A Predictive Modeling Approach to Estimating Seismic Retrofit Costs

Juan F. Fung,[a)] Siamak Sattar,[b)] David T. Butry,[b)] and Steven L. McCabe[b)]

This paper presents a methodology for estimating seismic retrofit costs from historical data. In particular, historical retrofit cost data from FEMA 156 is used to build a generalized linear model (GLM) to predict retrofit costs as a function of building characteristics. While not as accurate as an engineering professional's estimate, this methodology is easy to apply to generate quick estimates and is especially useful for decision makers with large building portfolios. Moreover, the predictive modeling approach provides a measure of uncertainty in terms of prediction error. The paper uses prediction error to compare different modeling choices, including the choice of distribution for costs. Finally, the proposed retrofit cost model is implemented to estimate the cost to retrofit a portfolio of federal buildings. The application illustrates how the choice of distribution affects cost estimates.

INTRODUCTION

A decision maker faced with the task of estimating the cost of a seismic retrofit for a single building will most likely hire an engineering consulting firm to evaluate the building and determine the cost of retrofitting the building.

Suppose the decision maker wants to obtain retrofit cost estimates for a portfolio of several hundred, or even several thousand, buildings. The costs and time associated with estimating retrofit costs for a large number of buildings may prevent the decision maker from making timely budgeting decisions. If a decision maker simply wants to know whether retrofitting the building portfolio is a good investment, there is a non-trivial cost associated with obtaining the information needed to make a decision.

This paper presents a predictive model to obtain retrofit cost estimates based on *historical* retrofit cost data. Given a set of building characteristics such as building size and age, the decision maker can predict the average retrofit cost for the building. Importantly, the predictive

[a)]National Institute of Standards and Technology, 100 Bureau Dr., Mailstop 8603, Gaithersburg, MD 20899, juan.fung@nist.gov

[b)]National Institute of Standards and Technology, 100 Bureau Dr., Gaithersburg, MD 20899

model is agnostic to the choice of retrofit *action*, making predictions using building characteristics as predictors.

The predictive model can be used for a single building or for a large portfolio of buildings, providing a quick way to estimate retrofit costs. While these predictions have higher uncertainty compared to estimates from an engineering consulting professional, upper and lower bounds on predictions are straightforward to obtain from the model and provide a measure of the degree of uncertainty. An application to an actual building portfolio illustrates the approach, as well as some of the modeling choices discussed in the paper.

BACKGROUND AND LITERATURE

The historical retrofit cost data used in this paper was originally collected for FEMA 156 (FEMA, 1994) and is freely available online. In particular, the data can be found as part of FEMA's archived Seismic Rehabilitation Cost Estimator (SRCE) software, (FEMA, 2013–2014). Further information on the data set is provided in the section "The Training Data."

Elements of the predictive-modeling methodology, inspired by FEMA 156 and FEMA 157 (FEMA, 1995), are developed by the authors in Fung et al. (2017). FEMA 156 made two major contributions: (1) the collection of a reliable data set of retrofit costs and building characteristics; and (2) methods for estimating average retrofit costs, including a linear regression model.

The linear regression model in FEMA 156 estimates average retrofit cost using building characteristics, such as building size and age, as predictors. FEMA 156 and FEMA 157 present the model, but do not test how well it performs in predicting retrofit costs. Fung et al. (2017) show that a model with a different set of predictors, outperforms the FEMA 156 model in the sense of having a lower prediction error in predicting retrofit costs. For instance, Fung et al. (2017) show that including an indicator for whether a building is deemed historic as an additional predictor also lowers prediction error.

In a series of papers, Jafarzadeh et al. (2013b,a, 2014) collect and analyze a database on retrofit costs for 158 public schools in Iran. Jafarzadeh et al. (2014) provides the associated data, as well as a detailed discussion of the data collection effort and a description of the data. While the database is fairly detailed and more modern than the SRCE data, it is not applicable for predicting retofit costs for buildings in the United States because building codes differ across countries.

Jafarzadeh et al. (2013b) analyze the data using standard linear regression, while Jafarzadeh et al. (2013a) apply artificial neural networks to predicting costs. The main objective of the papers is to explore which predictors matter most for retrofit costs (Jafarzadeh et al., 2013b) and the parameterizations that are likely to minimize prediction error (Jafarzadeh et al., 2013a). While neural networks are a promising direction for predicting retrofit costs, it is difficult to draw conclusions on the application of neural networks for cost prediction from Jafarzadeh et al. (2013a). First, it is unclear whether a neural network can provide much improvement over standard linear regression with a set of only 158 training samples. More importantly, the authors use the hold-out method for both model selection and evaluation, potentially problematic approaches as discussed below.

More recently, Nasrazadani et al. (2017) collected their own database of 167 retrofits of masonry school buildings in Iran. The authors use Bayesian linear regression in order to predict retrofit costs. The main objective is to compare retrofit costs for three retrofit actions for a given level of expected gain in performance, measured as the change in lateral strength after retrofit. In addition to the performance gain, Nasrazadani et al. (2017) find that a building's pre-retrofit value is an important predictor of retrofit cost.

Finally, while the focus of this paper is on seismic risk *mitigation* (i.e., pre-event), several recent papers study actual repair and retrofit costs for buildings that are damaged in the aftermath of a seismic event. Di Ludovico et al. (2017a,b) obtained estimates of repair costs for residential buildings damaged in the 2009 L'Aquila earthquake in central Italy. Del Vecchio et al. (2018) compare actual repair costs from buildings damaged in L'Aquila to predictions based on FEMA P-58. Finally, Cremen and Baker (2019) validate component-level loss predictions based on FEMA P-58 to actual losses experienced in the 2011 Christchurch earthquake in New Zealand.

CONTRIBUTIONS

This paper extends the methodology developed in Fung et al. (2017) in three key directions:

- **Model assumptions**: The predictive methodology developed in Fung et al. (2017) is based on a standard linear regression model. This paper uses a *generalized linear model* (GLM) and compares GLM to standard linear regression, measuring model performance in terms of prediction error. The key advantage of GLM over standard linear regression is that predictions are more easily interpretable, as discussed below.

- **Model selection and evaluation**: Fung et al. (2017) evaluate models using the "hold-out" method, where a subset of the data is held out when fitting (or "training") a model. Thus, the hold-out method does not use all of the available data to train a model. Moreover, Fung et al. (2017) focus on model *evaluation*, assessing how well a model performs, rather than model *selection*, choosing the best among several competing models. This paper uses nested K-fold cross-validation to perform both model selection and model evaluation. A key advantage of K-fold cross-validation is that it uses all of the available data to train each model.

- **Outcome of interest**: Fung et al. (2017) focus on predicting total construction costs associated with a retrofit, which can include costs associated with repairing damage due to other environmental impacts and with removing hazardous material in addition to costs of structural mitigation. Fung et al. (2018) apply the methodology to predict *structural* retrofit costs only. The current paper compares model performance in predicting structural and total costs.

In addition, the impacts of some of the modeling assumptions on retrofit cost predictions are illustrated on an actual building portfolio. The results demonstrate the flexibility and applicability of the predictive modeling approach for seismic risk mitigation. In particular, the illustration of key modeling decisions and the tradeoffs associated with those decisions should be valuable for owners of building portfolios during the planning phase (that is, pre-evaluation) of a potential seismic retrofit program. Such easily obtainable order of magnitude estimates can assist in making the decision to pursue further action, such as an evaluation.

MODEL DEVELOPMENT

This section develops the predictive retrofit cost-estimating model, including a discussion of the predictors, and describes nested K-fold cross-validation, used for model selection and evaluation.

