Exploring the Efficiency of BIGDATA Analyses in SHM

THOMAS J. MATARAZZO, S. GOLNAZ SHAHIDI
and SHAMIM N. PAKZAD

ABSTRACT

In SHM, BIGDATA is currently perceived as the result of special applications such as long-term monitoring, dense sensor arrays, or high sampling rates. Along the development of novel sensing techniques as well as advances in sensing devices and data acquisition technology, it is expected that BIGDATA will become more easily obtained. In a previous study, the evaluation of selected SHM procedures exemplified computational challenges for BIGDATA. It was concluded that processing BIGDATA in SHM can be prohibitive, especially when incorporating more sensors into the analysis. This paper focuses on the relationship between sensor network size and information extracted by an SHM procedure, e.g., system identification or damage detection. An application is presented to study the accuracy and effectiveness of these procedures as more sensors are included in the datasets.

INTRODUCTION

The definition of BIGDATA remains a subject for debate. In 2001, Laney introduced the three V's model: Volume, Variety, and Velocity as distinctive features of BIGDATA [1]. Ever since, this definition has been modified to include qualitative features such as Variability, Veracity, and Value [2]. However, this is not the only way computer scientists define BIGDATA. Ward and Barker [3] present a survey of definitions of BIGDATA by leading computer technology companies, and concluded that BIGDATA definitions included at least one of the following critical factors: size, complexity, and processing tools.

Such BIGDATA definitions are rather incomplete; with fast advancing technology, computer scientists are only one group of scholars dealing with the notion of BIGDATA. A BIGDATA problem can be encountered in any data-driven study or decision-making process ranging from sociology or political science to biology or physics [4, 5]. Huge-scale machine learning, compressed sensing, and optimization problems are also considered BIGDATA problems.

Department of Civil and Environmental Engineering, Lehigh University, ATLSS Engineering Research Center, 117 ATLSS Drive, Imbt Labs, Bethlehem, PA 18015, U.S.A.

In this context, BIGDATA encompasses problem sizes where traditional solution techniques cannot reasonably be applied [6]. A more recent report on the importance of BIGDATA by Backer and Laney in 2012, focuses on the "innovative processing solutions" [7] and thus seems more applicable to a broader perspective than data science. Therefore, BIGDATA is about the tools and techniques to deal with emerging unprecedented problems' dimensions over all disciplines.

This paper focuses on BIGDATA in Structural Health Monitoring (SHM); a field that aims to interpret structural vibration data into information about structural characteristics. Such information is valuable in damage identification [8], model calibration [9], and vulnerability assessment of constructed structures [10]. With rapid development in sensing technology, SHM problem sizes rapidly increase in terms of time and space parameters. The authors previously investigated the sensitivity of the several SHM techniques in regard to growth of sensor network size and collected data samples [11]. It was observed that for certain data volumes, computational requirements of system identification (SID) techniques were orders of magnitude higher than pre-processing and data-driven damage detection methods. Also, it was determined that computational time is significantly more sensitive to the number of sensors in the network than the number of samples. However, it is not clear whether an increase in data size necessarily guarantees an increase in structural information. Therefore, this analysis of BIGDATA in SHM focuses on the relationship between sensor network size and information extracted by an SHM procedure, e.g., SID or damage detection.

BIGDATA PROBLEM DESCRIPTION

In the context of SHM, BIGDATA problems are not sufficiently described by data dimensions. The scalability and associated computational costs of an SHM procedure define the suitability for processing a very large dataset [11]. Many output-only SID algorithms [12-18] are not scalable procedures; their computational requirements typically increase cubically as with sensor channels and linearly with samples. Similar trends were observed for damage detection techniques using AR, SVR, and ARX models [11].

In consideration of modern computational limits, which cannot be identified uniformly, this paper considers potential gains in information by introducing more sensors into a data set and analysis. It is assumed that the large efforts required for processing BIGDATA would, in one way or another, provide a deeper insight in the analysis; an insight that is exclusively provided by BIGDATA, or in this case, a large number of sensors. While there are numerous SHM applications that would greatly benefit from this BIGDATA insight, e.g., long-term monitoring, finite element model updating, there are also many research questions that can be sufficiently addressed without BIGDATA. Part of the BIGDATA problem is determining for which questions BIGDATA is an appropriate response.

