CAN BIG DATA APPROACHES HELP EARTHQUAKE ENGINEERING IN UNDERDEVELOPED COUNTRIES?

In Ho Cho, Ikkyun Song, and Raymond K. W. Wong
Iowa State University
Ames, Iowa, USA

Abstract

Traditionally, innovations in earthquake engineering result from experimental researches and related in-depth analyses. However, for underdeveloped countries, such approaches may not be directly applicable due to financial issues and different practices. This paper suggests a Big Data approach by which expensive test-based processes may be replaced or largely complemented by data-driven knowledge-harvesting processes. We denote "Big Data approach" herein as a purely large data-driven research paradigm rooted in advanced computing power and strong data analyses, which have little restriction to the real-world database's pathological problems of immense data size, sparseness, biasness, and a multitude of variables. By virtue of persistent efforts of global earthquake engineering community, decades-long experimental databases are becoming available to global researchers. Following the Big Data approach, we predict complex responses of the community database by use of the so-called generalized additive model (GAM), which enables flexible and reliable predictions with multiple variables. We showed that the advanced computing power integrated with GAM is a key enabling factor of the big data approach, and also asserted easy interpretability and promising prediction capabilities of GAM in applications to a wide spectrum of RC shear walls in the database. Also, we demonstrate the prediction capability of GAM based on two popular choices of smoothers (cubic spline regression and thin plate regression spline) and present interesting performance comparisons, guideline and limitations. Results suggest that if researchers of underdeveloped countries adopt the Big Data approach to their database or practice, they may better achieve cost-effective earthquake engineering, code renovations and beyond.

Introduction

Data-driven research is becoming a promising next-generation research paradigm (e.g., drug design (Ebejer et al. 2013), seismology (Bozorgnia et al. 2014), and even cosmology (Kamdar et al. 2016)). Notable advances in computing power enable researchers to apply such trends to extremely large data, dubbed Big Data hereafter. Big data are often characterized by big volume, large velocity (fast changes), and complex relations among many variables. In relation to natural hazards, global earthquake engineering research communities established important foundations for such data-driven discoveries (e.g., *NEEShub* (Hacker et al. 2011); NSF Cyber-infrastructure for natural hazards (Rathje et al., 2015)). However, data have not been actively used to improve the predictive and preventive ability of the earthquake engineering fields. Often, identification of problematic structural issues occurs at the post-disaster phase. Real experiment-based approaches are instrumental for understandings complex interplay among structural variables and performance variables, but limited financial resources may pose critical challenges to underdeveloped countries. Furthermore, substantial biasness, sparseness, and missing values of database must be overcome (Figure 1). This paper seeks to aid researchers in underdeveloped countries by providing an advanced non-parametric statistical technique, which is suitable for big data in our fields. With existing reinforced concrete shear wall (RCSW) database (Figure 2), we demonstrate how data can be used to rigorously "predict" untested structures' responses and to unravel the hidden significances of variables, notably without any prejudice. In particular, this paper expounds upon an advanced, non-parametric technique called generalized additive model (GAM), e.g. Hastie and Tibshirand (1990). GAM holds excellent accuracy and efficiency, and allows remarkable flexibility in terms of the distributions of the response variable and its relationship to the predictor variables. This paper supports the proposed

statistical approaches by providing systematic investigations and comparisons against real experimental results and high-precision computer simulations. We hope that earthquake engineering community in underdeveloped countries will substantially benefit from the novel capabilities of the proposed statistical approach. If prudently applied to their own databases, underdeveloped countries' researchers will be able to analyze, identify and even predict critical issues of their earthquake engineering practices, quickly and cheaply. Thereby, the proposed methods will substantially complement typical earthquake engineering in underdeveloped countries and also in the developed countries.

Difference from Traditional Statistical Methods. The notable difference of the proposed big data-driven approach from previous methods is twofold: first, the proposed statistical prediction allows "unspecified relationships" among variables of database, and the learning process is solely based on the raw data with a flexible additive model assumption. Second, the present statistical learning and prediction tasks have little restriction to the number of variables. Traditionally, statistical methods are usually used to confirm a researcher's pre-defined relationship. However, the proposed statistical approach assumes no pre-defined relations, but seeks to find the hidden relationship of variables and significance of variables.