PREDICTIVE MODEL

Suppose a decision maker has information on building characteristics, such as building size and age, for each building in a portfolio. The decision maker would like to know the cost to retrofit each building, given the building characteristics. That is, the decision maker would like to use the building characteristics, X, as *predictors* to predict retrofit cost, Y, as $\hat{Y} = \hat{f}(X)$.

The goal of prediction is to obtain an estimator $\hat{f}$ that can produce predictions $\hat{Y}$ for *any* input X (James et al., 2013). Prediction is applicable when the objective is to obtain an outcome of interest that is not easily obtainable. The estimator $\hat{f}$ is obtained using existing information on the relationship between X and Y; for instance, from buildings that have already been retrofitted. Once the estimator $\hat{f}$ is obtained, the decision maker can simply plug in any X to obtain predictions for the buildings of interest.

FEMA 156 proposes a predictive model of retrofit cost (in dollars per square foot), $\hat{Y} = \hat{f}(X)$, where X includes all of the predictors given in Table 1, except for the Historic indicator.

In addition, the FEMA 156 model includes *interactions* between each predictor and building type, b. Interactions capture the combined effects of two (or more) predictors; e.g., the interaction between building type and age captures the possibility that retrofit costs for older unreinforced masonry (URM) buildings are different than retrofit costs for newer URM buildings as well as for older buildings of other types. Such full-interaction models, in which each predictor is interacted with building type, effectively result in a large number of predictors. In practice, it is equivalent to training separate models for each building type.

Table 1. Definition of outcome, Y, and set of predictors, X, used in this paper.

Variable	Definition
Y	Retrofit cost (in dollars per square foot)
s	Seismicity (e.g., peak ground acceleration)
p	Performance objective (e.g., life safety)
b	Building type (e.g., unreinforced masonry, wood frame)
Area	Building area (in square feet)
Age	Building age (in years)
Stories	Number of above and below ground stories
Occup	Occupancy during retrofit (e.g., vacate occupants from building)
Historic	Is building deemed historic? (yes or no)

Fung et al. (2017) show that full interactions (equivalently, separate regressions by building type) are unnecessary. In particular, Fung et al. (2017) show that a simpler model, without building-type interactions, results in lower prediction error. The only interaction that Fung et al. (2017) include is the interaction between seismicity and performance objective. The logic is that the combined effects of seismicity and the performance objective also affect retrofit costs.

In addition, Fung et al. (2017) include a Historic indicator to account for the fact that historic buildings are treated differently (often, uniquely) in a retrofit. The results in Fung et al. (2017) suggest that this model outperforms the FEMA 156 model. In this paper, the same set of predictors is used to train the predictive models.

CHOOSING AN ESTIMATOR

This paper considers two types of estimators. The first is obtained from standard linear regression, which assumes that the outcome of interest is normally distributed. The second is obtained from a Generalized Linear Model (GLM), which relaxes the assumption that the outcome is normally distributed. It should be noted that the choice of an estimator is not a choice of a model for the true data generating process for Y. In other words, it is not meant to be an accurate, physical model of Y. Rather, the goal is to obtain an estimator that will predict $E[Y|X]$ with high accuracy.

Standard linear regression, or Ordinary Least Squares (OLS), combines a linear estimator, $\hat{Y} = \hat{f}(X) = X\hat{\beta}$, with the assumption that the outcome, Y, follows a normal distribution, conditional on X (to be precise, the distributional assumption is only required for inference; for instance, for deriving confidence and prediction intervals). The estimator is called linear because it is a linear combination of the predictors, X.

The standard linear regression model estimated in Fung et al. (2017) is:

$$\begin{aligned} \ln(Y) = \beta_0 + \beta_s + \beta_p + \beta_{sp} + \beta_b + \beta_1 \ln(\text{Area}) + \beta_2 \ln(\text{Age}) \\ + \beta_3 \ln(\text{Stories}) + \beta_4 \text{Occup} + \beta_5 \text{Historic} + \epsilon \end{aligned} \tag{1}$$

where Y is assumed to follow a log-normal distribution, i.e., $\ln(Y)|X$ follows a normal distribution with $E[\ln(Y)|X] = X\beta$, for the coefficient vector $\beta = \{\beta_0, \beta_s, \ldots, \beta_5\}$, and $E[\epsilon|X] = 0$, i.e., the predictors X are uncorrelated with any unobserved variation in Y. Note that the error term, ϵ, captures the effect of predictors that are not included in the model.

The Generalized Linear Model (GLM) is an extension of standard linear regression that uses a linear predictor but does not assume the outcome follows a normal distribution. In particular, the outcome may follow any distribution in the exponential family, which includes the normal distribution (Coxe et al., 2013).

A GLM has three components: (1) a random component, which refers to the distribution of the outcome, $Y \sim F_{Y|X}$; (2) a systematic component, which refers to the linear combination of predictors, $X\beta$; and (3) an invertible link function, which specifies the relationship between the random and systematic components, $g(E[Y|X]) = X\beta$. Note that the link function, g, allows for potentially *nonlinear* relationships between the mean of Y and the predictors, X.

The standard linear regression model is a special case of a GLM. Consider the model given in Eq. (1). The random component is the normal distribution of the outcome variable, $\ln(Y)|X \sim N(X\beta, \sigma^2)$ where $\sigma^2 \equiv Var(\epsilon|X)$. The systematic component refers to the right-hand side of Eq. (1), the linear combination of predictors. Finally, note that the link function is the identity function, since $g(E[\ln(Y)|X]) = E[\ln(Y)|X] = X\beta$.

More generally, one can use knowledge about the outcome of interest in choosing a GLM specification. For instance, if the outcome of interest is cost, a reasonable choice of distribution might be *skewed* to the right, with much of the mass concentrated on smaller values of cost and a long right tail. Note that while the normal distribution is symmetric around the mean, the log-normal distribution is right-skewed.

Moreover, the distribution of cost should be *non-negative*. Note that while normally distributed variables can take positive or negative values, the OLS model assumes $Y|X$ follows a log-normal distribution and, thus, must be non-negative. However, this forces the outcome of interest to be $\ln(Y)$ rather than Y itself.

Alternative choices of distribution in the exponential family include the gamma distribution and the inverse normal distribution. Each satisfies the desired properties: the distributions are right-skewed and random variables can only take positive values. An advantage of using a gamma or inverse normal distribution, rather than the log-normal distribution, is that the outcome of interest is Y itself rather than $\ln(Y)$. Other common distributions in the exponential family do not satisfy the desired properties.

Since $X\beta$ may take postive or negative values, a suitable link function must ensure that $E(Y|X)$ is positive. For a model with gamma or inverse normal distribution, the natural logarithm is a suitable link function. That is, $g(E[Y|X]) = \ln(E[Y|X]) = X\beta$, or:

$$\begin{aligned}\ln(E[Y|X]) = \beta_0 + \beta_s + \beta_p + \beta_{sp} + \beta_b + \beta_1 \ln(\text{Area}) + \beta_2 \ln(\text{Age}) \\ + \beta_3 \ln(\text{Stories}) + \beta_4 \text{Occup} + \beta_5 \text{Historic}\end{aligned} \tag{2}$$

The difference between Eq. (1) and Eq. (2) is that Eq. (1) is a model for the mean of the log, $E[\ln(Y)|X]$, while Eq. (2) is a model for the log of the mean, $\ln(E[Y|C])$, which are not equivalent in general. An important implication is that applying the exponential function to the right-hand side of Eq. (2) yields mean cost, $E[Y|X]$, which is directly interpretable in dollars per square foot. In contrast, applying the exponential function to the right-hand side of Eq. (1) yields $\exp\{E[\ln(Y)|X]\}$, which is much more difficult to interpret. This is the main advantage of the gamma and inverse normal distributions for modeling costs.

