International Conference on Smart Infrastructure and Construction 2019 (ICSIC):
Driving data-informed decision-making

ice Publishing

DeJong, Schooling and Viggiani
ISBN 978-0-7277-6466-9
https://doi.org/10.1680/icsic.64669.381

CSIC Cambridge Centre for Smart Infrastructure & Construction

THE IMPORTANCE OF ANALYSING DATA FROM INSTRUMENTED INFRASTRUCTURE

F. D.-H. Lau [1,2] and Niall M. Adams[2,3]*

[1]*Lloyd's Register Foundation Programme on Data-Centric Engineering, The Alan Turing Institute, London, UK*
[2]*Department of Mathematics, Imperial College London, London, UK*
[3]*Data Science Institute, Imperial College London, London, UK*
** Corresponding author*

ABSTRACT Data acquired from instrumented infrastructure are at the heart of structural health monitoring. Engineers use information extracted from these data to improve their understanding of how the structures respond to stimuli. Little attention and emphasis, however, has been given to analysing the data before using a particular model. Analysing the data using appropriate statistical tools reveals structure that should inform the types of models deployed. In this paper, we emphasise the importance of statistically analysing data before fitting any model. These ideas are illustrated using data collected from a fibre-optic sensor network installed on railway bridges. These data capture the space-time response of the bridge through a sensor network. During periods of rest, we reveal that the collective sensor data exhibit a distinct dynamic latent structure, which is attributable only to the sensor system. To our knowledge this latent structure has never been documented before. This case study will illustrate that analysing data before implementing any procedure plays a vital role in structural health monitoring applications and if not done correctly, or at all, may lead to erroneous and misleading conclusions.

1. Introduction

In recent decades there has been a rapid increase in the number of physical structures instrumented with sensor networks (Bowers *et al*., 2016; Butler *et al*., 2016; Glisic *et al*., 2005; Lau *et al*., 2018a; Measures *et al*., 1992). These sensor networks are being deployed to facilitate the monitoring of structures to detect changes in their behaviour. The aims of structural health remain the same but the way questions are being approached has changed. Operators of instrumented infrastructures are now faced with numerous statistical and data science challenges that arise when collecting data using a sensor network (Lau *et al*., 2018b). We discuss some of these challenges through examples of railway bridges installed with a fibre-optic Bragg sensor network. In particular, we emphasise the importance of examining the data carefully before proceeding with statistical modelling. Analysing the data using appropriate statistical tools reveals structure that should inform the types of models deployed. Further, we describe how such an analysis can be used to guide model choices.

A major challenge in these settings is the analysis of spatio-temporal data. Such data capture the space-time response of the bridge through the sensor network, during periods of rest and train passage events. In both periods, we demonstrate that the individual sensors exhibit a clear banding pattern and temporal variations. During periods of rest, we reveal that the collective sensor data exhibit a distinct dynamic latent structure, which is attributable only to the sensor system. To our knowledge this latent structure has never been documented before and thus has not been directly incorporated into any models of such data. Modelling of these features and the latent structure, which explains a significant proportion of the variation in the data, is vital to understanding the baseline response of the sensor network. During train passage events this latent structure changes dramatically, the data being dominated by the event signal rather than the background sensor signal.

In Section 2 we discuss the data from a railway bridge instrumented with an integrated fibre-optic sensor network. Then in Section 3 we analyse the data remarking on notable features and discuss the statistical modelling implications. In Section 4 we discuss ways to use statistical models to detect changes in the structure behaviour.

2. Sensor Data

The sensor data considered throughout this paper is produced by a distributed network of fibre-optic strain sensors. These sensors use Bragg (National Instruments, 2018) gratings which refract light at discrete locations along the length of an optical fibre (Fibre-Bragg Grating or FBG). The sensor gratings are spaced 1 metre along the fibre-optic cable and each cable is 20 meters long. When the cable is subjected to strain the wavelength of refraction shifts. The fibre-optic cable also responds to variations in temperature due to the thermal expansion of the optical fibre. This is one of the environmental factors that contributes to the temporal variations observed in the data. Each sensor has an individual offset that allows the sensor analyser to distinguish deformations at the different Bragg locations. A sophisticated algorithm, converting the deformation to wavelength, involves a peak detection procedure that is used to determine the dominant wavelength at each time instance (Micron Optics, 2010). The change in

wavelength readings can be converted into the change of strain, through a linear transformation, which is the unit we shall work with throughout this paper.