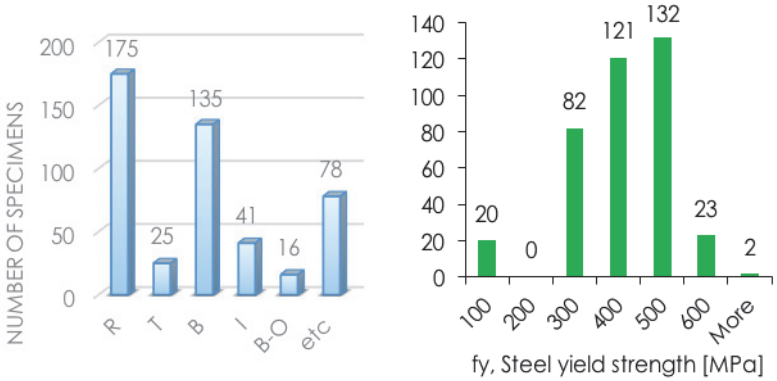

Figure 1. Sparseness and biasness revealed from 470 real experiments of RC shear wall database (from NEESHub and international literature).

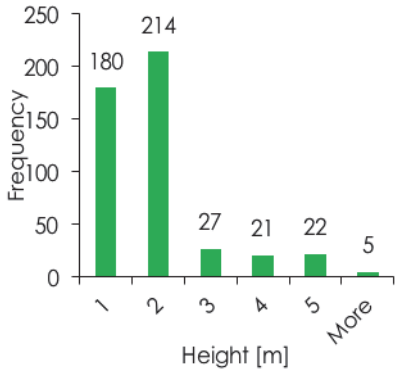

Figure 2. Type of RCSW (R: Rectangular; B: Barbell-; B-O: Barbell-shaped with opening; T- and I-shaped; etc: all other RC walls).

Summary of Statistical Theory

A generalized additive model (GAM) (Hastie and Tibshirani 1990) is a non-parametric extension of the well-known generalized linear model in which covariates enter the model through unspecified smooth functions. The general form of GAM can be represented as:

$$g(\mu_i) = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \cdots \quad (1)$$

where g is a smooth link function; $\mu_i \equiv \mathrm{E}(Y_i \mid x_i)$, Y_i is a response variable and from some exponential family of distribution (e.g., normal, binomial, or gamma distribution); x_i is i^{th} vector of data points comprising multiple variables, $x_i = \{x_{1i}, x_{2i}, \ldots\}$; f_i are smooth functions of covariates x_{ji} (Wood 2006). For instance, Y_i would be the maximum shear force of i^{th} RCSW specimen, and then x_i may be {length$_i$, height$_i$, AxialForce$_i$, …}. The GAM depends on sum of unspecified smooth functions rather than pre-specified forms of x_i, which imparts unique flexibility to GAM. To glean the central notion of the GAM, the following descriptions involve one variable and normal distribution case (extensions are straightforward as in Wood, 2006). Then, the GAM is $\mathrm{E}(Y \mid x) = f(x)$, and the smooth function f can be approximated as

$$f(x) = \sum_{j=1}^{q} b_j(x)\beta_j \tag{2}$$

where $b_j(x)$ is the jth basis function and β_j is an unknown parameter. Fitting the model can be done by maximizing the corresponding likelihood with a penalty term given by $\lambda \int \left[f''(x) \right]^2 dx$ where λ is smoothing parameter. Too large λ value leads to an over-smoothed estimate while too small λ value results in an under-smoothed estimate. We can choose λ value which enables to fit model appropriately by minimizing generalized cross validation (GCV) score, V_g (Golub et al. 1979):

$$V_g = \frac{n \sum_{i=1}^{n} \left(y_i - \hat{f}_i \right)^2}{\left[tr(I - A) \right]^2} \tag{3}$$

where $\hat{f}$ is the estimate from fitting to all the data, A is the corresponding smoothing matrix (Wood 2006). It should be noted that the relevant library in R automatically optimizes λ, and thus explicit control on λ is generally unnecessary. A basis for spline should be chosen to construct a GAM. There are two popular types of basis used in GAM: (a) Thin plate regression splines (TPRS) (Wood 2003) and (b) Cubic regression spline (CRS) (Wood 2006). TPRS is suitable for any number of covariates, and is notably "knot-free" (i.e. requiring no knot-location selection). Yet, CRS requires the knot location selection and is restricted to a single variable. Generally, TPRS requires more computational cost than CRS. As an illustrative example, Fig. 3 compares four regression models (Black = Linear; Red = Parabolic; Green = GAM&CRS; Yellow = GAM&TPRS) with 470 real RCSW data. Fig. 3 presents a good example of the flexibility of GAM applied to the complex real-world database. On one hand, cubic spline is a curve constructed by combining a number of cubic polynomial sections. Those sections join at a certain point, called "knot," of which location should be pre-selected for the cubic spline basis. The cubic polynomial sections are joined such that the entire spline becomes continuous up to second derivative. One the other hand, thin plate spline (Duchon 1977) can be used for multiple covariates. Thin plate spline function, f, can be obtained by minimizing

$$\| y - f \|^2 + \lambda J_{md}(f) \tag{4}$$

where y is the vector of y_i data and $f = [f(x_1), f(x_2), \ldots, f(x_3)]^{\mathrm{T}}$. $J_{md}(f)$ is a penalty functional measuring the "wiggliness" of f, and λ is a smoothing parameter controlling the tradeoff between data fitting and smoothness of f. In Eq. (5), $\hat{f}$ is the function that can minimizes Eq. (4). As marked in Eq. (5), the first terms are related to wiggliness while the second terms are independent of wiggliness.