With more sensors available during data collection, the measurement space can contain a dense spatial grid, thus yielding more spatial information, e.g., high-resolution mode shape estimates. Furthermore, frequency content observed by a sensor network is highly dependent on individual sensor locations within the structure. It is, however, unclear whether a proposed dense spatial grid (more sensors) will necessarily improve individual frequency or damping estimates. In other words, does redundant frequency data increase estimation accuracy?

Similar questions can be posed in regards to data-driven damage detection methods. An ideal data-driven damage detection methodology would provide information about time, location, extent, and severity of damage. However, many of the existing techniques accomplish the first two or three of these goals, because such methods statistically test the features extracted from measured signals that might not have a physical interpretation to assess damage severity. Dense sensor networks are deemed more accurate in identifying the location and extent of damage. However, there is no clear limit to the information supplied by spatially dense sensor grids. Also, could one successfully detect the location and extent of damage by processing only a portion of the sensor network data?

In this paper, we aim to draw attention to these issues regarding computational complexity of SHM procedures, and investigate the potential value of information gained by using denser sensor networks in SHM projects. This investigation is performed through a numerical example of a bridge girder, where several sensor deployment setups are considered.

APPLICATIONS OF BIGDATA IN SHM

To exemplify BIGDATA in SHM, multiple sensing schemes are considered for the monitoring of a 500 DOF flexible, simple beam. In each sensing scheme, the simulated sensors are located uniformly across the structure; a higher number of sensors indicates a higher spatial resolution during measurement and enables the collection of more spatial information. The following sections implement SID and damage detection procedures using data that is considered to be very large in their respective applications. The information extracted by these SHM procedures will be compared among different sensing schemes to assess potential benefits of using BIGDATA in either application.

System Identification Performance with BIGDATA

The first part of this BIGDATA investigation included five sensing schemes, each with uniform spacing: 8, 16, 32, 64, and 128 sensors. The 500 DOF beam structure was excited by random white noise and for each sensing scheme, noisy DOF acceleration responses were selected according to a sensing scheme for the dataset, resulting in five data matrices with increasingly large dimensions. SID was performed on each dataset using the Structural Modal Identification Toolsuite (SMIT) [19] which is available for free download at http://smit.atlss.lehigh.edu. More specifically in SMIT, the ERA-NExT algorithm was chosen to construct the stabilization diagrams, requiring fifty repeated identifications (even model orders two through one hundred) for each dataset (sensing scheme). The computational time required for SID of each sensing scheme is provided in Table I. As discussed earlier, as more sensors are included in SID, the computational requirements grow cubically.

The identification of the first eight structural modes was targeted, for which, the smallest sensing scheme of 8 sensors was sufficient. For this same research goal, the use of 64 sensors represents a very large dataset and relatively, 128 sensors yield

BIGDATA. Sensing schemes that utilize more sensors will permit a higher resolution for estimated mode shapes, but will the accuracy of frequency or damping estimates be improved?

In Table II, frequency estimates for all eight modes are given for each sensing scheme. A higher accuracy was observed for higher modes in each scheme. When compared to other sensing schemes, there is no clear trend for the frequency accuracy.

TABLE I. COMPUTATIONAL REQUIREMENTS OF FIFTY ERA-NEXT IMPLEMENTATIONS FOR EACH SENSING SCHEME.

8 Sensors	**16 Sensors**	**32 Sensors**	**64 Sensors**	**128 Sensors**
14 sec	1 min 39 sec	13 min	2 hrs 4 min	28 hrs 35 min

TABLE II. ERA-NEXT FREQUENCY ERRORS (%) AND MAC VALUES.

Mode	**8 Sensors**		**16 Sensors**		**32 Sensors**		**64 Sensors**		**128 Sensors**	
	error	*MAC*	*error*	*MAC*	*error*	*MAC*	*error*	*MAC*	*error*	*MAC*
1	- 1.13	0.996	- 0.94	0.996	- 0.87	0.996	- 1.00	0.996	- 0.97	0.996
2	- 1.19	0.995	- 1.16	0.995	- 1.15	0.995	- 1.15	0.995	- 1.17	0.995
3	0.80	0.999	0.81	0.999	0.83	0.999	0.87	0.999	0.84	0.999
4	0.06	1.000	0.06	1.000	0.06	1.000	0.06	1.000	0.06	1.000
5	0.03	0.999	0.03	0.999	0.04	0.999	0.03	0.999	0.03	0.999
6	- 0.02	1.000	- 0.02	0.999	- 0.02	0.999	- 0.02	0.999	- 0.02	0.999
7	0.04	1.000	0.03	1.000	0.03	1.000	0.03	1.000	0.04	1.000
8	- 0.03	1.000	-0.03	1.000	- 0.03	1.000	- 0.03	1.000	- 0.03	1.000