In this work, we consider data from a railway bridge instrumented with FBG sensor networks. This steel-concrete half-through railway bridge located in Staffordshire UK is a 26.8 metre composite instrumented with 80 FBG sensors. The sensor network was installed within the bridges' main girder during construction. The sensor records are measured at a frequency of 250Hz. This bridge was completed in March 2016.

For brevity, we report results only for this bridge. However, a second instrumented bridge which is different in material and lifespan produces data with very similar properties. This second bridge is a masonry viaduct in Leeds retrofitted with an FBG sensor network. The sensor records are measured at a frequency of 1000Hz. These two instrumented structures represent physical assets at the start and end of their lifespans.

3. Analysis

Before analysing data, it is important to consider the objective of the analysis. In the setting of structural health monitoring, the main aim is to detect evidence of deterioration over time. This can be tackled in a number of ways. On way of doing this from a purely data driven point of view is to characterise the response of the bridge and the sensor network to stimuli.

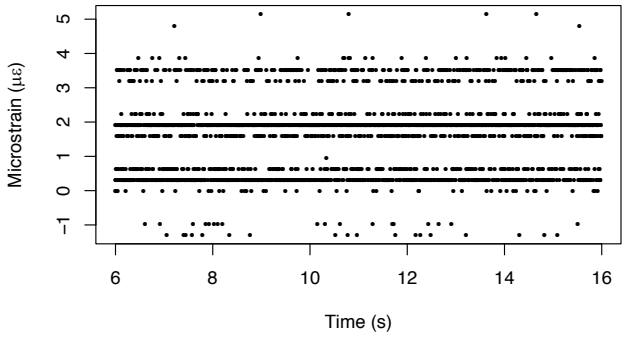
The first pitfall when beginning an analysis of data is failing to examine the data closely. Since there is a vast amount of data automatic plotting procedures can suppress structure. As noted earlier, the mechanism used to generate the records is complicated and induces specific structure in the observed data, as illustrated in Figure 1(a). Notably, this figure displays a period under which the bridge was subject to no stress. Observe first the banding structure which is an artefact of the measurement system, and second that the measurements change randomly from tick to tick. Collectively, this variation is sensor *noise*. This is observed for all sensors. The physics governing the behaviour of the bridge would suggest that the data should be constant during these rest periods. The key point here is that the sensor system is subject to its own noise process, which does not depend obviously on the status of the bridge. At present we are not unable to fully explain this banding structure.

Figure 1(b), like Figure 1(a), reports the strain values from a single sensor over a longer period. The banding pattern is less apparent; an artefact of the scale of the strain axis. From Figure 1(b) it is clear that temporal variation is present. In part, this variation over time can be attributed to temperature effects. This temporal variation raises a challenge in the analysis of the data, since the data cannot be treated as independent or identically distributed. A simple procedure to remedy this is given in Lau *et al*. (2018b). Further, it is clear that the distribution of these data is not Gaussian, and in fact, not clearly continuous.

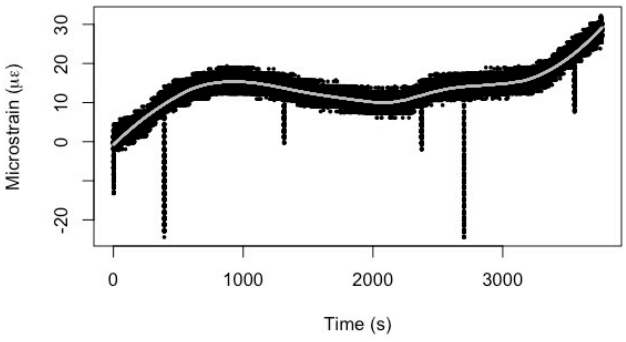
Studying sensor records individually does not capture the space-time phenomena of interest. As we have noted, the bridge and the sensor network are a spatially-distributed object which exhibits a response to temporal stimulus. This raises a number of challenges conceptually and statistically. Considering sets of sensors over time makes plotting difficult. To explore structures over sensors and time we can use simple summaries such as heatmaps of correlation matrices. Figures 2 and 3 display heatmaps of correlation matrices over *space* and *time* for a rest period and a train passage event. The data from rest periods and train passage events need to considered separately. If considered together, the strong correlation structure exhibited separately in time and over the sensors, displayed in Figure 3, would dominate; the response from a train passage event completely overwhelms the sensor noise. Figure 2(a) shows irregular correlation patterns in space. The correlation over time, Figure 2(b), exhibits some regular structure – a periodic pattern every 0.2 seconds.

