CAN BIG DATA APPROACHES HELP EARTHQUAKE ENGINEERING IN UNDERDEVELOPED COUNTRIES?

In Ho Cho, Assistant Prof, CCEE; Black & Veatch Faculty
Ikkyun Song, Graduate RA, CCEE
Raymond K. W. Wong, Assistant Prof, STAT
Iowa State University, **USA**

16th US-JAPAN-NZ Workshop
On the Improvement of Structural Engineering and Resiliency
Todaiji Temple Cultural Center in Nara, Japan
June 27-29, 2016

Era of Big Data

Formidably Complex Data

Existing Database

Simulated Database

Human-like (better) Decision

Artificial Intelligence

Machine Learning

Super Computer

Engineering Big Data

1. Immense Size and Volume ➤ needs Parallel/Cloud computing
2. Complex inter-relations ➤ advanced statistics/AI
3. High dimensionality (Multiple dimensions/Many variables)
4. High velocity (not in current study), i.e. rapid change of data
5. Pursue "Data-driven" discovery
6. Data may foretell "hidden" relations or problematic issues

Difference from Traditional Statistical Methods

1. A few variables (predictors) to describe a response
2. Simple relations
3. Needs a pre-specified relation among variables
4. Statistical methods are often used to confirm the pre-defined relations
5. Hard to find "hidden" relations or problems

Why are Eng. Big Data important for Our Fields

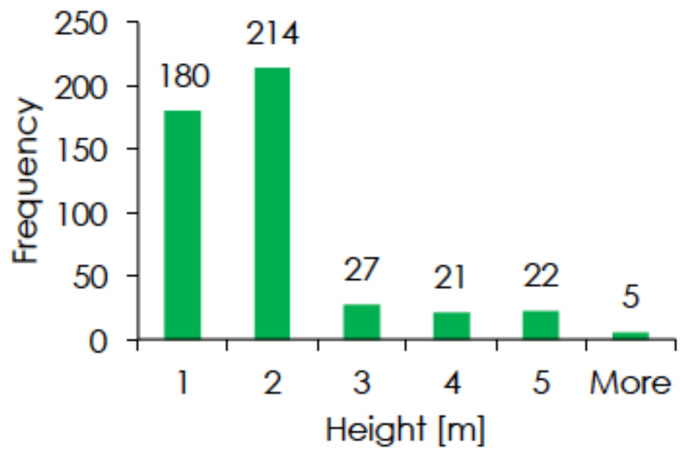
1. Global communities have established DB: e.g., RCSW DB of
    - ACI 445-B Shear Wall Database: https://datacenterhub.org/resources/142
    - SERIES Wall Database: https://datacenterhub.org/resources/355 High velocity (not in current study), i.e. rapid change of data
    - BRI Wall Database: https://datacenterhub.org/resources/14087

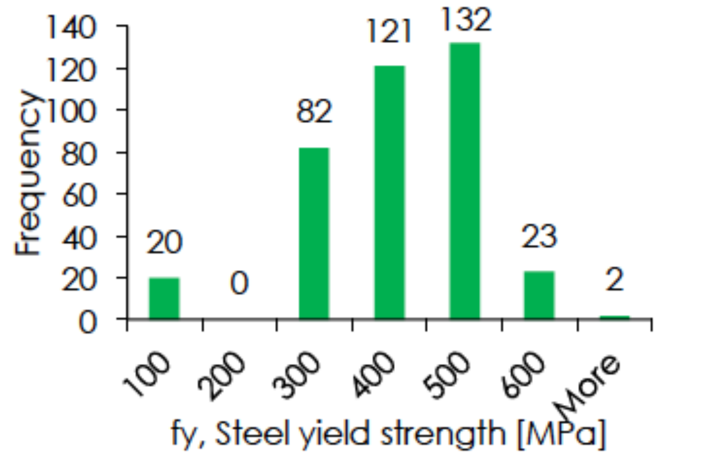
2. Our DB share the Big Data characteristics (size, complexity, etc.)
3. Engineering community seeks to foresee hidden problems
4. Strong need for Big Data-oriented methods, algorithms, etc.

5. Underdeveloped countries may have different DB and practices
6. The proposed methods are cost-effective compared to real tests

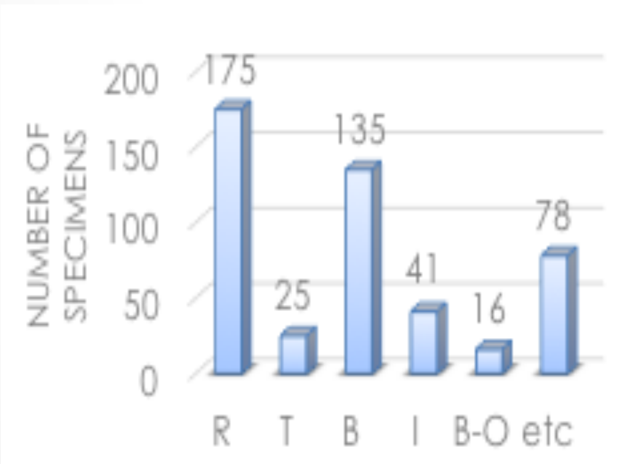
Introduction: Statistical Issues of Community DB

