. https://doi.org/10.1061/(ASCE)TE.1943-5436.0000725.

Fay, R. E. 1996. “Alternative paradigms for the analysis of imputed survey data.” *J. Am. Stat. Assoc.* 91 (434): 490–498. https://doi.org/10.1080/01621459.1996.10476909.

FEMA. 2020. “FEMA flood map service center: Welcome!” Accessed September 29, 2016. https://www.fema.gov/.

Ferrari, P. A., P. Annoni, A. Barbiero, and G. Manzi. 2011. “An imputation method for categorical variables with application to nonlinear principal component analysis.” *Comput. Stat. Data Anal.* 55 (7): 2410–2420. https://doi.org/10.1016/j.csda.2011.02.007.

Gelman, A., I. Van Mechelen, G. Verbeke, D. F. Heitjan, and M. Meulders. 2005. “Multiple imputation for model checking: Completed-data plots with missing and latent data.” *Biometrics* 61 (1): 74–85. https://doi.org/10.1111/j.0006-341X.2005.031010.x.

Little, R. J., and D. B. Rubin. 2014. *Statistical analysis with missing data*. Hoboken, NJ: Wiley.

Luettich, R. A., and J. J. Westerink. 2004. “Formulation and numerical implementation of the 2D/3D ADCIRC finite element model version 44. XX.” Accessed April 16, 2018. https://adcirc.org/files/2018/11/adcirc_theory_2004_12_08.pdf.

Macabuag, J., T. Rossetto, I. Ioannou, A. Suppasri, D. Sugawara, B. Adriano, F. Imamura, I. Eames, and S. Koshimura. 2016. “A proposed methodology for deriving tsunami fragility functions for buildings using optimum intensity measures.” *Nat. Hazards* 84 (2): 1257–1285. https://doi.org/10.1007/s11069-016-2485-8.

Massarra, C. C., C. J. Friedland, B. D. Marx, and J. C. Dietrich. 2019. “Predictive multi-hazard hurricane data-based fragility model for residential homes.” *Coastal Eng.* 151 (Sep): 10–21. https://doi.org/10.1016/j.coastaleng.2019.04.008.

Meng, X.-L. 1994. “Multiple-imputation inferences with uncongenial sources of input.” *Stat. Sci.* 9 (4): 538–558. https://doi.org/10.1214/ss/1177010269.

Nguyen, C. D., J. B. Carlin, and K. J. Lee. 2017. “Model checking in multiple imputation: An overview and case study.” *Emerging Themes Epidemiol.* 14 (1): 8. https://doi.org/10.1186/s12982-017-0062-6.

Pita, G. L., R. Francis, Z. Liu, J. Mitrani-Reiser, S. Guikema, and J.-P. Pinelli. 2011. “Statistical tools for populating/predicting input data of risk analysis models.” In *Proc., 1st Int. Conf. on Vulnerability, Uncertainty, and Risk*, 468–476. Reston, VA: ASCE.

Raghunathan, T. E., J. M. Lepkowski, J. Van Hoewyk, and P. Solenberger. 2001. “A multivariate technique for multiply imputing missing values using a sequence of regression models.” *Survey Methodol.* 27 (1): 85–96.

Rubin, D. B. 1978. “Multiple imputations in sample surveys-a phenomenological Bayesian approach to nonresponse.” In *Proc., Survey Research Methods Section of the American Statistical Association*, 20–34. Alexandria, VA: American Statistical Association.

Stuart, E. A., M. Azur, C. Frangakis, and P. Leaf. 2009. “Multiple imputation with large data sets: A case study of the Children’s Mental Health Initiative.” *Am. J. Epidemiol.* 169 (9): 1133–1139. https://doi.org/10.1093/aje/kwp026.

Van Buuren, S. 2012. *Flexible imputation of missing data*. New York: CRC Press.

Westerink, J. J., R. A. Luettich, J. C. Feyen, J. H. Atkinson, C. Dawson, H. J. Roberts, M. D. Powell, J. P. Dunion, E. J. Kubatko, and H. Pourtaheri. 2008. “A basin-to channel-scale unstructured grid hurricane storm surge model applied to southern Louisiana.” *Mon. Weather Rev.* 136 (3): 833–864. https://doi.org/10.1175/2007MWR1946.1.

White, I. R., P. Royston, and A. M. Wood. 2011. “Multiple imputation using chained equations: Issues and guidance for practice.” *Stat. Med.* 30 (4): 377–399. https://doi.org/10.1002/sim.4067.

Zhu, H., J. G. Ibrahim, and X. Shi. 2009. “Diagnostic measures for generalized linear models with missing covariates.” *Scand. J. Stat.* 36 (4): 686–712. https://doi.org/10.1111/j.1467-9469.2009.00644.x.

Zijlema, M. 2010. “Computation of wind-wave spectra in coastal waters with SWAN on unstructured grids.” *Coastal Eng.* 57 (3): 267–277. https://doi.org/10.1016/j.coastaleng.2009.10.011.