ESTIMATING PREDICTION ERROR

Root Mean Squared Error (RMSE) is used as the measure of prediction error for new data, known as out-of-sample prediction error. Let $\hat{\beta}$ denote the coefficient vector obtained from training the model (i.e., from estimating $\hat{f}$). The RMSE is estimated from a set of data of size m that is **not** used to train the model as:

$$RMSE \equiv \left(\frac{1}{m}\sum_{i=1}^{m}(\hat{\beta}^T x_i - g(Y_i))^2\right)^{\frac{1}{2}} \tag{3}$$

Note that RMSE as defined in Eq. (3) is estimated on the scale of the link function applied to the outcome variable, $g(Y)$. Alternatively, RMSE may be estimated using the inverse of the link function, with terms $(g^{-1}(\hat{\beta}^T x) - Y)^2$. For instance, if $g = \ln$, then $g^{-1} = \exp$ and RMSE is on the original scale of the outcome, Y, as shown in Eq. (4).

$$RMSE \text{ (dollars) } \equiv \left(\frac{1}{m}\sum_{i=1}^{m}(g^{-1}(\hat{\beta}^T x_i) - Y_i)^2\right)^{\frac{1}{2}} \tag{4}$$

To obtain a reasonable estimate of RMSE, Eq. (3) should be estimated on data that has not been used to train the model. Otherwise, the model will already be familiar with the data and estimates of model performance will be biased. However, the purpose for estimating RMSE must also be taken into account. In particular, the data used to evaluate how a particular model will perform on new data (model evaluation) should not be used to also *compare and select* different models (model selection); (Guyon and Elisseeff, 2003). The potential problem is that data used for model selection is inadvertently used to train the model, biasing estimates of model performance.

K-fold cross-validation is a method for estimating RMSE. The idea is to split the training data into K mutually exclusive subsets, or "folds," iteratively using each fold as a test set while using the remaining $K - 1$ folds as training data (Arlot and Celisse, 2010). A key advantage of this approach is that it uses all of the available data to train and to test the model. The estimate of Eq. (3) is obtained by averaging RMSE estimates across each fold.

In order to perform model selection and model evaluation together, this paper uses nested K-fold cross-validation (Krstajic et al., 2014). Nested K-fold cross-validation applies K-fold cross-validation for model selection *before* applying K-fold cross-validation for model evaluation. Loosely, this may be thought of as performing an additional K-fold cross-validation within each of the K folds, so that model selection is nested within model evaluation. The idea is to select the best model first and to evaluate its performance *after* selection. The procedure guarantees that different parts of the data are used for each step, and still uses all of the available data to train and to test models. RMSE estimates are obtained by averaging across all iterations.

THE TRAINING DATA

As mentioned in the introduction, the training data used in this paper was originally collected for FEMA 156. The publicly available version of the data (the SRCE data) includes 1978 buildings, compared to the 2088 collected for FEMA 156. The SRCE data set is missing an important building characteristic that is used in FEMA 156: building occupancy class. Nevertheless, the discussion in FEMA 156 suggests this data set should be representative of commercial and residential buildings in the United States and Canada.

Table 2 presents summary statistics for retrofit costs and for the non-categorical building characteristics that appear in Eq. (1) and Eq. (2), including building area (in square feet), building age (in years), and building height (in stories). Note that total costs are 44% higher than structural costs on average and have much higher variability. The building characteristics area, age, and stories also exhibit large variability and cover a broad spectrum of buildings. Finally, Table 2 also presents a measure of seismicity, PGA, which is discussed below.

Table 2. Summary statistics for outcomes of interest and select predictors in the training (SRCE) data, with N = 1526 excluding Canadian buildings (1 ft = 0.3048 m).

Variable	Minimum	Mean	Median	Maximum	Standard deviation
Structural cost ($/sq ft)	0.49	36.03	23.33	675.42	44.74
Total cost ($/sq ft)	0.49	52.13	28.84	1688.55	81.95
Area (1000 sq ft)	0.15	68.98	28.67	1430.30	113.26
Age	2.00	44.29	40.00	153.00	22.13
Stories	1.00	3.12	2.00	38.00	2.99
PGA (g_n)	0.01	0.27	0.36	0.58	0.15

Note that the sample size in Table 2, $N = 1526$, corresponds to buildings within the contiguous United States. In particular, it excludes 187 buildings in Canada and 14 buildings in the US territories due to the challenges in obtaining consistent seismic hazard data across all locations (no buildings in Alaska or Hawaii appear in the SRCE data). Another 190 buildings are excluded due to missing data (e.g., building age).

Costs in the original SRCE data are normalized to average construction costs in California for 1993. Following Fung et al. (2017), this paper presents costs normalized to 2016 national average construction costs, using the Engineering News Record's Building Construction Index (BCI) (ENR, 2017). It is worth noting that retrofit engineering practice has evolved since the SRCE data was collected, likely decreasing the rate of growth in retrofit costs relative to the growth in the material and labor costs represented by the BCI.

Total construction costs include costs of structural mitigation, as well as additional costs triggered by the retrofit, including: (1) costs associated with compliance with the Americans with Disabilities Act of 1990 (ADA 1990); (2) costs associated with removal of asbestos and other hazardous material; (3) costs associated with repairing damage or deterioration; and (4) non-structural mitigation costs. Total costs vary greatly. Moreover, it is difficult to say how they correspond to costs today (e.g., compliance with ADA 1990). In contrast, structural costs are the construction costs associated with the retrofit of structural components.

Fig. 1 presents a histogram for structural retrofit costs in the SRCE data. Note the distribution is right-skewed, with a very long and thin right tail. While the histogram approximates the *unconditional* distribution of costs, it nevertheless illustrates the properties a cost distribution is expected to exhibit.

The measure of seismicity provided in the SRCE data is based on outdated seismic hazard maps from ATC-3 (ATC, 1978). In practice, decision makers will use more recent seismic hazard maps, such as those produced by the US Geological Survey (USGS). The measure of seismicity used in this paper is based on USGS peak ground acceleration (PGA) as a fraction of standard gravity (g_n) with a 10% probability of exceedance in 50 years (USGS, 2014). In particular, PGA is averaged at the county level, because that is the finest location information provided in the SRCE data, and weighted by Census-tract population as a proxy for building density (Fung et al., 2017). There is no claim that this is the best measure of regional seismicity for a building. A decision maker may choose to use another measure if that is desired. For instance, another measure that may contrinute to retrofit cost is seismicity at the time the building

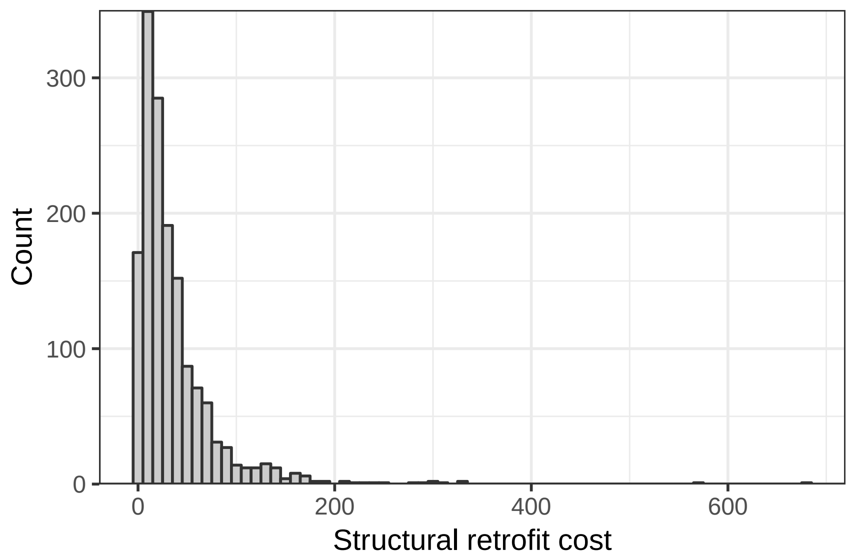

Figure 1. Histogram of structural retrofit costs from the training (SRCE) data. Costs are in dollars per square foot (1 ft = 0.3048 m).

was constructed (since this data is not readily available for the SRCE data, it is expected that building age acts as a proxy for this information). The important point is to be able to capture the variation in seismic risk each building faces.