Figure 1 Strain data from an individual sensor

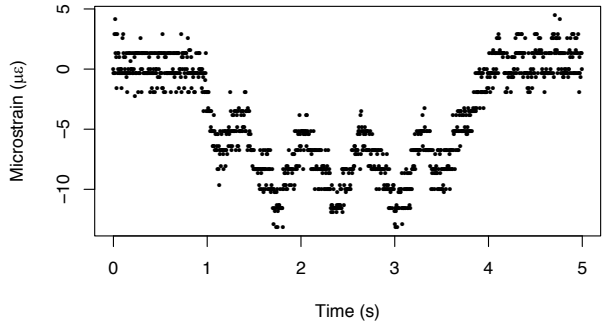
(a) Data extract from a rest period



(b) Long sensor stream for a single sensor consisting of nearly 1 million records (approximately 1 hour). A smoother (grey) is added to show temporal variation and extreme spikes correspond to train passage events.



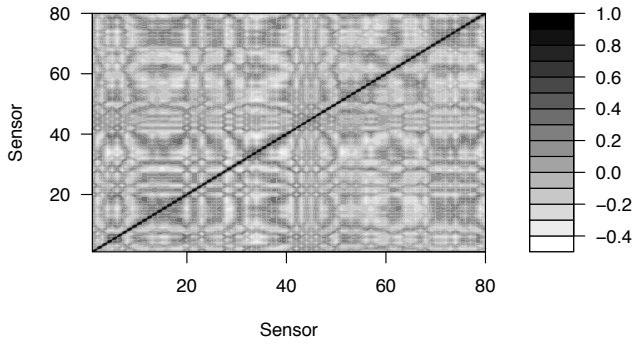
(c) Data extract containing a train passage event recorded by a single sensor.



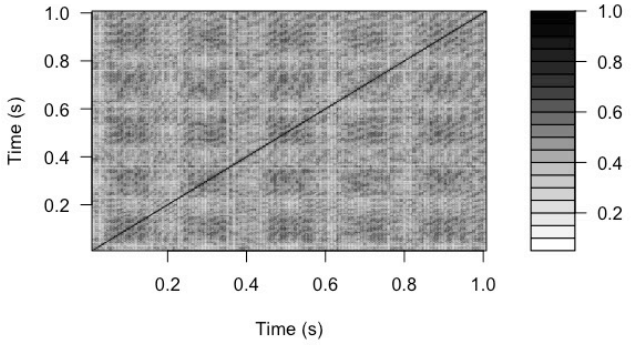
In contrast, the correlation structures during train-passage events have a regular and stronger pattern. Over space, the correlations exhibit a clear "chessboard" pattern, where every group of 20 sensors are highly correlated with each other and the groups are either highly positively or negatively correlated – see Figure 3(a). The explanation for this relates to the symmetric placement of the fibre optic cables on the bridge. The configuration of the sensor network is described in detail in Lau *et al*. (2018b).

Figure 2 Heatmaps of correlations for sensor data during a period of rest

(a) Heatmap of correlation of rest data over space



(b) Heatmap of correlation of rest data over time



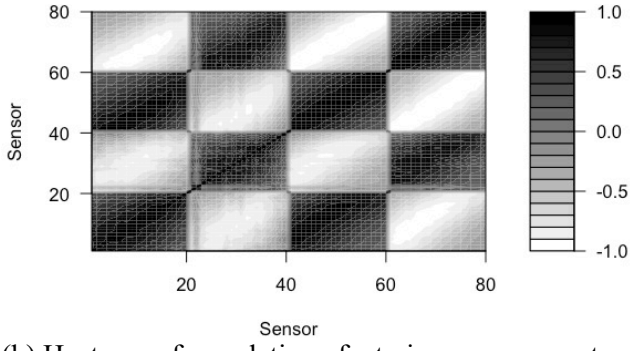
A train passage event manifests markedly different structure in the data and thus should be treated separately in reasoning about correlation. It is clear that the sensors cannot be treated as independent over space or time. The high correlation between the sensors explains why the linear model in Lau *et al*. (2018b) produces highly accurate predictions.

The collective time-space response of the bridge sensor system is most strikingly revealed in latent representations. For instance, when performing principal component analysis (Jolliffe, 1986) on data from periods of rest, the first 2 principal components explain 40% of the variance in the data. Put differently, an appreciable amount of information from the data is well represented in a 2D latent space. More interestingly still, and unreported as far as we know, is the specific configuration of points in this space – see Figure 4. Again, note that the *physics* that governs the response of the bridge to stimuli is not what is being captured by this latent 2D structure. This structure is an artefact of the sensor network, which is present in all rest periods, subsets of sensors, and over both bridges.