$$\hat{f}(x) = \underbrace{\sum_{i=1}^{n} \delta_i \eta_{md}\left(\| x - x_i \| \right)}_{\textit{wiggly components}} + \underbrace{\sum_{j=1}^{M} \alpha_j \phi_j(x)}_{\textit{Zero wiggly terms}} \tag{5}$$

where δ_i and α_j are coefficients to be determined, ϕ_j are linearly independent polynomials spanning the null space of J_{md}, and η_{md} are the basis functions. Thin plate regression splines seek to find the balance by reducing the wiggly components of Eq. (5) and retaining the zero wiggly terms in Eq. (5). In this fashion, thin plate regression splines are regarded as a powerful method that has little restriction to the burdensome "knot location" determination and many variables. For detailed formulations, see (Wood 2006).

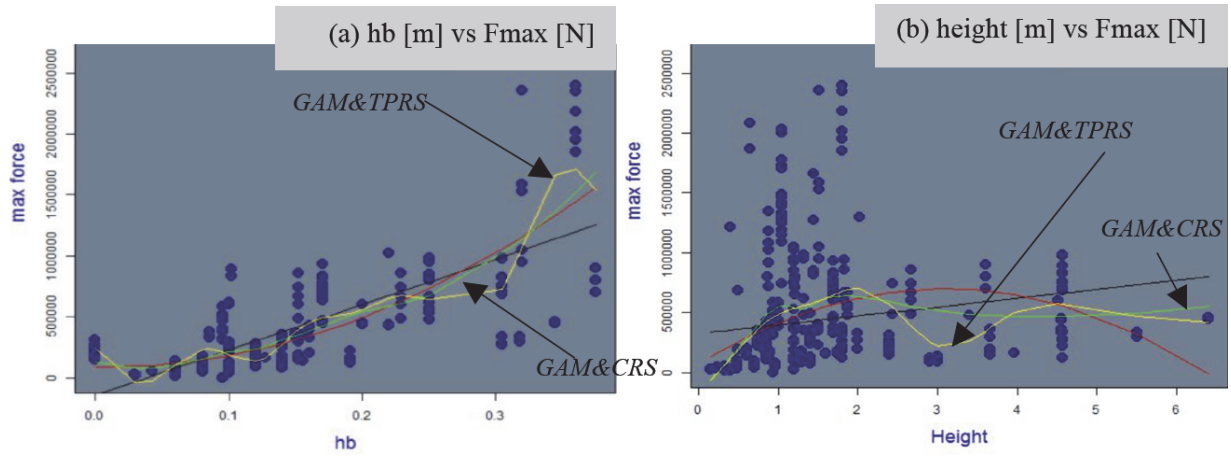

Figure 3. Example of one-dimensional regressions of 470 real RC wall data: (a) hb (thickness of boundary element) versus Fmax; (b) wall height versus Fmax. [Regression type: Black = Linear; Red = Parabolic; Green = GAM&CRS; Yellow = GAM&TPRS].

Metrics for Prediction Comparisons. In this study, we adopted several metrics to quantitatively compare the prediction performances: CVE/CVE$_b$ = Ratio between cross validation error (CVE) and base cross validation error (CVE$_b$); ρ = Pearson correlation; R^2= Coefficient of determination. These were adopted to measure how accurately the GAM fits real data after leaning. This choice of metrics is based on the comparable study on machine learning (Kamdar et al. 2016). In short, the higher CVE/CVE$_b$, ρ , or R^2, the more accurate prediction. In particular, CVE is defined as the summation of $(y^i_{experiment} - y^i_{predicted})^2/N$ over i =1, ⋯ , N, where N is number of data, $y^i_{experiment}$ is the i^{th} measured response, and $y^i_{predicted}$ is i^{th} predicted response according to the cross-validation procedure. The base CVE$_b$ is defined as the summation of $(y^i_{experiment} - y^i_{mean,predicted})^2/N$ over i =1, ⋯ , N, where $y_{mean,predicted}$ is mean of predicted values. ρ is defined as $cov(y_{predicted}, y_{experient})/(\sigma_{ypredicted} \times \sigma_{yexperiment})$. R^2 is given by $1 - \sum_{i=1}^{N} (y^i_{experiment} - y^i_{mean,predicted})^2 / \sum_{i=1}^{N} (y^i_{experiment} - y^i_{predicted})^2$. For comparison among different statistical models, we use CVE/CVE$_b$ as a primary metric and also refer to other metrics as reference.