Sparseness and Biasness



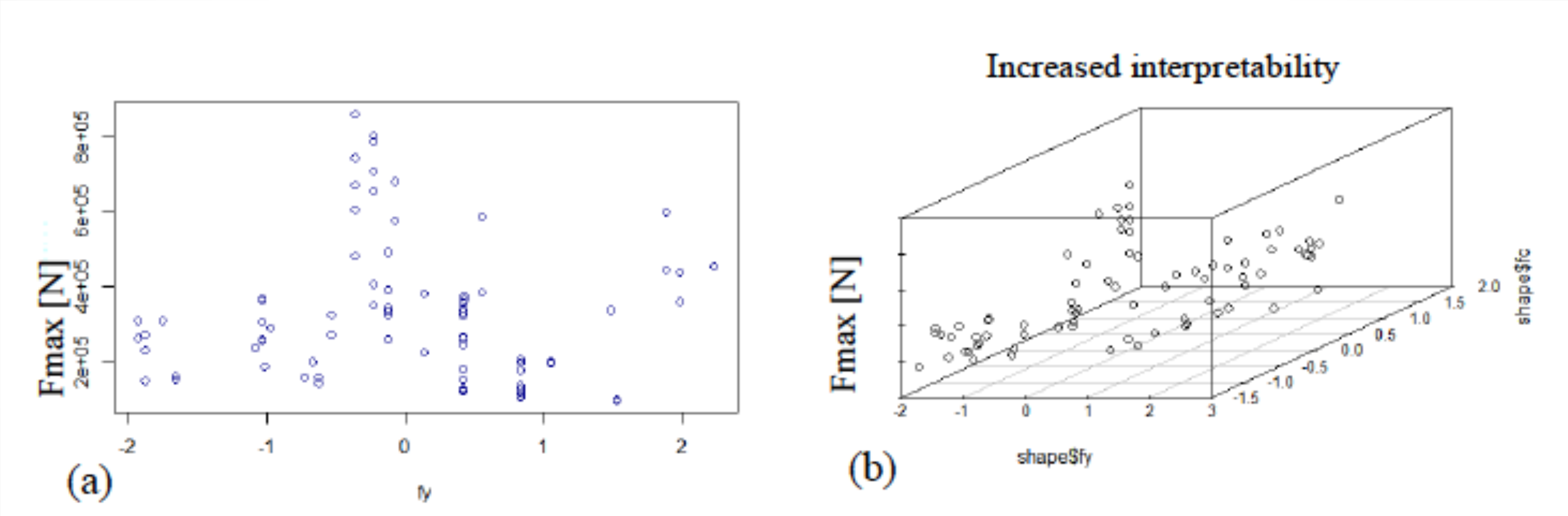


Revealed from 470 real experiments of RC shear wall database (collected from *NEESHub*, international reports, and literature).



R: Rectangular RCSW
T: T-shaped RCSW
B: Barbell-shaped RCSW
I: I-shaped RCSW
B-O: Barbell-shaped wall with Opening

Introduction : High Dimensionality & Interpretability



Change in the interpretability of database with increasing dimensionality:

(a) two-dimensional (2D) scatter plot of standardized f_y (steel yield strength of longitudinal bars) and F_{max} (maximum shear force)
(b) 3D plot of F_{max}, the standardized f_y, and the standardized f_c' (concrete strength)

Generalized Additive Model (GAM)

1. Pioneering works of Hastie and Tibshirani (1990).
2. A non-parametric extension of generalized linear model
3. Covariates enter into the model through smooth functions
4. No need for pre-defined relation among variables
5. Focus on flexible, powerful prediction capability

General form of GAM is given by (Wood 2006)

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

Where g is a smooth link function, f is a smooth function

$$\mu_i \equiv \mathbb{E}(Y_i|\boldsymbol{x}_i)$$

Yi is *i*-th response (from exponential family distributions)

xi is *i*-th vector of data points

In our case, *Yi* is the *i*-th RC shear wall (RCSW)'s maximum force

$$\boldsymbol{x}_i = \{length_i, height_i, AxialForce_i \ \ldots\}$$

GAM: Cubic Regression Spline

Fitting of the model is done by maximizing likelihood with a penalty term of

$\lambda \int [f''(x)]^2 dx$ where λ = smoothing parameter

Two popular smooth functions:

Cubic Regression Spline (CRS) & Thin Plate Regression Spline (TPRS)

(1) Cubic Regression Spline (CRS)

- Constructed by connecting cubic polynomial sections.
- "Knot" locations are pre-selected
- e.g., cubic spline functions of Gu (2013) are given by

$$b_1(x) = 1, b_2(x) = x, \text{and } b_{i+2}(x) = R(x, x_i^*) \text{ for } i = 1, 2, \cdots, \text{p} - 2$$

where

$$R(x, x^*) = [(x^* - 1/2)^2 - 1/12][(x - 1/2)^2 - 1/12]/4$$
$$-[(|x - x^*| - 1/12)^4 - 1/2(|x - x^*| - 1/12)^2 + 7/240]/24$$

x^* = the know location

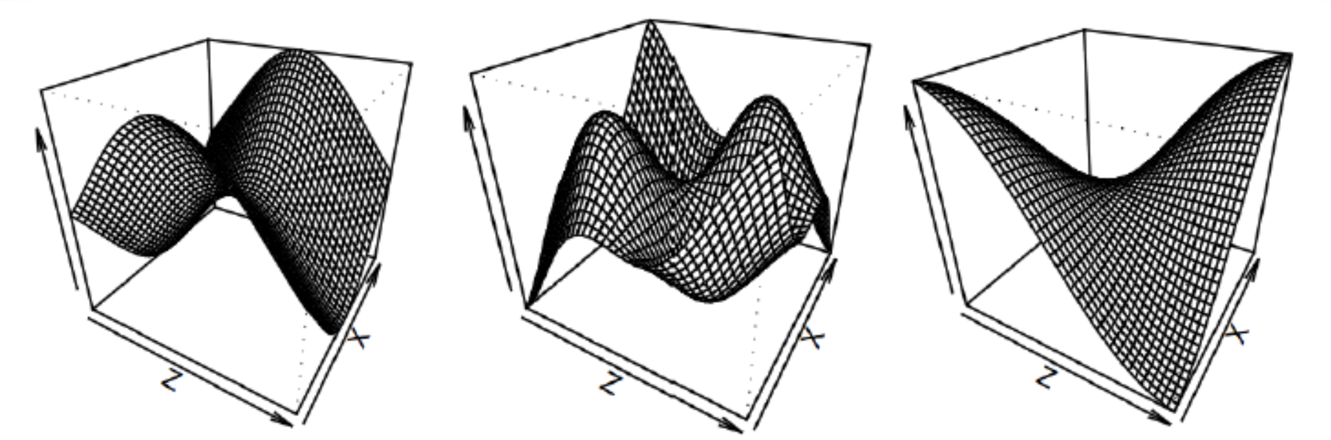
GAM: Thin Plate Regression Spline

(2) Thin Plate Regression Spline (TPRS)