As more sensors were included in the analysis, the individual frequency errors fluctuated slightly, but did not improve overall. For brevity, corresponding SID damping errors are not displayed explicitly; they were similar in behavior to the frequency errors in Table II. In other words, the estimation accuracy of frequency and damping is not necessarily restricted by the sensing scheme; in this case, the accuracy is limited by the SID procedure, which may rely on other data attributes, e.g., sampling rate or data length.

The modal assurance criterion (MAC) value was computed for each sensing scheme and is displayed in Table II. The MAC values were invariant to the sensing schemes, indicating that the overall mode shape accuracy was retained as more sensors were added. However, it is important to acknowledge the extra value of high-resolution mode shapes, provided by denser sensing schemes, which may be required to properly assess certain research questions. Estimated mode shapes are compared directly to the exact values in Figure 1. The mode shapes identified from sensing schemes with 8, 32, and 128 sensors are superimposed. The results in Figure 1 further demonstrate that rich spatial information can be accurately extracted from BIGDATA. The computational efforts, listed in Table I, indicate that even for simple structures, high-resolution mode shapes come at a high price. When deciding if BIGDATA would be beneficial to the analysis, it is necessary to first establish what level of mode shape detail is required, based on the specific research goals and the complexity of the structure. In this example, since the structure is simple and the goal is simply identification, the redundancy of the data is evident beyond 32 sensors.

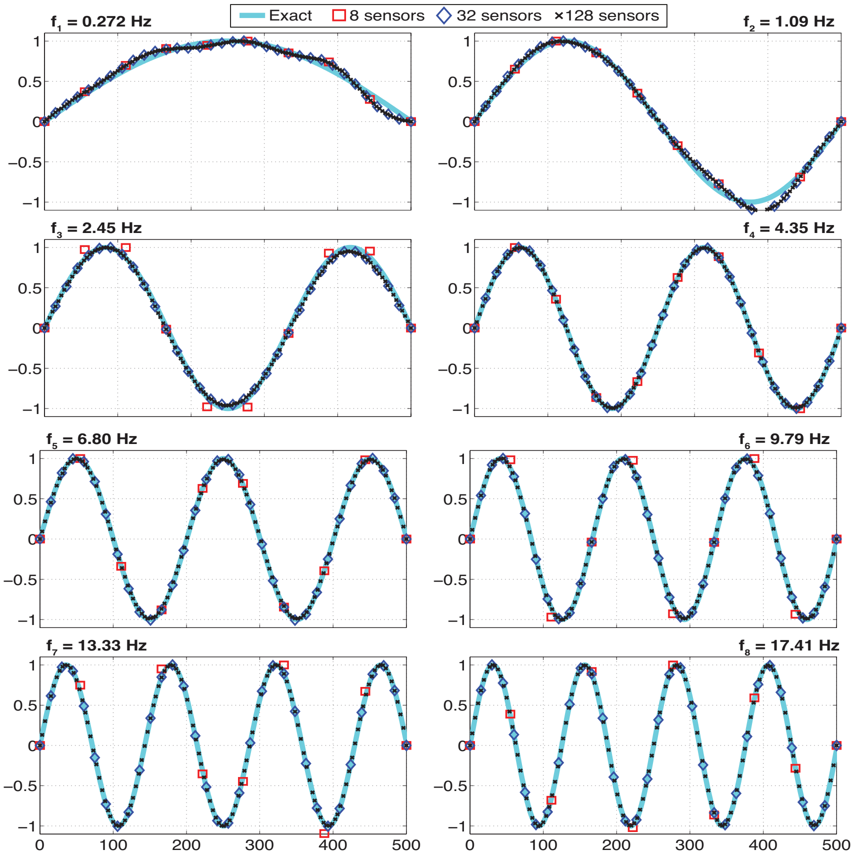

Figure 1. Eight identified mode shapes using ERA-NExT method.