Prediction with GAM

To demonstrate the predictive power of GAM, we used the existing database of RCSW experiments. This study assumes the Gamma distribution in light of the domain-specific nature of data (i.e., real, positive values in existing values). For the link function among variables, we chose *log* since it can easily incorporate multiplicative connections of engineering variables. For the smooth function-related parameter, k (i.e. the number of basis dimensions in smooth functions), we adopted 7 throughout the statistical studies as done by by (Wood 2006). We began with raw data of 10 variables, notably with no prejudices regarding relations or relative significance of these variables. The 10 variables of the existing RCSW database are: axial force ratio (denoted by afr), wall thickness (thickness), boundary element's thickness (hb) and boundary element's width (bb), wall height (height), wall length (length), primary

reinforcing bar's yield strength (fy) and diameter (dia), concrete compressive strength (fc), and boundary element's reinforcement ratio (bderr). Target response is the maximum shear resistance, F_{max}.

Like other prediction methods (e.g., Machine Learning), key procedure is "cross-validation" which consists of three tasks: (1) Exclusion of one real wall specimen, (2) Construction of a GAM by learning the remaining walls, and (3) Prediction of the test (excluded) wall's response using the constructed GAM. These processes are repeated throughout all walls. The difference between the predicted F_{max} using GAM and the original value of the omitted wall specimen directly represents how precisely GAM can predict the target response. For the systematic presentation of prediction power, the so-called Q-Q plots were used to compare the scaled response of real experiments and predicted values (see Figure 4). In the Q-Q plot, the more straight linear path means the more precise prediction.

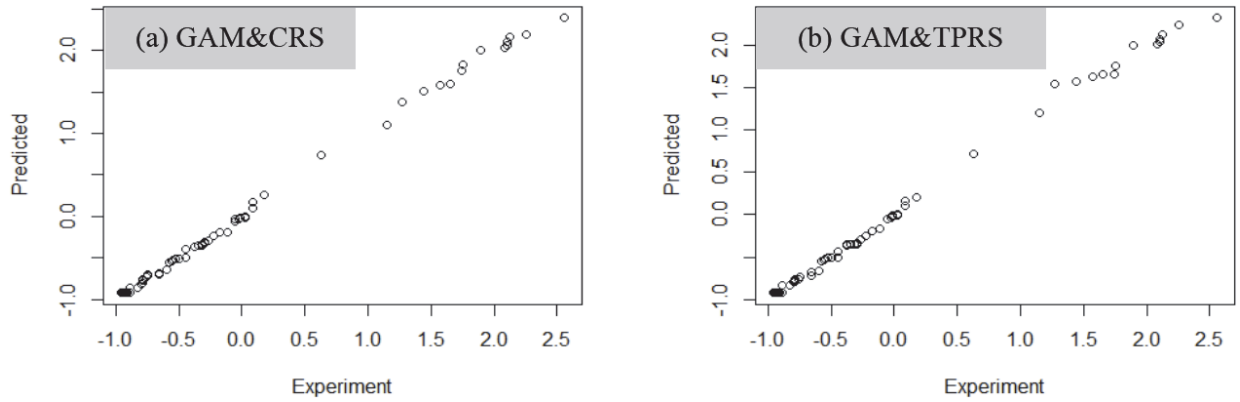

Figure 4. Q-Q plot of real test data and the predicted value using (a) GAM-CRS; (b) GAM-TPRS.

Importantly, although the statistical models use no prejudices or weighting factors, the predicted responses using two GAMs show good agreements with real experimental data (Fig. 4). The promising prediction accuracy is commonly found in both GAM-CRS and GAM-TPRS. It should be noted that all the statistical predictions in Fig. 4 are made by the "best" statistical models, which are explained what follows.

Constructing a Best GAM with a Given Number of Variables. Selection of variables is critical for prediction models, yet without any prejudice, we are uncertain which variables should be included in the GAM. To address this issue, we explain here how to find a best GAM model that can most accurately predict the target response with a given number of variables. We first constructed all possible combinations of variables. Each case, we compared CVE/CVE_b to determine the best combination. Tables 1 and 2 summarize the best combination of a given number of variables: e.g., amongst all two-variable combinations, GAM-CRS selects height and hb (second row of Table 1) as the best combination. These comparisons are focusing on only the prediction accuracy of the given statistical setting and assumptions. Overall, Table 1 and Figure 5 show that the best combination for GAM-CRS is the combination of six variables (marked by bold letter). In Table 2 and Figure 5, the best combination for GAM-TPRS has the seven variables (marked by bold letter). Interestingly, axial force ratio is constantly identified as the statistically important variable: i.e. the present data-driven learning pinpoints the same issue raised by many researchers' in-depth investigations (e.g., Wallace et al. 2012; Westenenk et al. 2012). This bears out the promising possibility of the systematic data-driven investigation.