Modelling this latent structure is not straightforward. For the *doughnut* structure of the representation, one may appeal to circular statistics (e.g. see Fisher, 1995). However, circular distributions such as the von Mises and Rayleigh (Forbes *et al*. 2011) do not directly achieve the *hole* in the latent representation. Further, even if a model that captures the latent structure is found, there still remains the question of how to project back into the manifest space of the data.

Figure 3 Heatmaps of correlations for sensor data during a period of rest

(a) Heatmap of correlation of a train passage event over space



(b) Heatmap of correlation of a train passage event over time

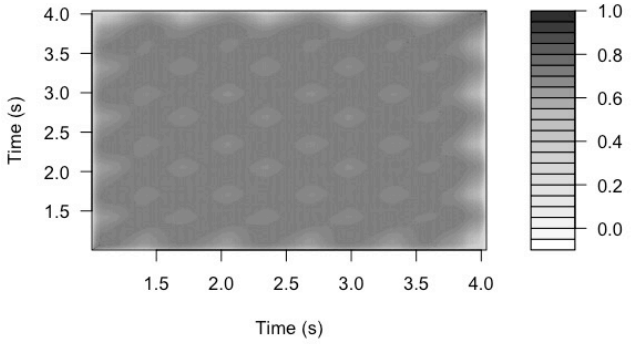
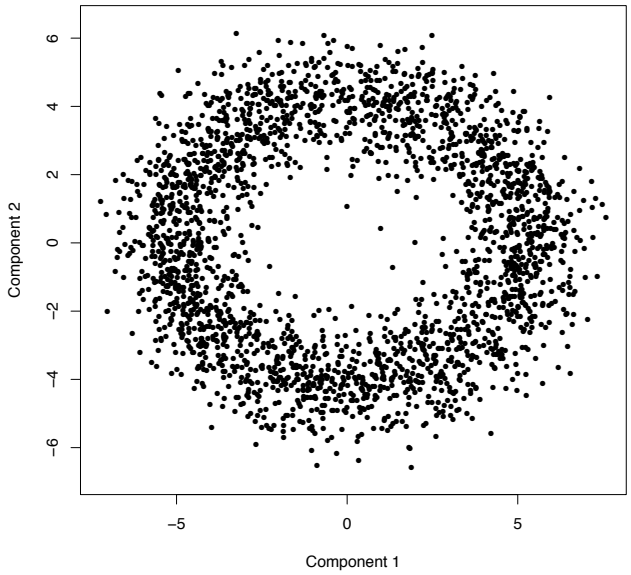


Figure 4 Latent space representation of sensor data. First two principal components.



4. Monitoring for changes

Suppose that a model can be constructed that captures the salient features described in Section 3. This model can provide the basis for detecting various types of changes of interest in structural health monitoring. The specific details of deployment depend upon the precise objectives for the asset.

The steel-concrete bridge is new and well-specified, therefore continuous monitoring is not necessarily required. In contrast the viaduct, being 150 years old, is at risk and merits continuous monitoring. A statistical model can be used in a sequential fashion, for continuous monitoring, such that each new record is tested and determined to be an anomaly. Any detection procedure will typically have a false detection rate i.e. a probability that the flagged anomaly is not truly an anomaly. When this procedure is repeatedly used, the overall probability of a false detection increases. This is a pertinent problem in our setting where data are revealed at a high frequency. Controlling false detection rates has been investigated (Benjamini and Hochberg, 1995; Benjamini and Yekutieli, 2001; Genovese and Wasserman, 2004) and applied to multiple data streams (Gandy and Lau, 2013). However, currently no procedure exists that is capable of controlling the false detection rate over an infinite number of data points.

Further, it is known for the viaduct that precise locations are more damaged than others, so monitoring for local changes is necessary. Monitoring for local or global changes dictates the type of model deployed.

Using the analogy to human health, deterioration may manifest as an increase in the recovery time of the system following a train event. In the case of the steel-concrete bridge, train passage event data could be used to support models designed to detect changes in the recovery time. Of course, that requires methods of automatically identifying the beginning and end of a train passage event.

Finally, and reiterating a core message of this paper, caution is required to distinguish the bridge response from the sensor network. This is most simply illustrated with an example. Suppose we are concerned with monitoring a specific location on an asset. Changes at that location will manifest as a result of either changes to the bridge response or, for example, sensor failure.