Damage Localization with BIGDATA

In this section, the effect of network size on the results of data-driven damage detection methods is studied. The simple beam described in the previous sections is used to simulate undamaged and damaged structural conditions. Different sensor schemes are assumed for the beam and corresponding acceleration datasets are passed through a damage detection framework. The *angle coefficient* damage feature is selected, which is a scalar function of ARX model coefficients and is computed using pairs of sensor data. A sequential two-sample t-test is also applied for change point analysis. In this change detection technique, vectors of damage features are divided into two parts, for which the difference in their mean values is tested. The performance of this damage detection framework [20] has been validated for damage detection in a small-scale steel frame under impact loading.

In this paper, four sensor schemes are selected with 10, 20, 50, and 100 sensors, distributed evenly along the beam and ARX coefficients are created using acceleration data from each neighboring sensor pair. Undamaged and damaged ambient vibration data is generated by applying random excitation to the 500 DOF beam model. The damage is introduced as a 20% reduction in the section moment of inertia for 5% of the beam length at its center. The nodal accelerations are then contaminated with noise to account for operational and environmental variability. Thirty sets of random vibration data from undamaged and damaged conditions of the beam are generated for hypothesis testing: fifteen datasets represent a known, undamaged case, and the remaining fifteen are collected from an unknown structural condition. In this framework, a successful damage detection process should show signs of damage at the fifteenth set of unknown features (split number fifteen in the sequential t-tests).

For each sensor scheme, acceleration signals from specific DOFs are collected. The damage detection analysis is performed using Damage Identification Toolsuite (DIT) developed by the authors and is available for free at http://dit.atlss.lehigh.edu [21]. DIT offers several combinations of data training models, damage sensitive features, and statistical tests.

Figures 2 and 3 show the DIT output for damage detection in each sensor scheme using 7th order ARX models, angle features, and t-tests. In these plots, the timing of a potential damage would correspond to that split number where test statistics are maximized. Maximum values in the test statistics indicate the largest difference between the means of the two splits.

Figure 2a shows that with 10 sensors on the beam, the occurrence of damage is concluded. The timing of damage is not distinct among the sensor pairings, nevertheless the largest peak in the test statistics is consistent with the correct timing of damage (split number fifteen). While the sensor pairs above the threshold include the true location of damage, the set up is not successful in accurately detecting the extent of damage; damage is localized to 18% of the length of beam at its center, i.e., locations sensors 4, 5, and 6. Figure 2b shows the results of damage detection using 20 sensors. In this case, timing of damage is precisely identified at split number fifteen, and damage is correctly localized to 15% of the beam at its center.

Figure 3a shows the results of damage detection using 50 sensors; damage is clearly identified at split number fifteen, localized to its true location at the center of the beam with 6% of the length of the beam. The last case is damage detection using 100 sensors. Results of this case are shown in Figure 3b. In this case, the timing, location, and extent of damage are identified accurately.

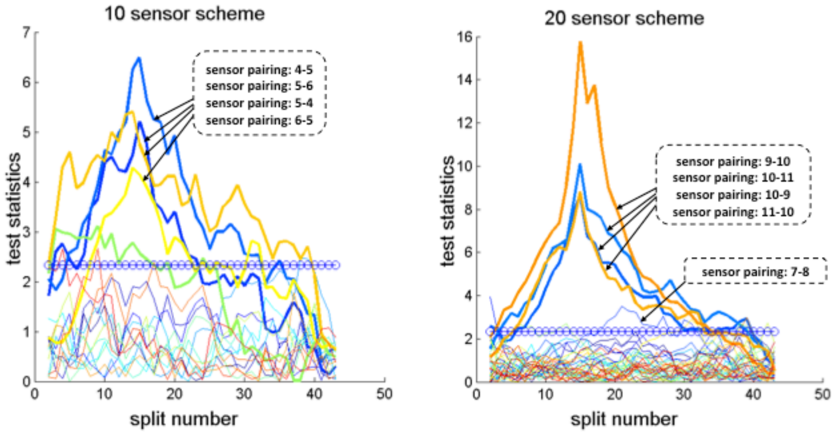

Figure 2. Damage detection results: (a) 10 sensor scheme (b) 20 sensor scheme.