Table 1. Best combination selection using GAM-CRS

NV	*NC*	*Best combination of variables (in each row, ordered by increasing p-value)*			*CVE/ CVE_b*	*Pearson*	R^2
2	45	height	hb		12.24	0.958	0.918
3	120	height	hb	dia	16.39	0.969	0.939
4	210	height	afr	hb dia	21.00	0.976	0.952
5	252	height	afr	dia hb fc	22.46	0.978	0.955
6	**210**	**afr**	**thickness**	**hb height fy dia**	**26.21**	**0.981**	**0.962**
7	120	afr	thickness	hb height fy dia fc	25.75	0.981	0.961
8	45	afr	height	fy bb length thickness dia hb	24.64	0.980	0.959
9	10	afr	height	fy bb thickness length dia hb fc	23.61	0.979	0.958
10	1	afr	height	fy bb length thickness dia hb bderr fc	4.63	0.918	0.784

(NV: number of variable; NC: number of total combinations)

Table 2. Best combination selection using GAM-TPRS

NV	*NC*	*Best combination of variables (in each row, ordered by increasing p-value)*			*CVE/ CVE_b*	*Pearson*	R^2
2	45	length	height		12.22	0.958	0.918
3	120	length	dia	afr	15.70	0.968	0.936
4	210	length	height	afr dia	20.89	0.976	0.952
5	252	afr	thickness	fy bderr length	23.32	0.978	0.957
6	210	afr	thickness	fy bderr length fc	22.92	0.978	0.956
7	**120**	**afr**	**height**	**fy thickness dia hb length**	**24.33**	**0.979**	**0.959**
8	45	afr	height	fy length thickness hb dia bderr	22.97	0.979	0.956
9	10	afr	length	bb fy height bderr dia thickness fc	23.93	0.979	0.958
10	1	afr	bb	height fy hb dia length thickness bderr fc	14.88	0.968	0.933

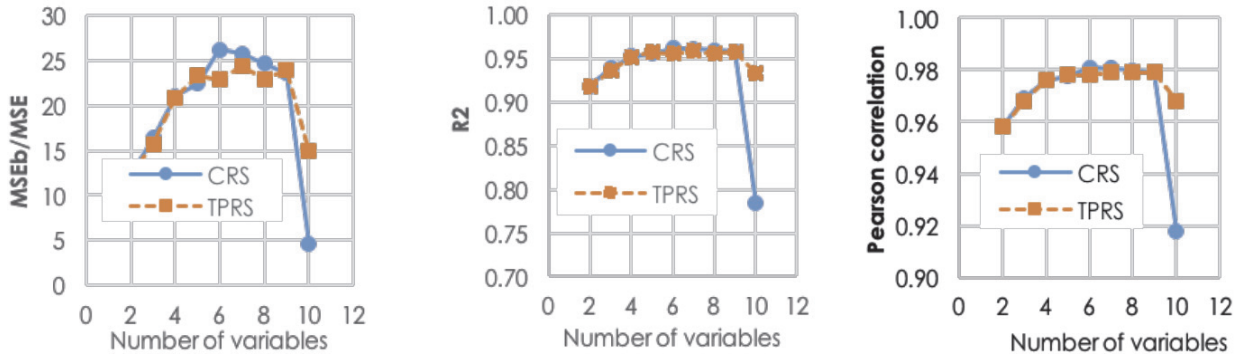

Figure 5. Variation of metrics used for best combination of variables.

Statistical Prediction Versus High-Precision Computer Simulations

There are important analogy and difference between the statistical prediction and high-precision computer simulations. Global researchers have established high-fidelity computational simulation platforms (e.g., *OpenSees* (McKenna et al. 2000); Orakcal and Wallace 2006; Vecchio and Collins 1986). In terms of the

analogy, both predictions can be used to "reproduce" responses of real experiments to a certain level of errors (Fig. 6). Here, we used the author's parallel multi-scale finite element analysis program (Cho 2013) named as VEEL (Virtual Earthquake Engineering Laboratory), of which accuracy and generality have proven thoroughly (Cho and Porter 2014a and 2014b). Difference is noteworthy: computer simulations are built upon engineering principles and explicit relationships among variables. Contrarily, statistical predictions are rooted in implicit interrelations among variables, requiring no pre-specified relations. Thus, statistical predictions may be directly used for the hidden relations of given data. Computer simulations can predict various responses (often continuous) spanning macroscopic and microscopic regimes while statistical predictions are restricted to "observed" responses (discrete) (see Fig. 6). But, simulations often require expensive computational cost whereas statistical predictions are relatively cheap. Thus, as long as sufficient level of accuracy is assured, the data-driven statistical predictions may as a cheap, reliable tool.