5. Summary and conclusion

Advances in technology are facilitating data collection from instrumented infrastructure at an unprecedented scale. This data deluge creates great opportunity for potential advances in structural health monitoring. However, the complexity of the data available provides new challenges for modelling. Among these challenges are: the volume of data, spatio-temporal correlation, unanticipated temporal variation, and complexity of the sensor noise process. The purpose of this paper is to provide a reminder about the importance of close examination of data prior to modelling and additionally to reveal specific latent data structures that occur in such systems.

Based on our exploration of the FBG sensor data, we have the following recommendations:

- First, inspect the data carefully. This will involve looking at graphical representations of the data. Simple features of the data may be overlooked when looking at numerical outputs of the data. This is especially important in situations where the data collection process is unclear. In our FBG data, by simply plotting the data for a single sensor, Figure 1(a), we see a banding pattern and that the data are clearly not continuous. These properties of the data can then be used to inform a modelling procedure. Further, plotting different parts of the data may reveal other properties. For instance, looking at the train passage event in Figure 1(c) suggests that there is no trend in the data: the microstrain values are approximately zero before and after the train event. However, looking over a longer time horizon, Figure 1(b) reveals a slow trend over time.

- Do not dismiss the sensor "noise" too quickly. Without other sources of information, the sensor data provides the only representation of the physical asset, even though it is corrupted by noise from the sensor analyser. Understanding the sensor noise will lead to more accurate statistical methods.

- Always remember what the data represents. In this work, the data is produced from a sensor network which is attached to a physical asset in the field. Therefore, the data captures the response of the physical asset *through* the sensor network which are both subject to environmental factors. Thus any conclusions based on this data refers to both the physical asset, the environment and the sensor network.

6. Acknowledgements

The authors would like to acknowledge the Cambridge Centre for Smart Infrastructure and Construction (CSIC), The Laing O'Rourke Centre at Cambridge and the Staffordshire Alliance (Network Rail, Volker Rail, Atkins and Laing O'Rourke) for providing the dataset used in the paper. The first author would like to acknowledge support from The Alan Turing Institute under the EPSRC grant EP/N510129/1 and the Turing-Lloyd's Register Foundation Programme for Data-Centric Engineering.

7. References

Measures RM *et al*. (1992) Fiber optic sensors for smart structures. *Optics and Lasers in Engineering* 16(2):127–152, 10.1016/0143-8166(92)90005-R.

Glisic B *et al*. (2005) Long-term monitoring of high-rise buildings using long-gauge fibre optic sensors. *In 7th International Conference on Multi-Purpose High-Rise Towers and Tall Buildings, Dubai, UAM, 10 - 11 December (on conference CD, paper #0416)*.

Bowers K *et al*. (2016) Smart infrastructure: Getting more from strategic assets. *Centre for Smart Infrastructure and Construction*.

Butler LJ *et al*. (2016) Integrated fibre- optic sensor networks as tools for monitoring strain development in bridges during construction. *The 19th Congress of IABSE Proceedings, Stockholm*, September 21-23, pages 1767– 1775.

Lau FDHL *et al*. (2018a) The role of statistics in data-centric engineering. *Statistics & Probability Letters* 136:58 – 62, 10.1016/j.spl.2018.02.035.

Lau FDHL *et al*. (2018b) Real-time statistical modelling of data generated from self-sensing bridges. *Proceedings of the Institution of Civil Engineers - Smart Infrastructure and Construction* 171(1):3–13, 10.1680/ jsmic.17.00023.

National Instruments (2018) Fundamentals of Fiber Bragg Grating (FBG) Optical Sensing. http: //www.ni.com/white-paper/11821/en/ (accessed 08/01/2018).

Micron Optics (2018) Sensing Instrumentation & Software user guide. http://www.sengenia.com/pdfs/Manual.pdf (accessed 01/08/2018).

Jolliffe IT (1986) Principal component analysis and factor analysis. In *Principal component analysis*. Springer, pp. 115–128, 10.1007/ 978-1-4757-1904-8 7

Fisher NI (1995) Statistical Analysis of Circular Data. Cambridge University Press.

C Forbes *et al* (2011) Statistical Distributions. Wiley.

Benjamini Y and Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289– 300.

Benjamini Y and Yekutieli D (2001) The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188

Genovese C and Wasserman L (2004) A stochastic process approach to false discovery control. *The Annals of Statistics*, 32:1035.

Gandy A and Lau FDHL (2013) Non-restarting cumulative sum charts and control of the false discovery rate. *Biometrika*, 100(1):261–268, 10.1093/biomet/ass06