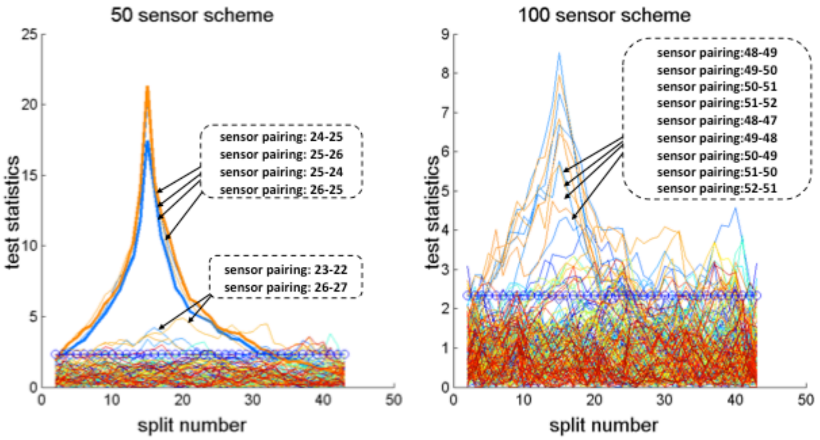

Figure 3. Damage detection results: (a) 50 sensor scheme (b) 100 sensor scheme.

SUMMARY, CONCLUSIONS, AND FUTURE WORK

Previous sections provided applications of BIGDATA in SHM for a simple beam structure. In SID, the smallest sensing scheme of 8 sensors was sufficient for identifying structural modal properties. As more sensors were included into the analysis, the required computational time grew cubically and estimated mode shapes became denser with no loss in accuracy. Additional sensors showed no clear improvement on frequency or damping estimates, indicating a limitation within the SID procedure. Therefore, for simple structures, the choice of BIGDATA in SID relies solely on the spatial resolution of the mode shapes solicited by the research question.

In damage detection, the occurrence of damage was accurately identified using the smallest sensing scheme of 10 sensors. The accuracy in estimation of location and extent of damage, however, increased with the number of sensing nodes used in the network. This increase became less noticeable when comparing schemes with 50 and 100 sensors, implying that any additional information would not be justified by the cost of using a denser sensing scheme. This result suggests that for accurate damage diagnosis in a BIGDATA setting, it is not required to process all the monitoring data; however, it is necessary to adopt certain feature extraction techniques and sensor selection schemes. While the datasets considered were very large with respect to the goals of each application, the data sizes were still manageable, in the sense that the SHM procedure was successful, i.e., the analyses may have been computationally expensive but remained possible. For more complex structures, corresponding BIGDATA sizes can be prohibitively large, causing SHM procedures fail. In anticipation of such cases, it would be beneficial to construct specific BIGDATA strategies aimed to reduce computational requirements of SHM procedures.

One strategy to reduce computational efforts for SID is through the use of offline (after data collection) Dynamic Sensor Networks (DSNs) [22]. In general, DSNs represent time-variant sensor configurations, e.g., a mobile sensor network, however, they can also be used to efficiently store an information-packed subset of BIGDATA. Offline DSNs are user-selected data matrices in which a vast amount of spatial information is condensed into a small size. With this technique, an equivalent amount of information is sought but at a significantly lower computational cost.

Computational expense of damage detection procedures can be reduced in several ways: (1) similar to the procedure presented in this paper, by applying novel feature extraction techniques to lower the dimension of problem, while damage detection performance is preserved, (2) through on-board data compression and transmission of compressed coefficients [23], and (3) by applying subset selection algorithms to accurately locate damage while only a subset of collected data are transmitted or processed [24-25].

ACKNOWLEDGEMENT

Research funding is partially provided by the National Science Foundation through Grant No. CMMI-1351537 by Hazard Mitigation and Structural Engineering program, and by a grant from the Commonwealth of Pennsylvania, Department of Community and Economic Development, through the Pennsylvania Infrastructure Technology Alliance (PITA).