- **Suitable for many covariates**
- **"Knot-free"**
- **Computationally more expensive than CRS**
- **Thin spline functions *f* (Duchon, 1977) are found by minimizing**

$$\|\boldsymbol{y} - \boldsymbol{f}\|^2 + \lambda J_{md}(\mathrm{f})$$ **where J_{md} means "wiggliness" of *f***

$$J_{md} = \int \cdots \int_{R^d} \sum_{v_1+\cdots+v_d} \frac{m!}{v_1! \cdots v_d!} \left(\frac{\partial^m f}{\partial x_1^{v_1} \ldots \partial x_d^{v_d}} \right)^2 dx_1 \cdots dx_d$$

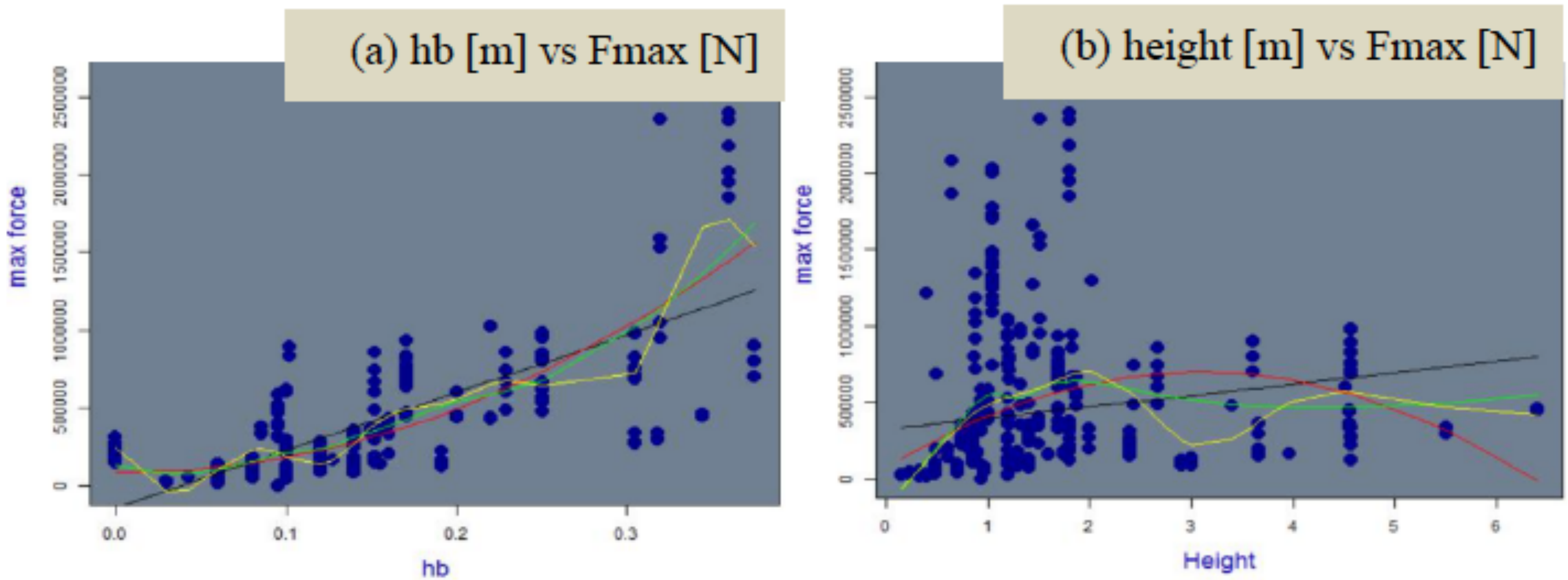


Example of thin plate spline basis function using 2 covariates (cited from Wood 2006)

GAM's Flexibility

Regression type

Black = Linear; **Red** = Parabolic; **Green** = GAM&CRS; **Yellow** = GAM&TPRS]



Example of one-dimensional regressions of 470 real RC wall data:

(a) hb (thickness of boundary element) versus Fmax
(b) wall height versus Fmax.

Metrics for Prediction Quality

Three metrics are used to compare **predictive power of the statistical methods (as done by Machine Learning-based works of Kamdar et al. 2016)**

The larger value, the more accurate prediction.

1. Cross-Validation Error (CVE) Ratio: CVE/CVE*b*

$$\mathrm{CVE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_{experiment}^{i} - y_{predicted}^{i}\right)^2$$

$$\mathrm{CVE}_b = \frac{1}{N}\sum_{i=1}^{N}\left(y_{experiment}^{i} - y_{mean,predicted}\right)^2$$

2. Pearson Coefficient: ρ

$$\rho = \frac{cov(y_{predicted}, y_{experiment})}{\sigma_{y_{predicted}} \times \sigma_{y_{experiment}}}$$