Limitation of Statistical Prediction

The limitation of the statistical prediction stems from the quality of data. If the database has little information of a certain type of structures, the statistical prediction tends to perform poorly. To quantitatively explain this "extrapolation" issue, we performed two case studies: (1) Statistical prediction after excluding WSH series (WSH1 through WSH6); (2) including WSH series. For each case, we used VEEL, GAM-TPRS and GAM-CRS for predicting F_{max} of RW1, RW2, WSH1 and WSH6. RW1 and RW2 represent typical rectangular RCSW in the database while WSH1 and WSH 6 represent special wall types residing on the boundary of database (Fig. 7). Table 3 compares the predicted responses by VEEL, GAM-TPRS, and GAM-CRS, each being normalized by experimental response. VEEL has no effect of exclusion of data points (second rows in Table 3). In general, the exclusion of WSH series in the learning process substantially weakens the accuracy of the statistical prediction (Table 3). Both GAM methods exhibit poor prediction of WSH1 and WSH6. It should be noted that even without WSH series, both GAMs accurately predict RW1 and RW2 (Table 3) since they are residing in the middle of the existing database (Fig. 7). After including WSH series (Table 3, right panel), the accuracy of both statistical prediction methods is notably improved for all wall specimens. However, there appears to be a trade-off. The statistical models need to cover wider ranges of database, the previously accurate specimens, RW1 and RW2 are showing slightly increased error.

Remarks on Parallel Processing of R & Rmpi Codes. Finding the best combination of a given number of variables is computationally intensive. Thus, we developed an algorithm-oriented parallel computing algorithm using *Rmpi* (Yu 2002). For load balance, we adopted the so-called "cyclic allocation" since as this type of problem's becomes bigger, cyclic allocation approaches the optimal scalability (Cho and Hall 2012). Fig. 8 shows a summary of parallel computing performance of the proposed parallel codes, exhibiting a reasonable scalability up to 4 slave processors. Fig. 8 shows running costs used for finding the best 5-variable combination (User code: the time spent on execution of user-defined codes; Total: the total running time). The developed *R&Rmpi* codes are available in the authors' paper [Song et al. 2016]..

Conclusions

As a big data-oriented remedy to underdeveloped countries' earthquake engineering, the generalized additive model (GAM) has been studied. Validations and applications to real-world RCSW database revealed a promising capability of the statistical prediction. Compared to the results of real tests and high-precision computer simulations, the statistical prediction appears to hold reasonable accuracy in reproducing responses of a wide range of RCSW specimens. Notably, those predictions were made without pre-specified relationships among variables. Results suggest that as far as statistical prediction accuracy is concerned, not all variables (i.e. structural attributes) are necessary, and that there may exist relative significances among some attributes. The proposed statistical approaches will shed light on the

new data-driven discovery in earthquake engineering, notably in underdeveloped countries, since the method is highly cheap, reliable and flexible. The new method may help accommodate arbitrarily new multivariate databases.

Acknowledgement

This research is supported by the research funding of Department of CCEE of ISU. Generous research funding from Black & Veatch is appreciated. The simulations are partially supported by the HPC@ISU equipment at ISU, some of which has been purchased through funding provided by NSF under MRI grant number CNS 1229081 and CRI grant number 1205413. Thanks are due to Professor Sri Sritharan for valuable discussion on earthquake engineering experiments.