PGA for the buildings in the training data is shown on the last line of Table 2. Note that PGA ranges from a minimum of 0.01 to a maximum of 0.58, with mean PGA of 0.27. Given PGA, the buildings are assigned to one of four seismicity *categories* as summarized in Table 3: "Low" seismicity corresponds to $PGA < 0.1$; "Medium" seismicity corresponds to $PGA \in [0.1, 0.2)$; "High" seismicity to $PGA \in [0.2, 0.4)$; and "Very High" seismicity is $PGA >= 0.4$. These values are based on FEMA 156's definitions. Note that buildings in the training data are roughly evenly distributed across each seismicity category.

Table 3. Definition of seismicity categories used in this paper, as a function of PGA (g_n), and their relative shares in the training (SRCE) data.

Seismicity	Lower bound	Upper bound	Count	Percentage
Low (L)	0.0	0.1	345	23%
Moderate (M)	0.1	0.2	339	22%
High (H)	0.2	0.4	414	27%
Very High (VH)	0.4	$\max PGA$	428	28%

The performance objective categories represented in the SRCE data are defined in FEMA 156 as follows: Life Safety (LS) "allows for unrepairable damage as long as life is not jeopardized and egress routes are not blocked;" Damage Control (DC) "protects some feature or function of the building beyond life-safety, such as protecting building contents or preventing the release of toxic material;" and Immediate Occupancy (IO) "allows only minimal post-earthquake damage and disruption, with some nonstructural repairs and cleanup done while the building remains occupied and safe." Since DC is no longer used and is not directly comparable to current performance objectives, this paper focuses on predicting LS and IO. Table 4 presents the number and percentage of each performance objective category in the SRCE data. Note that over half of the training data corresponds to a target performance objective of LS.

Table 4. Performance objective by number and percentage in the training (SRCE) data.

Performance objective	Count	Percentage
LS	822	53.9%
DC	444	29.1%
IO	260	17.0%

The term Occup in Table 1 represents what happens to occupants during retrofit construction and is defined as follows: In-place (IP) means that work is scheduled around normal hours of occupancy; Temporarily removed (TR) means that occupants are moved to another room in the building during construction; and Vacant (V) means that the building is completely vacated during construction. In terms of construction costs, completely vacating the building is the lowest-cost option, while leaving occupants in-place is the most expensive option. However, in practice occupant relocation costs would also be taken into account, potentially making the decision to vacate the building less cost-effective. Table 5 summarizes each occupancy category for the training data. The majority of cost estimates in the training data represent retrofits for which occupants have been temporarily removed.

Building historic status, represented by the variable Historic, simply indicates whether or not a building is deemed historic (and, therefore, must be treated carefully during a retrofit). In particular, Historic buildings are expected to cost more to retrofit than other buildings, all else equal, because of the need to preserve the structure and character of the building. In the training data, 155 buildings, or 10.2%, deemed historic.

Finally, Table 6 presents the 15 building construction types in the SRCE data. Following

Table 5. Occupancy during retrofit by number and percentage in the training (SRCE) data.

Occupancy during retrofit	Count	Percentage
V	267	23.2%
TR	647	56.3%
IP	235	20.5%

FEMA 156, each building type is assigned to one of eight building *groups*, based on structural similarities, as shown in Table 6. Strictly speaking, Eq. (1) and (2) each use building group rather than type.

Table 6. Building types, groups, and their shares in SRCE data.

Building Group	Building type	Model name	Count	Percentage
1	URM	Unreinforced Masonry	459	30.08%
2	W1	Wood Light Frame	40	2.62%
	W2	Wood (Commerical or Industrial)	47	3.08%
3	PC1	Precast Concrete Tilt Up Walls	51	3.34%
	RM1	Reinforced Masonry with Metal or Wood Diaphragm	51	3.34%
4	C1	Concrete Moment Frame	103	6.75%
	C3	Concrete Frame with Infill Walls	254	16.64%
5	S1	Steel Moment Frame	74	4.85%
6	S2	Steel Braced Frame	28	1.83%
	S3	Steel Light Frame	11	0.72%
7	S5	Steel Frame with Infill Walls	107	7.01%
8	C2	Concrete Shear Wall	247	16.19%
	PC2	Precast Concrete Frame with Infill Walls	12	0.79%
	RM2	Reinforced Masonry with Precast Concrete Diaphragm	10	0.66%
	S4	Steel Frame with Concrete Walls	32	2.10%

MAIN RESULTS

This section presents predicition error estimates for several modeling choices, using K-fold cross-validation with $K = 10$ folds. RMSE estimates are obtained by averaging prediction error across the folds. Coefficient estimates, $\hat{\beta}$, are shown in Appendix Table A2.

CHOICE OF DISTRIBUTION

One of the first modeling choices a decision maker will make is the choice of distribution. As shown in Fig. 1, a model for cost should use a distribution that is right-skewed and that guarantees the outcome of interest does not take negative values.

This section considers a GLM with systematic component given in Eq. (2) and the choice of distribution for the outcome. The choices considered are the normal distribution with link function $g = \ln$; the gamma distribution with link $g = \ln$; and the inverse normal distribution with link $g = \ln$.

The choice is therefore between a symmetric distribution for Y, which allows Y to take positive or negative values, and a skewed distribution for Y, which restricts Y to take only positive values. Note that the normal distribution with log link is not equivalent to standard linear regression, since it is a model for $\ln(E[Y|X]) = X\beta$. Table 7 presents the estimates of RMSE and its standard deviation, σ_{RMSE}, assuming the outcome of interest is structural cost. Estimates of RMSE and σ_{RMSE} are presented on the original (dollars per square foot) scale of the outcome, as in Eq. (4).

Table 7. Model selection for three choices of outcome distribution, when the outcome is structural cost in dollars per square foot (1 ft = 0.3048 m). RMSE estimates and standard deviation of RMSE estimates, σ_{RMSE}, are in dollars per square foot. The GLM with gamma distribution, with the lowest RMSE, is the chosen model.

Model	RMSE	σ_{RMSE}
GLM-Gamma	40.43	9.75
GLM-Normal	41.66	8.97
GLM-Inverse Normal	41.87	9.13

The GLM with gamma distribution has the lowest RMSE and can therefore be interpreted as the preferred model. However, the RMSE for the other GLMs are only about 3% to 3.5% larger. It is difficult to conclude that the gamma model is superior, especially when considering σ_{RMSE}. Given that these are noisy estimates of RMSE, it is plausible that a different draw of the data could result in choosing the normal or inverse normal model.