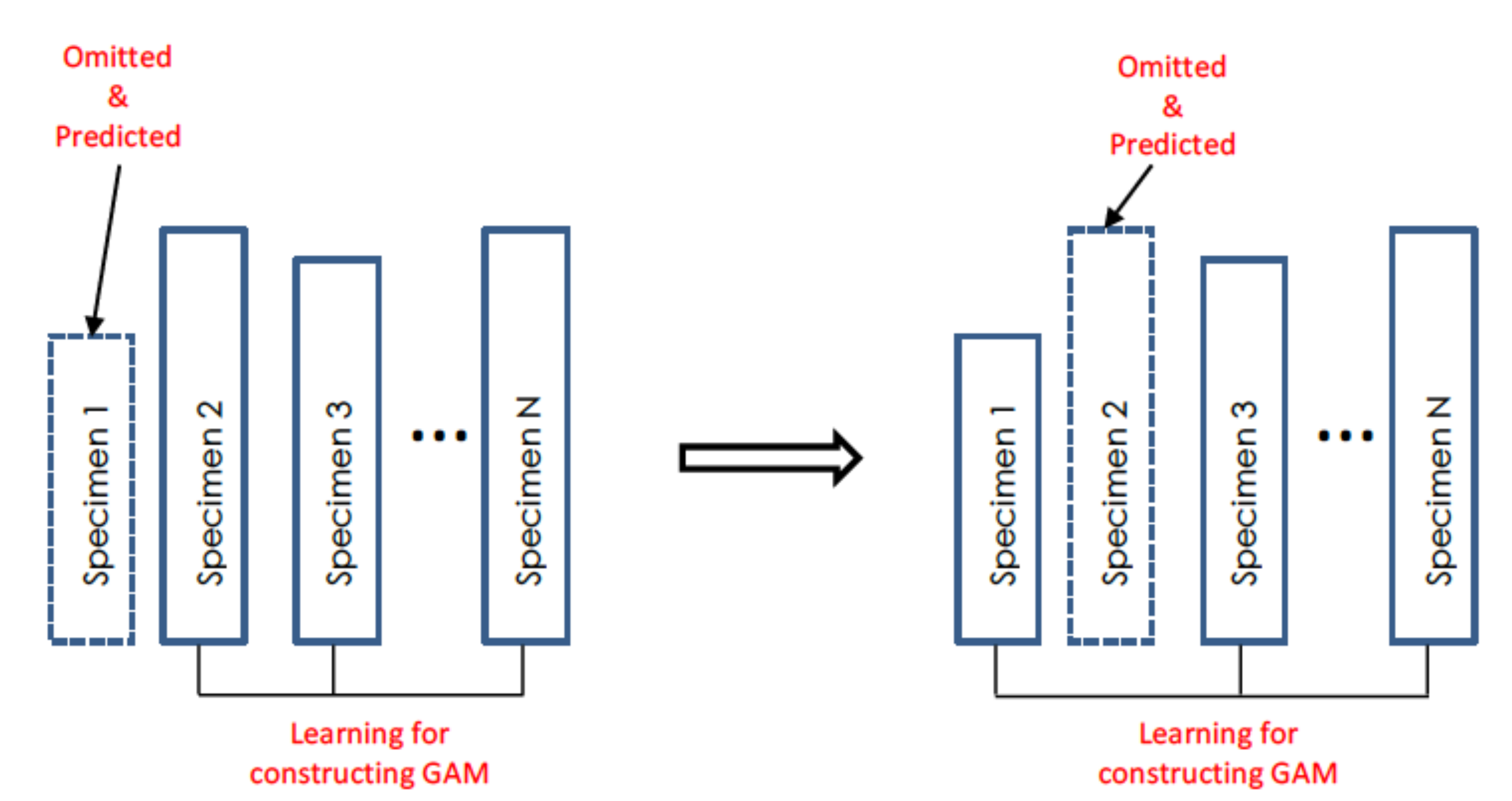
3. Coefficient of Determination: R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{N}\left(y_{experiment}^{i} - y_{mean,predicted}\right)^2}{\sum_{i=1}^{N}\left(y_{experiment}^{i} - y_{predicted}^{i}\right)^2}$$

Prediction with GAM using Cross-Validation

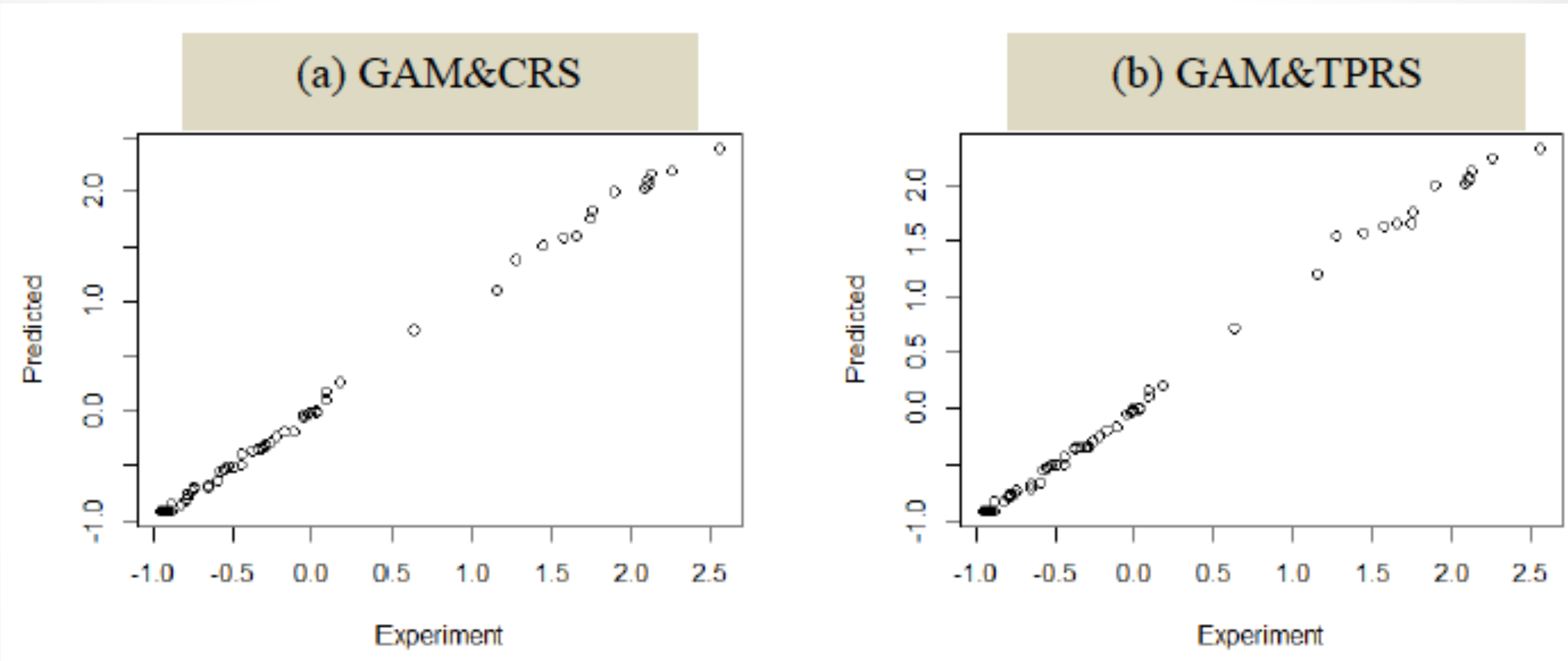
Three Steps of Cross-Validation

1. Remove a specimen (data point)
2. Learn the remaining specimens (remaining data)
3. Predict the removed specimen



Overall Prediction Power of GAM

- **Used the "best" model of GAM (shall be discussed next sections)**
- **GAM setting: logarithmic link, Gamma distribution of response**
- **Two smooth functions, CRS and TRPS, were separately used**



Q-Q plot of real experimental data and the predicted value
(a) Using GAM-CRS
(b) Using GAM-TPRS

Construct a "Best" GAM

Which variables must be included in the GAM for best prediction?