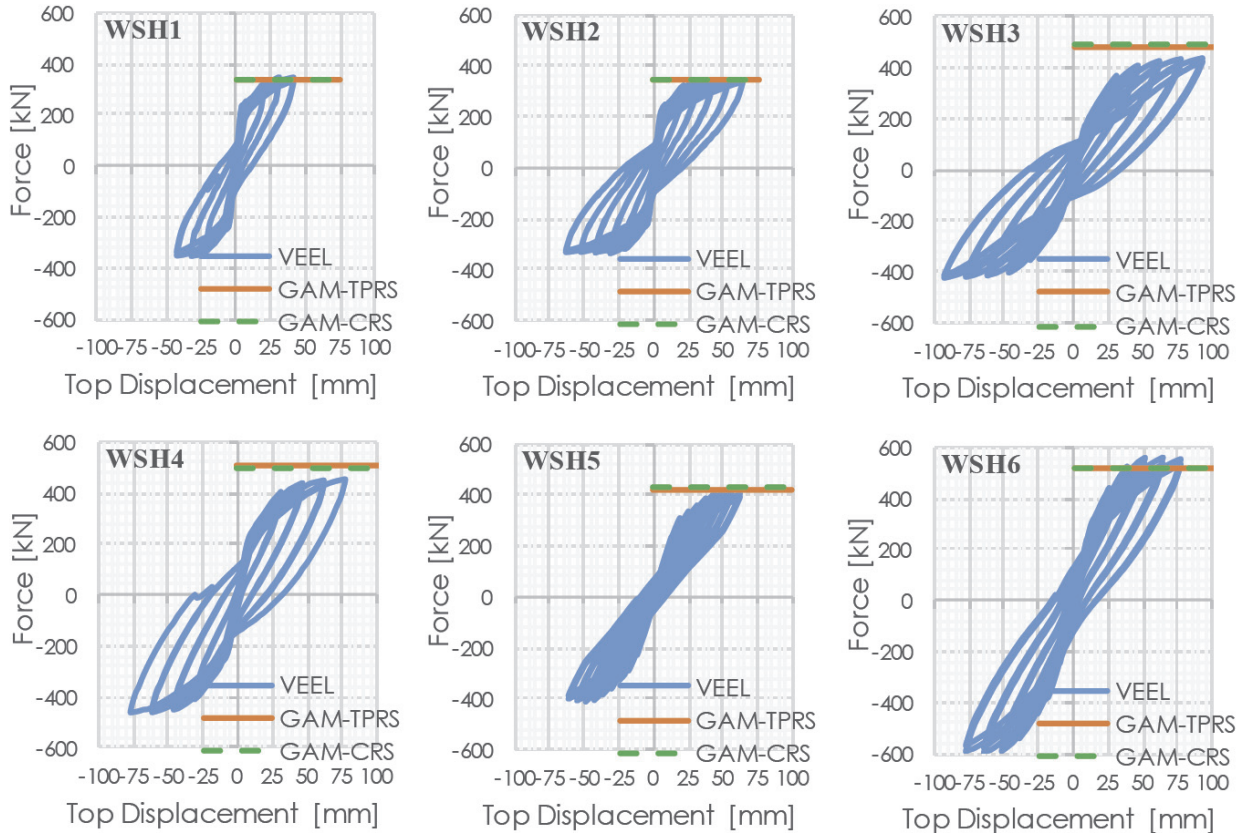

Figure 6. Comparison of WSH series: (Top 6 panels) Experimental results cited from Dazio et al. (2009); (Bottom 6 panels) predictions from high-precision computational simulations (VEEL), GAM using TPRS, and GAM using CRS. VEEL's prediction errors are less than 5% for all walls (see Song et al. 2016).

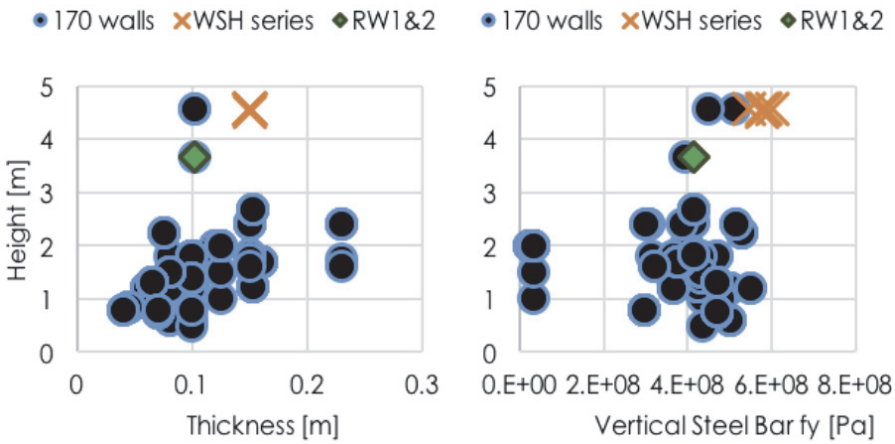

Figure 7. Scatter plot of rectangular RCSW specimens showing the ranges of database.

Figure 8. Parallel processing cost for finding the best combination out of 252 cases.

Table 3. Predictions without/with WSH series (F_{max} is normalized by that from experiment)

Without WSH's	*RW1*	*RW2*	*WSH1*	*WSH6*	*With WSH's*	*RW1*	*RW2*	*WSH1*	*WSH6*
VEEL/Experiment	1.10	1.01	1.05	0.95	VEEL/Experiment	1.10	1.01	1.05	0.95
GAM-TPRS/Ex.	1.00	1.00	4.50	24.44	GAM-TPRS/Ex.	1.08	0.94	1.01	0.88
GAM-CRS/Ex.	0.99	0.99	0.58	0.60	GAM-CRS/Ex.	1.08	0.94	1.02	0.87

References

American Concrete Institute (ACI), 2005, "Building Code Requirements for Structural Concrete (ACI 318-05) and Commentary (ACI 318R-05)," American Concrete Institute, Detroit, Michigan.

Bozorgnia, Y., Abrahamson, N. A., Atik, L. A., Ancheta, T. D., Atkinson, G. M., Baker, J. W., Baltay, A., Boore, D. M., Campbell, K. W., and Chiou, B. S. J., 2014, "NGA-West2 research project," Earthquake Spectra, Vol. 30, pp. 973-987.