The lesson is that model selection is not always straightforward. Models should be chosen for interpretability and for coherence with their context. In this case, retrofit costs are highly skewed and should not take negative values. While the results do not strongly suggest that the GLM with gamma distribution is superior, they do suggest that this choice of distribution is reasonable. The inverse normal distribution, which is more skewed than the gamma distribution, would also be reasonable.

OLS VS GLM

Another decision is whether it is necessary to use a GLM, or whether a standard linear regression model (OLS) would perform just as well. This section evaluates each of these models for their expected out-of-sample performance.

Strictly speaking, the outcome in Eq. (1), $\ln(Y)$, is not the same as the outcome in Eq. (2), Y. Thus, model selection as performed in the preceding section cannot be performed. This is because each model has a different task: the task of the OLS model is to predict $E[\ln(Y)|X]$, while the task of the GLM is to predict $\ln(E[Y|X])$.

Table 8 presents the results of model evaluation, that is, estimating each model's expected out-of-sample performance, for the standard linear regression model, as well as the GLM models presented in the preceding section, when the outcome is structural cost in dollars per square foot. The results suggest that the GLM with gamma distribution is expected to have the best out-of-sample performance for its prediction task than any of the other models.

Table 8. Model evaluation: expected out-of-sample performance for OLS and three GLM specifications, when the outcome is structural cost in dollars per square foot (1 ft = 0.3048 m). Actual and predicted mean cost, RMSE, and σ_{RMSE} are in dollars per square foot. The GLM with gamma distribution is expected to have the best out-of-sample performance for its prediction task than any of the other models.

Model	Predicted cost	RMSE	σ_{RMSE}
Actual cost	36.03	-	-
GLM-Gamma	36.19	40.42	11.18
GLM-Normal	33.18	41.48	10.31
GLM-Inverse Normal	37.37	41.76	10.73
OLS	24.65	42.30	12.44

For ease of presentation, predictions, RMSE estimates, and σ_{RMSE} are given in dollars per square foot. OLS estimates in dollars per square foot are obtained by a naive application of the exponential function to predicted values. However, it should be noted that this naive transformation is biased (Moran et al., 2007).

The table also presents the actual mean structural retrofit cost in dollars per square foot from the training data (the same value from Table 2). The results of model evaluation suggest that the GLM with gamma distribution makes predictions closest to the true value. Moreover, the GLM with normal distribution tends to underestimate costs, while the GLM with inverse normal distribution tends to overestimate costs. OLS, on the other hand, severely underestimates costs.

For comparison, Table 9 presents predicted cost and prediction error on the log scale. The results in Table 9 suggest that the standard linear regression model has the lowest RMSE and thus is better at its prediction task than the GLMs are at their prediction tasks, in contrast to the results in Table 8. It might be tempting to conclude that the standard linear regression model, given in Eq. (1) is *better* than the GLM given in Eq. (2).

Table 9. Model evaluation: expected out-of-sample performance on log scale, when the outcome is structural cost. The results suggest that OLS provides the best out-of-sample performance in its task, predicting $E[\ln(Y)|X]$, than the other models in their tasks, predicting $\ln(E[Y|X])$.

Model	Predicted log cost	RMSE	σ_{RMSE}
Actual log cost	3.05	-	-
GLM-Gamma	3.39	1.06	0.06
GLM-Normal	3.07	1.07	0.07
GLM-Inverse Normal	3.41	1.08	0.06
OLS	2.98	0.97	0.04

Nevertheless, in addition to the fallacy that these RMSE estimates may be used for model selection, one must again be sure to use a model that is appropriate for the problem. While OLS may appear to be better, it may not actually fit the purpose of the problem.

In the present context, a GLM is recommended over the standard linear regression model because it is more easily interpretable in dollar terms. In particular, predictions of $E[Y|X]$ can be obtained easily from $\ln(E[Y|X])$ by the simple application of the exponential function, $\exp$. In contrast, predictions of $E[Y|X]$ cannot be obtained as easily from $E[\ln(Y|X)]$, though methods to transform back to the original dollar scale exist. Moran et al. (2007) presents and compares several methods, including the naive transformation.

The results illustrate not only how easily predictions can be obtained on the dollar scale, but also provide a sense of how well each GLM performs in predicting structural retrofit costs. In particular, note that the gamma and inverse normal models tend to slightly overestimate costs, while the normal model tends to underestimate costs. The gamma model's predictions have the lowest prediction error among the three models. In particular, the prediction error for the normal model is about 3% larger than that for the gamma model. In a later section, the application to federal buildings illustrates how this tradeoff between bias (accuracy) and variance (precision) manifests in practice.

TOTAL OR STRUCTURAL COST

Thus far, it is assumed that the outcome of interest is structural retrofit cost. However, a decision maker may be more interested in predicting the total construction cost. Table 10 presents estimates of expected out-of-sample performance of predicting each of these two outcomes, using the GLM with gamma distribution.

The results suggest that the GLM-gamma can predict structural cost more accurately than it can predict total cost (that is, with a roughly 60% lower RMSE). Moreover, note that the RMSE estimate for structural cost has a smaller standard deviation than the RMSE estimate for total cost, and thus the estimate of prediction error is less noisy.

Table 10. Predicted cost and expected out-of-sample performance for the GLM-Gamma in predicting structural construction cost and total construction cost. All values in dollars per square foot (1 ft = 0.3048 m). Predicted values and RMSE estimates suggest the GLM-Gamma is better at predicting structural cost than it is at predicting total cost.

Model	Acutal cost	Predicted cost	RMSE	σ_{RMSE}
Structural cost	36.03	36.19	40.42	11.18
Total cost	52.13	57.81	75.56	28.52

The results should be interpreted carefully. As in the preceding section, model selection is not appropriate because the outcomes are different. The results of model evaluation only suggest that predicting structural cost results in less uncertainty than predicting total cost: predicted values have lower RMSE and RMSE estimates have smaller variance.

The choice between structural and total cost will depend on the outcome of interest to the decision maker and the decision maker's objective for obtaining cost estimates. In the present context, predicting structural cost is recommended due to the lower uncertainty in prediction.

Structural cost estimates may be used as an order of magnitude approximation to cost in the planning stage. Estimates of total construction cost may be obtained by scaling strutural cost estimates up, as discussed in Fung et al. (2018). Table 11, which presents summary statistics for the ratio of total to structural cost, may be used as a reference for scaling up estimates. For instance, on average, total construction costs are double the structural costs.

Table 11. Summary statistics: ratio of total to structural construction costs in training (SRCE) data.

1st quartile	Mean	Median	3rd quartile	Max	s.d.
1	2.01	1.01	1.24	320	8.88

APPLICATION TO FEDERAL BUILDINGS

This section presents an application of the methodology for estimating seismic retrofit costs for a portfolio of buildings owned and leased by federal government agencies within the contiguous United States.

The application is motivated by Executive Order (EO) 13717, which asks "each executive department and agency...to enhance resilience by reducing risk to the lives of building occupants and improving continued performance of essential functions following future earthquakes." The estimates in this paper are not meant to be used for budget decisions. Rather, the paper presents a range of estimates that provide a sense of the expected order of magnitude, as well as the degree of uncertainty associated with the estimates.

The application illustrates the impact of an important modeling decision: the choice of distribution. The predictions presented in this section use a GLM with log link function, $g = \ln$, in order to easily present cost estimates in dollars per square foot. Moreover, the outcome of interest is assumed to be structural retrofit cost. The three choices of distribution for the outcome considered are the normal, the inverse normal, and the gamma.