Number of total possible combinations: e.g., when 4 variables are used

$$\frac{10!}{4!\,(10-4)!} = 210$$

For completely Data-Driven Prediction,

- **No pre-specified relation among variables**
- **No prejudice on importance of each variable**

- **1 target response: maximum shear resistance, F_{max}**
- **Start with 10 variables from DB**
  1. axial force ratio (denoted by afr)
  2. wall thickness (thickness)
  3. boundary element's thickness (hb)
  4. boundary element's width (bb)
  5. wall height (height)
  6. wall length (length)
  7. primary reinforcing bar's yield strength (fy)
  8. bar diameter (dia)
  9. concrete compressive strength (fc)
  10. boundary element reinforcement ratio (bderr)

Constructing a "Best" GAM

Best GAMs using CRS for a given number of variables

# of Vari.	# of Comb.	Best combination of variables (p-values)	CVE/CVE_b	Pearson	R^2
2	45	height(6.24e-11) hb(1.85e-05)	12.24	0.958	0.918
3	120	height(<2e-16) hb(3.71e-11) dia(0.00272)	16.39	0.969	0.939
4	210	height(<2e-16) afr(3.11e-13) hb(5.51e-10) dia(1.57e-08)	21.00	0.976	0.952
5	252	height(<2e-16) afr(1.73e-13) dia(5.51e-08) hb(5.59e-06) fc(0.292)	22.46	0.978	0.955
6	210	afr(<2e-16) thickness(<2e-16) hb(1.27e-11) height(9.51e-08) fy(7.01e-08) dia(3.26e-06)	**26.21**	**0.981**	**0.962**
7	120	afr(<2e-16) thickness(<2e-16) hb(1.76e-11) height(1.01e-07) fy(2.69e-07) dia(4.00e-06) fc(0.719)	25.75	0.981	0.961
8	45	afr(<2e-16) height(<2e-16) fy(<2e-16) bb(3.07e-10) length(6.60e-09) thickness(1.9e-08) dia(7.38e-05) hb(0.163)	24.64	0.980	0.959
9	10	afr(<2e-16) height(<2e-16) fy(<2e-16) bb(7.85e-10) thickness(5.37e-08) length(1.00e-08) dia(9.89e-05) hb(0.171) fc(0.707)	23.61	0.979	0.958
10	1	afr(<2e-16) height(<2e-16) fy(1.15e-13) bb(5.63e-08) length(2.58e-07) thickness(2.0e-06) dia(0.00999) hb(0.10544) bderr(0.64389) fc(0.72648)	4.63	0.918	0.784