Cho, I., 2013, "Virtual Earthquake Engineering Laboratory Capturing Nonlinear Shear, Localized Damage and Progressive Buckling of Bar," Earthquake Spectra, Vol. 29, pp. 103-126.

Cho, I., and Hall, J. F. 2012, "Parallelized implicit nonlinear FEA program for real scale RC structures under cyclic loading," *Journal of Computing in Civil Engineering*, pp. 26, 356-365.

Cho, I., and Porter, K., 2014a, Multilayered Grouping Parallel Algorithm for Multiple-level Multiscale Analyses, *International Journal for Numerical Methods in Engineering*, Vol. 100(12), pp. 914-932.

Cho, I., and Porter, K., 2014b, Structure-Independent Parallel Platform for Nonlinear Analyses of General Real-Scale RC Structures under Cyclic Loading, *ASCE Journal of Structural Engineering*, Vol. 140, SPECIAL ISSUE: Computational Simulation in Structural Engineering.

Dazio, A., Beyer, K., and Bachmann, H., 2009, "Quasi-static cyclic tests and plastic hinge analysis of RC structural walls," *Engineering Structures*, Vol. 31, pp. 1556-1571.

Duchon, J., 1977, "Splines minimizing rotation-invariant semi-norms in Sobolev spaces in Constructive theory of functions of several variables," *Springer*, pp. 85-100.

Ebejer, J. P., Fulle, S., Morris, G. M., and Finn, P. W., 2013, "The emerging role of cloud computing in molecular modelling," *Journal of Molecular Graphics and Modelling*, Vol. 44, pp. 177-187.

Golub, G. H., Heath, M., and Wahba, G., 1979, "Generalized cross-validation as a method for choosing a good ridge parameter," *Technometrics*, Vol. 21, pp. 215-223.

Hacker, T. J., Eigenmann, R., Bagchi, S., Irfanoglu, A., Pujol, S., Catlin, A., and Rathje, E., 2011, "The neeshub cyberinfrastructure for earthquake engineering," *Computing in Science & Engineering*, Vol. 13, pp. 67-78.

Hastie, T. J., and Tibshirani, R. J., 1990, "Generalized additive models, (Vol. 43)," *CRC Press*.

Kamdar, H. M., Turk, M. J., and Brunner, R. J., 2016, "Machine learning and cosmological simulations–I. Semi-analytical models," *Monthly Notices of the Royal Astronomical Society*, Vol. 455, pp. 642-658.

McKenna, F., Fenves, G., Scott, M., and Jeremic, B., 2000, "Open system for earthquake engineering simulation (OpenSees)." Pacific Earthquake Engineering Research Center, University of California, Berkeley.

Orakcal, K., and Wallace, J. W., 2006, "Flexural modeling of reinforced concrete walls-experimental verification," *ACI Structural Journal*, Vol. 103, pp. 196.

Rathje, E., Pinelli, J. P., Stanzione, D., Padgett, J., and Dawson, C., 2015, "NSF Abastract #1520817, Natural Hazards Engineering Research Infrastructure: Cyberinfrastructure," *National Science Foundation CMMI*.

Song, I., Cho, I., and Wong, K. W. R., 2016, "An advanced statistical approaches to data-driven earthquake engineering," *Earthquake Spectra* (under review).

Thomsen IV, J. H., and Wallace, J. W., 2004, "Displacement-based design of slender reinforced concrete structural walls-experimental verification," *Journal of Structural Engineering*, Vol. 130, pp. 618-630.

Vecchio, F. J., and Collins, M. P., 1986, "The modified compression-field theory for reinforced concrete elements subjected to shear," *ACI J.*, Vol. 83, pp. 219-231.

Wallace, J. W., Massone, L.M.,Bonelli, P., Dragovich, J., Lagos, R., Lüders, C., and Moehle, J., 2012, "Damage and Implications for Seismic Design of RC Structural Wall Buildings," *Earthquake Spectra*, Vol. 28(S1), pp. S281–S299.

Westenenk, B., de la Llera, J.C., Besa, J.J., Jünemann, R., Moehle, J., Lüders, C., Inaudi, J.A., Elwood, K.J., Hwang, S.J., 2012, "Response of Reinforced Concrete Buildings in Concepción during the Maule Earthquake," *Earthquake Spectra*, Vol. 28(S1), pp. S257–S280.

Wood, S., 2006, "Generalized additive models: an introduction with R," CRC press.

Wood, S. N., 2003, "Thin plate regression splines," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, Vol. 65, pp. 95-114.

Yu, H., 2002, "Rmpi: parallel statistical computing in R," *R News*, Vol. 2, pp. 10-14.