THE FEDERAL BUILDING DATA

Data on federally-owned and -leased buildings is available to *federal employees only* from the General Services Administration (GSA) Federal Real Property Profile (FRPP) (GSA, 2018). The FRPP is a centralized database of the federal government's inventory of land, building, and structure assets located throughout the United States and abroad. Each agency submits data on its assets annually.

The goal of this application is to obtain retrofit cost predictions for an actual building portfolio, using FRPP building data obtained by the authors for Fiscal Year 2015 (FY15). Table 12 provides some summary statistics for the FY15 FRPP data, including average hazard level, total number of buildings, and average square footage by seismicity. The data only includes buildings within the contiguous United States.

Table 12. Total number of buildings, mean PGA, mean building area, percent of buildings owned (rather than leased) and percent of buildings deemed historic, by seismicity category (1 ft = 0.3048 m). Based on FRPP building data for FY15.

Seismicity	Total Buildings	Mean PGA (g_n)	Mean Area (sq ft)	Percent Owned	Percent Historic
L	100403	0.04	9956	87.4%	12.69%
M	12397	0.14	4935	92.08%	12.2%
H	8725	0.31	8045	89.68%	11.99%
VH	2930	0.47	12321	93.38%	12.46%

In addition, the table lists the percent of buildings that are owned by the reporting agency (versus those that are leased by the agency) and the percentage of buildings that are deemed historic. Buildings deemed historic are those for which the FRPP Historic Status indicator lists the building as either a National Historic Landmark (NHL), National Register Eligible (NRE), or National Register Listed (NRL) building (GSA, 2015). Otherwise, the building is deemed non-historic.

PROXIES FOR BUILDING AGE, HEIGHT, AND TYPE

Some of the building characteristics needed for the predictive model are not collected for the FRPP (that is, the FRPP does not ask for this information). In particular, the following key building characteristics are not collected for the FRPP: (1) Building age or year built; (2) Number of stories or building height; (3) Building construction type.

Nevertheless, reasonable predictions for retrofit costs can be obtained by using the data in the FRPP and making some assumptions about the data that is not available. In practice, building owners and other decision makers should be able to easily obtain more complete and accurate information on building characteristics and thus obtain more accurate predictions when applying the predictive modeling approach.

Fung et al. (2018) develop an approach for obtaining *proxies* for the predictors that are not collected in the FRPP. It should be noted that this approach is not advocated as part of the methodology; rather, it is one way to circumvent the data limitations.

Three disparate sources are used to proxy for building age, height, and type, as shown in Table 13. Note that while data on Age is available at the state level, data on Height and Type are only available at the Census and Hazus Region levels, respectively. Hazus categorizes the 50 states and the District of Columbia into three Hazus Regions (East, Midwest, West) (FEMA, 2012). Census categorizes the 50 states and the District of Columbia into four Census regions (Northeast, Midwest, South, West).

Table 13. Data sources and category values for building age, height, and type proxies. "Depends on" means these values should be known in order to determine the appropriate proxy.

Characteristic	Depends on	Values
Age	Census Region	{Pre-1950, 1950-1970, Post-1970}[1]
Height	Census Region	{Low-Rise, Mid-Rise, High-Rise}[2]
Type	Age, Height, Hazus Region	See Table 2[3]

[1] Source: Census, American Community Survey (ACS) 1-year estimates for 2010.
[2] Source: Energy Information Administraion, Commerical Buildings Energy Consumption Survey (CBECS) for 1999.
[3] Source: FEMA, Hazus 2.1, General Building Stock (GBS), Tables 3A.2-3A.15 FEMA (2012).

It is worth noting that the proxy for general building age is based on *housing* age. While imperfect, the Census data on housing age is the most comprehensive source for building age that covers the entire United States.

Proxies for age, height, and type are drawn from a sampling distribution as suggested in Fung et al. (2018). For a given building in the FRPP with observed characteristics x, sample the unobserved characteristics $z = \{\text{Age, Height, Type}\}$ as:

$$\text{Age, Height, Type}|x \sim p(\text{Type}|\text{Age, Height}, x)p(\text{Height}|x)p(\text{Age}|x) \tag{5}$$

where $p(Z|x)$ represents the distribution of random variable Z conditional on $X = x$. Eq. (5) represents the joint distribution of Age, Height, and Type, conditional on x. In this case, the relevant x is the building location (e.g., Census region or Hazus region).

Since these features can be sampled at random, the procedure is repeated 1000 times. Thus, for each building, building age, height, and type are sampled 1000 times, resulting in 1000 "pseudo"-inventories. Predictions are generated for each "pseudo"-inventory. As a result of this sampling procedure, prediction intervals are easily obtained by computing empirical quantiles for the predictions.

IS THERE A PENTALTY FOR DISCRETIZING AGE AND HEIGHT?

The proxies for building age and height in Table 13 are *categorical*; that is, age and height are grouped into a small number of distinct categories. A natural question is whether there is a penalty to using categorical, rather than continuous, measures of age and height. This section compares prediction error for: (1) continuous building age and height; and (2) categorical building age and height.

Table 14 presents prediction error estimates, using K-fold cross-validation with $K = 10$, assuming the outcome of interest is structural cost. The results suggest that there is almost no penalty (in terms of prediction error) for using categorical, rather than continuous, age and height. Although the results suggest using categorical measures may actually *improve* performance, caution should be taken in light of the large standard error associated with prediction error estimates.

Table 14. Model evaluation: expected out-of-sample performance for GLM with three choices of distribution when the outcome is structural cost in dollars per square foot (1 ft = 0.3048 m). RMSE estimates suggest each model performs better when predicting with categorial rather than continuous building age and height.

Model	Distribution	RMSE	σ_{RMSE}
categorical	Gamma	40.69	13.86
continuous	Gamma	40.77	14.27
categorical	inverse	41.24	13.27
continuous	inverse	41.31	13.90
categorical	Normal	41.47	13.59
continuous	Normal	42.01	13.26

COST ESTIMATES FOR FEDERAL BUILDINGS

This section presents structural retrofit cost estimates using proxies for building age, height, and type, as well as 95% prediction intervals. Predictions are based on the GLM with gamma distribution, which appears to outperform the other models based on the preceding results. All costs normalized to 2016 US dollars.

Table 15 presents cost estimates for each of the four seismicity levels in Table 3. Note that costs for VH-seismicity buildings are the highest, and prediction intervals are the widest. Interestingly, retrofit-cost predictions for L-seismicity buildings are slightly higher than those for either M- or H-seismicity buildings. The results appear counterintuitive at first glance: one might expect that the cost of retrofitting L-seismicity buildings would be the lowest, on average.

However, retrofit cost is a function of multiple unobserved parameters, including: the building code the existing building is designed for; the target performance in the new code; and the existing versus the desired seismic detailing. The pattern of average retrofit costs being higher for L-seismicity buildings than for M- and H-seismicity buildings reflects the pattern in the training data, as shown in Table 16.

Table 15. Predicted average structural cost and 95% prediction intervals in dollars per square foot (1 ft = 0.3048 m), with proxies for building age, height, and type, by seismicity category for GLM with gamma distribution.

Seismicity	Lower bound	Mean cost	Upper bound
L	12.47	24.96	43.10
M	10.56	20.07	35.81
H	10.40	19.82	35.32
VH	16.42	31.03	55.69

Table 16. Average structural cost in dollars per square foot (1 ft = 0.3048 m), as well as the top and bottom 2.5% of costs, in the training (SRCE) data. Note the pattern of average retrofit costs being higher for Low seismicity buildings than for Medium and High seismicity buildings.