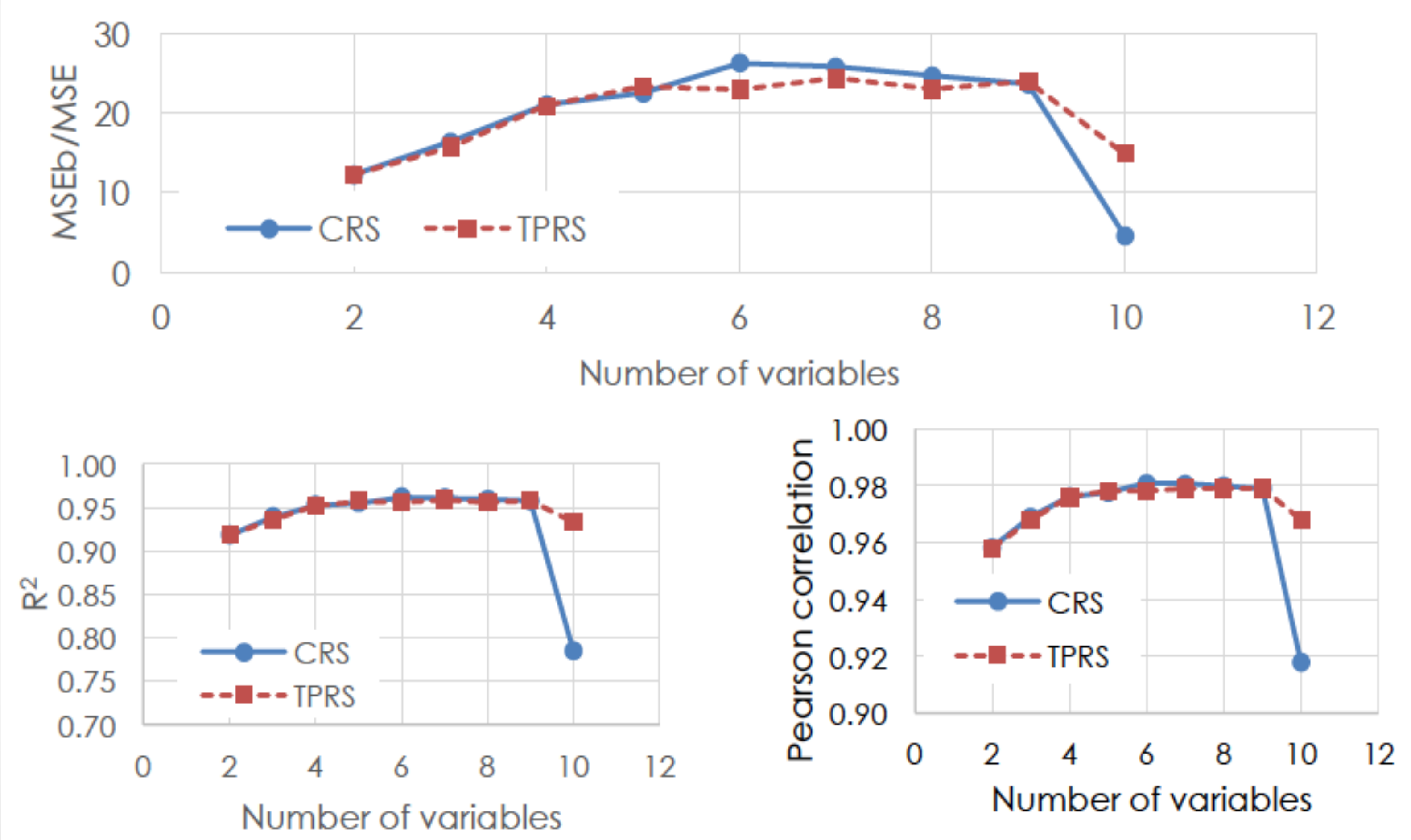
Constructing a "Best" GAM

Best GAMs using TPRS for a given number of variables

# of Vari.	# of Comb.	Best combination of variables (p-values)			CVE/ CVE_b	Pearson	R^2
2	45	length(5.91e-11)	height(1.59e-09)		12.22	0.958	0.918
3	120	length(<2e-16)	dia(<2e-16)	afr(2.11e-11)	15.70	0.968	0.936
4	210	length(<2e-16) dia(1.51e-11)	height(<2e-16)	afr(1.18e-13)	20.89	0.976	0.952
5	252	afr(<2e-16) bderr(1.43e-06)	thickness(2.06e-09) length(0.00033)	fy(3.24e-07)	23.32	0.978	0.957
6	210	afr(<2e-16) bderr(2.17e-06)	thickness(3.76e-09) length(0.00044)	fy(9.12e-07) fc(0.84103)	22.92	0.978	0.956
7	120	afr(<2e-16) thickness(6e-04) length(0.211003)	height(4.53e-05) dia(0.002263)	fy(0.000306) hb(0.010451)	**24.33**	**0.979**	**0.959**
8	45	afr(<2e-16) length(1.40e-05) dia(0.2323)	height(<2e-16) thickness(0.0152) bderr(0.6818)	fy(6.41e-06) hb(0.1574)	22.97	0.979	0.956
9	10	afr(<2e-16) fy(1.21e-07) dia(0.730411)	length(<2e-16) height(0.000944) thickness(0.767511)	bb(5.34e-08) bderr(0.018341) fc(0.792632)	23.93	0.979	0.958
10	1	afr(<2e-16) fy(0.000865) length(0.699968) fc(0.888840)	bb(6.25e-05) hb(0.001342) thickness(0.77149)	height(3.68e-04) dia(0.248087) bderr(0.875868)	14.88	0.968	0.933

Constructing a "Best" GAM

Variations of three metrics with varying number of variables



Engineering Interpretation from GAM predictions

Without any prejudice and pre-specified relationship among variables

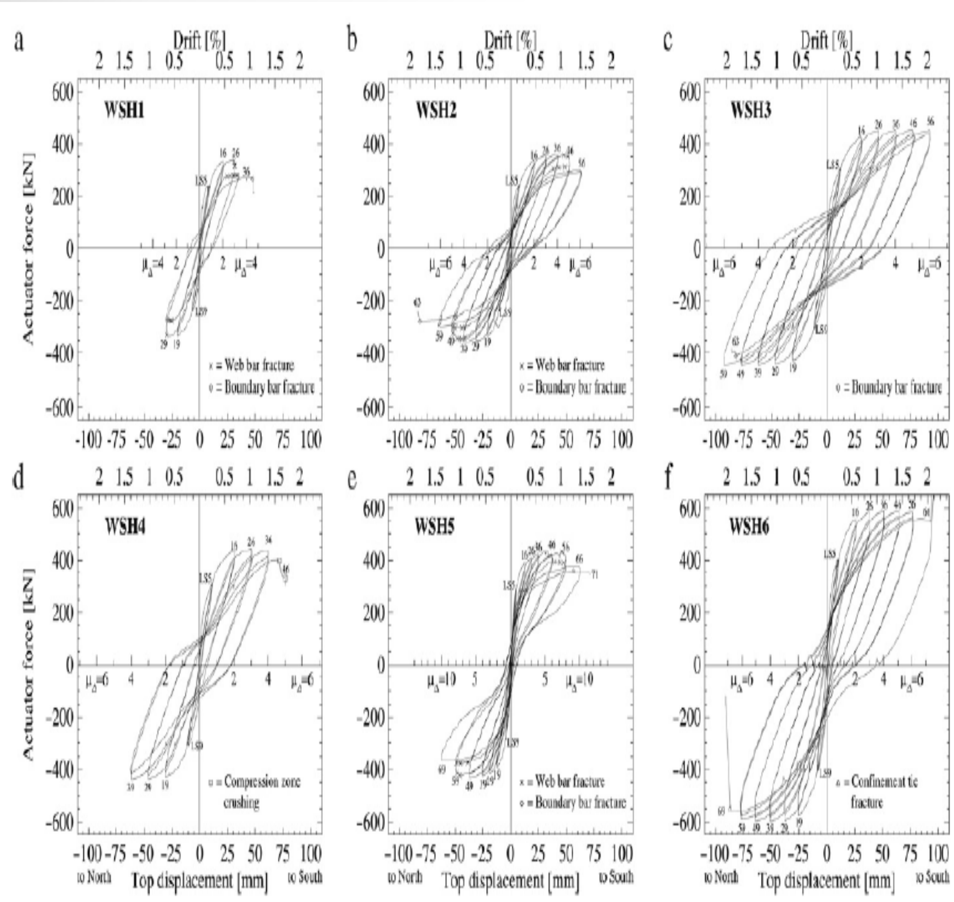
- Both methods identify the same five variables as indispensable to accurate prediction