REFERENCES

1. Laney, D. 2001. “3D data management: Controlling data volume, velocity and variety.” In *META Gr. Res. Note 6*.
2. Ivanov, T., Korfiatis, N., & Zicari, R. V. 2013. “On the inequality of the 3V’s of Big Data Architectural Paradigms: A case for heterogeneity.” *arXiv.org*, *cs.DB*. Retrieved from http://arxiv.org/abs/1311.0805v2\npapers3://publication/uuid/64A31441-C122-4219-91B5-10A799BB7FFE
3. Ward, J. S., & Barker, A. 2013. “Undefined By Data: A Survey of Big Data Definitions.” *arXiv.org*. Retrieved from http://arxiv.org/abs/1309.5821\npapers3://publication/uuid/63831F5F-B214-46D5-8A86-671042BE993F
4. Boyd, D., & Crawford, K. 2012. “Critical Questions for Big Data.” *Information, Commun. Soc.*, *15*(5), 662–679. doi:10.1080/1369118X.2012.678878
5. Marx, V. 2013. “Biology: The big challenges of big data.” *Nature*, *498*(7453), 255–260. doi:10.1038/498255a
6. Richtárik, P., & Takáč, M. 2012. “Parallel coordinate descent methods for big data optimization.” *arXiv Prepr. arXiv1212.0873*, *1*(June), 35. doi:10.1007/s10107-015-0901-6
7. Beyer, M. A., & Laney, D. 2012. *The importance of “big data”: a definition*. Stamford, CT.
8. Matarazzo, T. J., Shahidi, S. G., Chang, M., & Pakzad, S. N. 2015. “Are Today’s SHM Procedures Suitable for Tomorrow's BIGDATA?” In *Proc. Soc. Exp. Mech. IMAC XXXIII, Orlando, FL. Struct. Heal. Monit. Damage Detect. Vol. 7* (pp. 59–65.). Springer International Publishing. doi:DOI: 10.1007/978-3-319-15230-1_7
9. James III, G. H., Carrie, T. G., Lauffer, J. P., James, G. H., Carne, T. G., & Ill, G. H. J. 1993. *The Natural Excitation Technique (NExT) for Modal Parameter Extraction From Operating Wind Turbines*. Albuquerque, NM.
10. Andersen, P. (n.d.). “Identification of Civil Engineering Structures using Vector ARMA Models.”
11. Chang, M., & Pakzad, S. N. 2013. “Observer Kalman Filter Identification for Output-Only Systems Using Interactive Structural Modal Identification Toolsuite (SMIT).” *J. Bridg. Eng.*, *19*(5), 1–11. doi:10.1061/(ASCE)BE.1943-5592.0000530
12. Van Overschee, P., & De Moor, B. 1992. “N4SID: Subspace Algorithms for the Identification of Combined Deterministic-Stochastic Systems.” *Automatica*, *30*(1), 75–93.
13. Overschee, P. Van, & Moor, B. De. 1991. “Subspace algorithms for the stochastic identification problem.” *[1991] Proc. 30th IEEE Conf. Decis. Control*. doi:10.1109/CDC.1991.261604

14. Matarazzo, T. J., & Pakzad, S. N. 2015. "Structural Identification using Expectation Maximization (STRIDE): An Iterative Output-Only Method for Modal Identification." *J. Eng. Mech.* doi:10.1061/(ASCE)EM.1943-7889.0000951
15. Shahidi, S. G., Nigro, M. B., Pakzad, S. N., & Pan, Y. 2014. "Structural damage detection and localisation using multivariate regression models and two-sample control statistics." *Struct. Infrastruct. Eng.*, (September 2014), 1–17. doi:10.1080/15732479.2014.949277
16. Shahidi, S. G., Yao, R., Chamberlain, M. B. W., Nigro, M. B., Thorsen, A., & Pakzad, S. N. 2015. "Data-driven Structural Damage Identification Using DIT." In *Proc. Soc. Exp. Mech. IMAC XXXIII, Orlando, FL. Struct. Heal. Monit. Damage Detect. Vol. 2* (pp. 219–226). Springer International Publishing.
17. Matarazzo, T. J., & Pakzad, S. N. 2015. "A State-Space Model for Time-Varying Sensor Networks." In *Proc. Tenth Int. Work. Struct. Heal. Monit.* Stanford, CA.
18. Mascarenas, D., Cattaneo, a., Theiler, J., & Farrar, C. 2013. "Compressed sensing techniques for detecting damage in structures." *Struct. Heal. Monit.*, *12*(4), 325–338. doi:10.1177/1475921713486164
19. Yao, R., Pakzad, S. N., Venkitasubramaniam, P., & Hudson, J. M. 2015. "Iterative spatial compressive sensing strategy for structural damage diagnosis as a BIG DATA problem." In *Proc. Soc. Exp. Mech. IMAC XXXIII, Orlando, FL. Struct. Heal. Monit. Damage Detect. Vol. 2* (pp. 185–190). Springer International Publishing.