Seismicity	Percentile: 2.5%	Mean	Percentile: 97.5%
L	2.18	29.4	92.9
M	1.22	27.9	97.1
H	1.90	25.1	120.4
VH	2.54	55.0	227.4

One takeaway for decision makers looking for a way to prioritze which buildings to retrofit first is that building seismicity is an important driver of costs. The pattern reflected in the data suggests that building seismicity is a proxy for the unobserved parameters that determine retrofit cost.

CONCLUSION

This paper presents a predictive modeling approach to estimating seismic retrofit costs from historical data. The predictive model can be used to obtain quick, order of magnitude estimates. However, obtained estimates are expected to have a higher degree of uncertainty than professional engineering consulting estimates. Moreover, while the approach is applicable to estimating retrofit costs for a single building, the high degree of uncertainty makes it more applicable for a portfolio of buildings.

Several modeling choices are available to decision makers. First, the paper explores the choice of distribution for the outcome of interest and suggests that a gamma distribution is used in the context of predicting costs. Second, with regard to the choice between standard linear regression and GLM, the recommendation is a GLM because predictions can be easily expressed in dollars per square foot. Third, the choice of total construction cost or structural retrofit cost for the outcome of interest will depend on the decision maker's objective. The lower degree of uncertainty in predicting structural retrofit costs motivates its recommendation as the preferred outcome of interest.

The application to an actual building portfolio illustrates how modeling choices affect cost estimates. In particular, the GLM with gamma distribution appears to provide better out-of-sample performance than the GLM with normal or inverse normal distributions, regardless of whether building age and height are categorical or continuous. The application illustrates an approach for obtaining proxies for predictors that are unavailable and produces cost estimates by seismicity category.

The results demonstrate the flexibility and applicability of the predictive modeling approach for seismic risk mitigation. In particular, the illustration of key modeling decisions and the tradeoffs associated with those decisions should be valuable for owners of building portfolios during the planning phase of a potential seismic retrofit program. An important modeling decision that is not addressed in this paper is the question of how to choose predictors for the model. This problem is known as *feature selection* and is beyond the scope of the current paper; Fung et al. (2017) and Fung et al. (2019) provide a more thorough treatment of the feature selection problem for predicting seismic retrofit costs.

Finally, this paper only considers construction costs. In addition to the construction costs mentioned in the section "The Training Data," retrofits are likely to incur other costs, including the costs to relocate building occupants, project financing costs, and the costs of disrupting work. The incorporation of these other costs is important for a complete picture of retrofit costs and should be studied in the future.

Disclaimer

NIST policy is to use the International System of Units (metric units) in all its publications. In this report, however, information is presented in U.S. Customary Units (inch-pound), as this is the preferred system of units in the U.S. earthquake engineering industry.

REFERENCES

Arlot, S. and Celisse, A., 2010. A survey of cross-validation procedures for model selection. *Statist. Surv.* **4**, 40–79. doi:10.1214/09-SS054.

ATC, 1978. *Tentative provisions for the development of seismic regulations for buildings: a cooperative effort with the design professions, building code interests, and the research community*, vol. 510. Department of Commerce, National Bureau of Standards.

Coxe, S., West, S. G., and Aiken, L. S., 2013. Generalized linear models. In Little, T. D. (ed.), *The Oxford Handbook of Quantitative Methods, Vol 2: Statistical Analysis*, pp. 26–51. Oxford University Press.

Cremen, G. and Baker, J. W., 2019. A Methodology for Evaluating Component-Level Loss Predictions of the FEMA P-58 Seismic Performance Assessment Procedure. *Earthquake Spectra* **35**, 193–210.

Del Vecchio, C., Di Ludovico, M., Pampanin, S., and Prota, A., 2018. Repair costs of existing RC buildings damaged by the L'Aquila earthquake and comparison with FEMA P-58 predictions. *Earthquake Spectra* **34**, 237–263.

Di Ludovico, M., Prota, A., Moroni, C., Manfredi, G., and Dolce, M., 2017a. Reconstruction process of damaged residential buildings outside historical centres after the LAquila earthquake: part IIheavy damage reconstruction. *Bulletin of Earthquake Engineering* **15**, 693–729.

Di Ludovico, M., Prota, A., Moroni, C., Manfredi, G., and Dolce, M., 2017b. Reconstruction process of damaged residential buildings outside historical centres after the LAquila earthquake: part I" light damage" reconstruction. *Bulletin of Earthquake Engineering* **15**, 667–692.

ENR, 2017. Engineering News Record: Historical Indices. `http://www.enr.com/economics/historical_indices`.

FEMA, 1994. *Typical Costs for Seismic Rehabilitation of Existing Buildings, Volume 1: Summary*. *Tech. Rep. FEMA 156*, Federal Emergency Management Agency.

FEMA, 1995. *Typical Costs for Seismic Rehabilitation of Existing Buildings, Volume 2: Supporting Documentation*. *Tech. Rep. FEMA 157*, Federal Emergency Management Agency.

FEMA, 2012. *Hazus Earthquake Model Technical Manual*, 2.1 edn.

FEMA, 2013–2014. SRCE: Seismic Rehabilitation Cost Estimator. `https://www.fema.gov/media-library/assets/documents/30220`.

Fung, J., Butry, D., Sattar, S., and McCabe, S., 2017. *A Methodology for Estimating Seismic Retrofit Costs*. *Tech. Rep. NIST 1973*, National Institute of Standards and Technology.

Fung, J., Butry, D., Sattar, S., and McCabe, S., 2018. *Estimating Structural Seismic Retrofit Costs for Federal Buildings*. *Tech. Rep. NIST 1996*, National Institute of Standards and Technology.

Fung, J., Sattar, S., Butry, D., and McCabe, S., 2019. Selecting Building Characteristics for Predictng Seismic Retrofit Costs of a Building Portfolio. In *Proceedings of the 2nd International Conference on Natural Hazards & Infrastructure*. National Technical University of Athens. Forthcoming.

GSA, 2015. *Federal Real Property Council: 2015 Guidance for Real Property Inventory Reporting*, version 3 edn.

GSA, 2018. FRPP MS: Federal Real Property Profile Management System. `https://www.realpropertyprofile.gov/FRPPMS`.

Guyon, I. and Elisseeff, A., 2003. An introduction to variable and feature selection. *Journal of machine learning research* **3**, 1157–1182.

Jafarzadeh, R., Ingham, J., Wilkinson, S., González, V., and Aghakouchak, A., 2013a. Application of Artificial Neural Network Methodology for Predicting Seismic Retrofit Construction Costs. *Journal of Construction Engineering and Management* **140**, 04013044.

Jafarzadeh, R., Ingham, J. M., and Wilkinson, S., 2014. A Seismic Retrofit Cost Database for Buildings with a Framed Structure. *Earthquake Spectra* **30**, 625–637.

Jafarzadeh, R., Wilkinson, S., Gonzalez, V., Ingham, J., and Amiri, G. G., 2013b. Predicting Seismic Retrofit Construction Cost for Buildings with Framed Structures Using Multilinear Regression Analysis. *Journal of Construction Engineering and Management* **140**, 04013062.

James, G., Witten, D., Hastie, T., and Tibshirani, R., 2013. *An introduction to statistical learning*, vol. 112. Springer.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S., 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics* **6**, 10.

Moran, J. L., Solomon, P. J., Peisach, A. R., and Martin, J., 2007. New models for old questions: generalized linear models for cost prediction. *Journal of evaluation in clinical practice* **13**, 381–389.