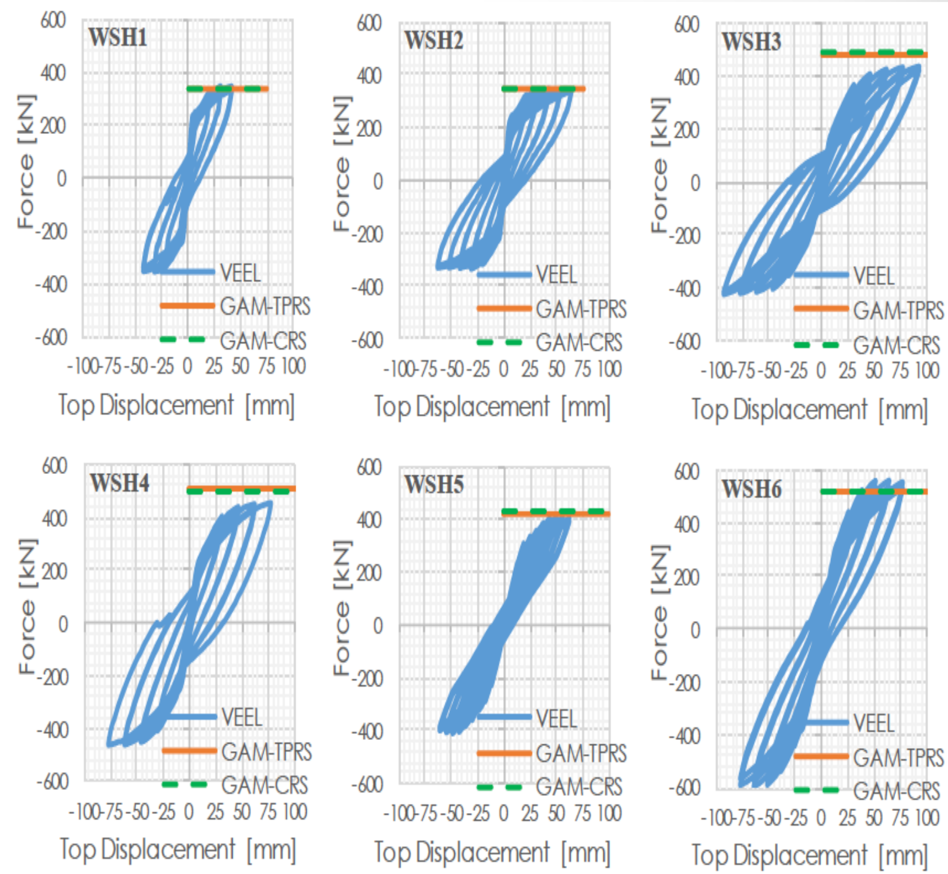
- Interestingly, the axial force ratio is identified as the most indispensable variable for rectangular RCSW prediction, which is aligned with the recent in-depth researches of (Wallace et al. 2012; Westenenk et al. 2012)

The proposed approach may help engineers and researchers to identify hidden roles of some ignored factors or even problematic issues

GAM versus High-Precision Computer Simulations

6 RC Wall tests (WSH series of Dazio et al. 2009)



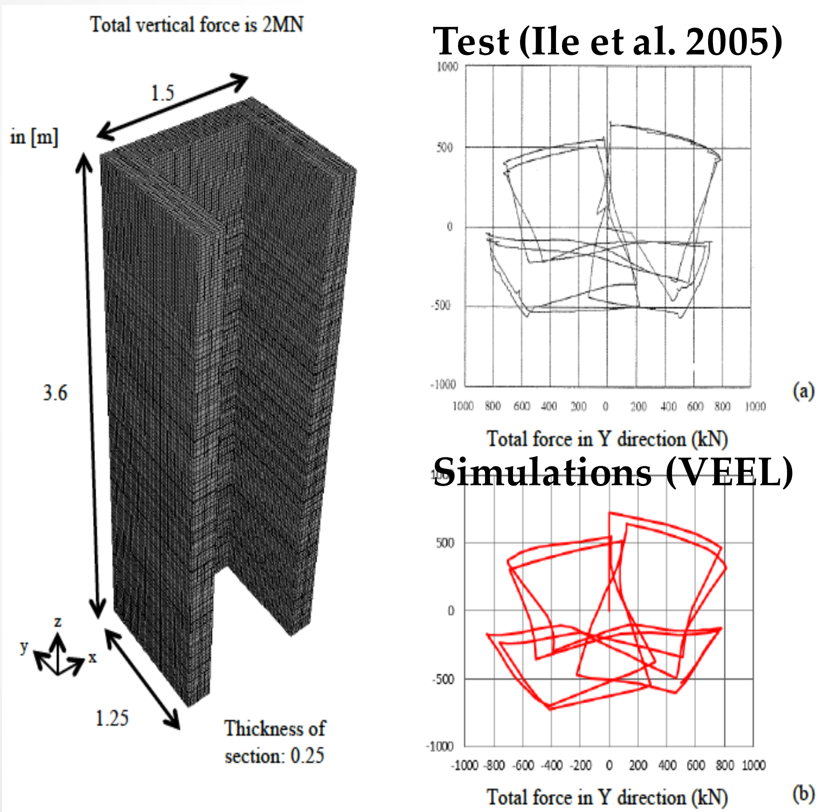


Experimental results cited from [Dazio et al. (2009)]

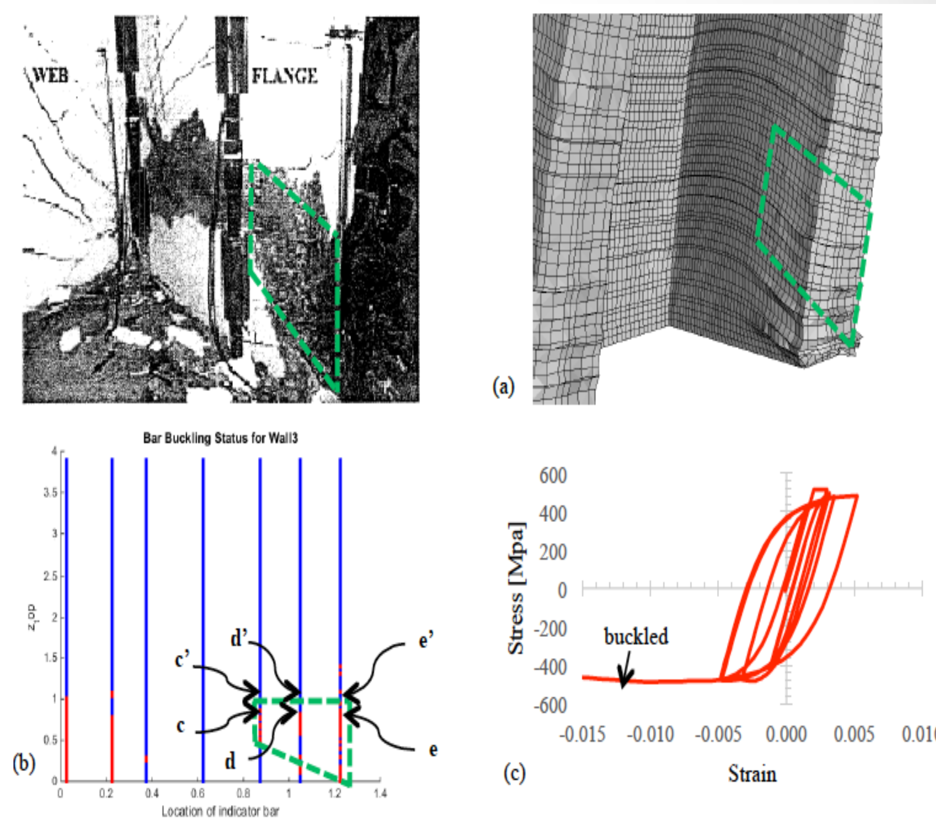
Predictions from high-precision computational simulations (VEEL), GAM-TPRS, and GAM-CRS

GAM versus High-Precision Computer Simulation

High-precision computer simulation can be used to enrich engineering Big Data for better prediction and investigations



Simulated bi-directional Force-Displacement response

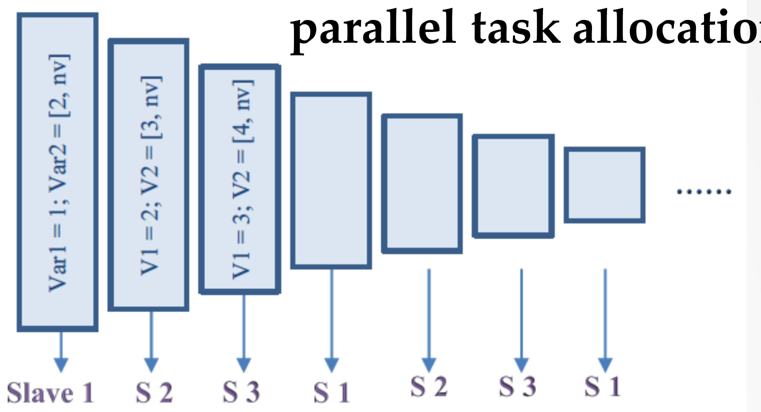


Simulated phenomena of progressive bar buckling & crushing

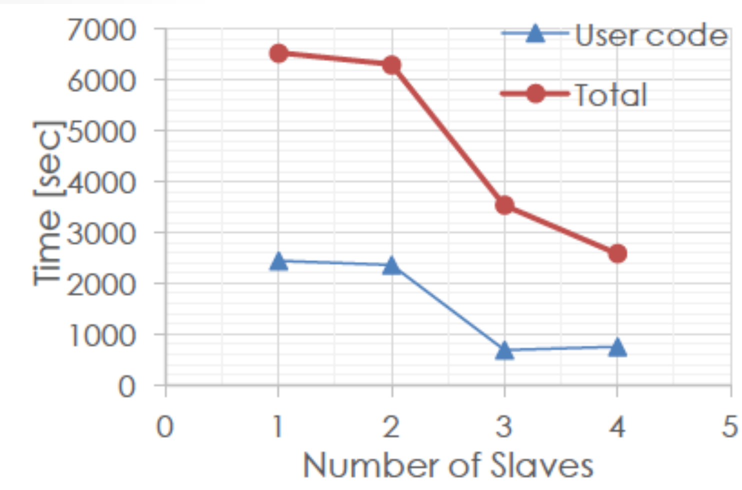
Super Computing using *R* & *Rmpi*

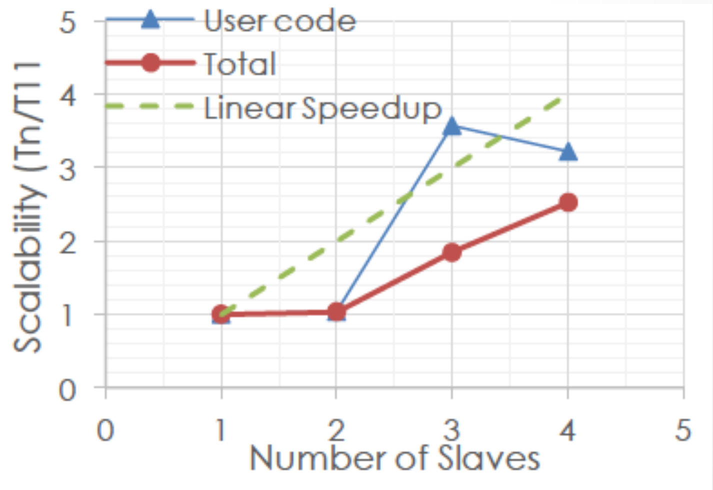
Serial and Parallel codes will be available of authors' paper (under review by *Earthquake Spectra*)

Problem-oriented cyclic parallel task allocation



Promising parallel performance of the developed codes





Conclusion

1. Highly flexible and general; can resolve high-dimensionality and complexity of engineering Big Data.

2. In light of economic benefit, this approach may aid EQE in underdeveloped countries as well as global engineers and researchers

3. To accelerate this transition, global data sharing, merging, utilization are significant
4. Convergence among Experiments, Engineering and BigData is critical

In next 5-10 years

The Big Data-oriented approaches will help

- Engineers to quickly check their designs
- Researchers to identify hidden problems and unravel relations
- Reduce unnecessary experiments
- Better focus on innovative new experiments

Acknowledgement

This research is supported by the research funding of Department of Civil, Construction, and Environmental Engineering of Iowa State University. Generous research funding from Black & Veatch is appreciated. The simulations of this paper is partially supported by the HPC@ISU equipment at Iowa State University, some of which has been purchased through funding provided by NSF under MRI grant number CNS 1229081 and CRI grant number 1205413. Special thanks are due to Professor John F. Hall for his productive guidance regarding nonlinear analysis methods, and also to Professor Sri Sritharan for valuable discussion on earthquake engineering experiments.

Thank You Very Much

Further discussion:

icho@iastate.edu