Nasrazadani, H., Mahsuli, M., Talebiyan, H., and Kashani, H., 2017. Probabilistic Modeling Framework for Prediction of Seismic Retrofit Cost of Buildings. *Journal of Construction Engineering and Management* **143**, 04017055.

USGS, 2014. 2014 National Seismic Hazard Maps. `https://earthquake.usgs.gov/hazards/hazmaps/`.

APPENDIX

This appendix presents supplementary material. Table A1 summarizes the shares of each building group in the training (SRCE) data.

Table A1. Building groups and their shares in SRCE data.

Building Group	Count	Percentage
1	459	30.08%
2	87	5.70%
3	102	6.68%
4	357	23.39%
5	74	4.85%
6	39	2.56%
7	107	7.01%
8	301	19.72%

Cofficient estimates obtained by training the models on the entire training data are presented in Table A2. The table presents coefficient estimates from training the standard linear regression model, Eq. (1), and the GLM with gamma distribution, for the two outcomes of interest: structural retrofit cost and total construction cost.

The coefficient estimates represent the estimator, $\hat{f}$, for each model and thus may be applied to obtain predictions, $\hat{Y} = \hat{f}(X_{new})$, for a set of predictors, X_{new}, representing a new building.

These coefficient estimates may only be applied to obtain predictions if the data in X_{new} has the same structure as the data used to train the model in this paper. Most importantly, the measure of seismicity for the new building should coincide with the measure of seismicity used to train the models: a population-weighted average of county-level PGA as described in Fung et al. (2017). If a decision maker would like to use a different measure of seismicity, the same procedure presented in this paper can be used.

Moreover, the results in this paper are obtained by training the models on the raw SRCE data, with costs normalized to 1993 California dollars. Predictions obtained from the trained models are then normalized to 2016 national dollars using the ENR Building Construction Index (ENR, 2017). This paper uses the index value $BCI_{2016} = 1.669$, as described in Fung et al. (2017).

Table A2. Coefficient estimates from training the OLS model in Eq. (1) and the GLM Eq. (2), for the two outcomes of interest. Standard errors in parentheses.

	Structural cost per sf	Structural cost per sf	Total cost per sf	Total cost per sf
	OLS	*glm: Gamma link = log*	*OLS*	*glm: Gamma link = log*
Area (β_1)	−0.182 (0.030)	−0.135 (0.029)	−0.105 (0.032)	−0.073 (0.039)
Age (β_2)	0.126 (0.067)	0.073 (0.066)	0.188 (0.072)	0.143 (0.088)
Stories (β_3)	0.266 (0.053)	0.204 (0.052)	0.266 (0.057)	0.265 (0.069)
Occupancy: TR (β_4)	0.218 (0.074)	0.152 (0.073)	−0.204 (0.080)	−0.252 (0.097)
Occupancy: IP (β_4)	−0.301 (0.094)	−0.249 (0.093)	−0.659 (0.101)	−0.596 (0.122)
BG: 2 (β_b)	−0.693 (0.141)	−0.423 (0.139)	−0.214 (0.151)	0.011 (0.183)
BG: 3 (β_b)	−0.445 (0.147)	−0.336 (0.145)	−0.314 (0.158)	−0.199 (0.192)
BG: 4 (β_b)	0.182 (0.103)	0.192 (0.102)	0.220 (0.111)	0.177 (0.135)
BG: 5 (β_b)	0.034 (0.159)	0.126 (0.157)	0.135 (0.171)	0.194 (0.208)
BG: 6 (β_b)	−0.924 (0.201)	−0.769 (0.199)	−0.833 (0.216)	−0.781 (0.263)
BG: 7 (β_b)	0.369 (0.139)	0.253 (0.137)	0.322 (0.149)	0.231 (0.181)
BG: 8 (β_b)	−0.216 (0.107)	−0.018 (0.106)	−0.065 (0.115)	0.174 (0.140)
Seismicity: M (β_s)	−0.040 (0.185)	−0.081 (0.183)	−0.037 (0.199)	−0.038 (0.242)
Seismicity: H (β_s)	−0.087 (0.185)	−0.189 (0.182)	−0.106 (0.198)	−0.065 (0.241)
Seismicity: VH (β_s)	0.423 (0.181)	0.314 (0.179)	0.386 (0.194)	0.465 (0.236)
Performance: DC (β_p)	−0.069 (0.193)	−0.151 (0.190)	0.006 (0.207)	0.011 (0.251)
Performance: IO (β_p)	0.133 (0.201)	−0.051 (0.199)	0.052 (0.216)	−0.121 (0.263)
Historic (β_5)	0.564 (0.109)	0.792 (0.108)	0.910 (0.117)	0.992 (0.143)
Census Region: Midwest (β_6)	−0.148 (0.151)	−0.139 (0.149)	−0.131 (0.162)	−0.065 (0.197)
Census Region: Northeast (β_6)	−0.110 (0.169)	−0.154 (0.167)	0.003 (0.182)	0.004 (0.221)
Census Region: South (β_6)	0.537 (0.137)	0.476 (0.136)	0.553 (0.147)	0.505 (0.179)
M x DC (β_{sp})	0.041 (0.250)	0.096 (0.246)	0.465 (0.268)	0.346 (0.325)
H x DC (β_{sp})	0.351 (0.255)	0.624 (0.252)	0.221 (0.274)	0.300 (0.333)
VH x DC (β_{sp})	0.200 (0.233)	0.343 (0.230)	0.186 (0.250)	0.044 (0.304)
M x IO (β_{sp})	0.090 (0.315)	0.103 (0.311)	−0.048 (0.338)	0.091 (0.411)
V x IO (β_{sp})	0.425 (0.296)	0.518 (0.292)	0.409 (0.318)	0.318 (0.386)
VH x IO (β_{sp})	0.372 (0.237)	0.584 (0.234)	0.717 (0.255)	0.605 (0.310)
Constant (β_0)	3.420 (0.440)	3.600 (0.434)	2.840 (0.472)	3.100 (0.573)
Observations	1,083	1,083	1,083	1,083

Notes: 'BG' means building group.
The terms 'M x DC,' ..., 'VH x IO' are interactions between seismicity and performance objective.
1 ft = 0.3048 m.

To illustrate, assume a decision maker has data for a new building X_{new} that conforms with the assumptions used to train the models in Table A2. Suppose Area = 1000, Age = 20, and Stories = 5. Moreover, suppose Occupancy is V, the building group (BG) is 1, Seismicity is L, the performance objective is LS, the building is *not* deemed historic (i.e., Historic = No), and the Census Region is the West. Thus, the coefficients for these latter predictors are all 0 and the prediction is based on:

$$X_{new}\hat{\beta} = \hat{\beta}_0 + \ln(\text{Area})\hat{\beta}_1 + \hat{\beta}_2 \ln(\text{Age}) + \hat{\beta}_3 \ln(\text{Stories}) \tag{6}$$

In particular, the predicted average structural cost (in dollars per square foot) using the GLM with gamma distribution is $\widehat{E[Y|X_{new}]} = \exp(X_{new}\hat{\beta}) = \exp\{3.6 - 0.135\ln(1000) + 0.073\ln(20) + 0.204\ln(5)\} \times BCI_{2016}$ = 24.89002 $\times BCI_{2016} =$ 41.55.

Finally, the coefficient estimates presented in this appendix should not interpreted as representing the "true" model of retrofit costs. Rather, they are the outcome of training the models in Eq. (1) and Eq. (2) for the specific purpose of making retrofit cost predictions as accurately as possible from observable building characteristics.