ice | proceedings

Proceedings of the Institution of Civil Engineers
Pages 1–10

Keywords: Data-Centric Engineering; Instrumented Infrastructure; Sensor System.

Real-time Statistical Modelling of Data Generated from Self-Sensing Bridges

F. Din-Houn Lau[1,2], **MSci (Hon), PhD**

Liam J. Butler[2,3], **PEng, PhD**

Niall M. Adams[1,5], **BSc (Hon), PhD, CStat**

Mohammed Z.E.B. Elshafie[3,4], **BSc (Hon), MPhil (Cantab), PhD (Cantab)**

Mark A. Girolami[1,2], **BSc (Hon), PhD, FRSE, RAEng**

[1]Department of Mathematics, Imperial College London
[2]The Lloyds Register Foundation Programme on Data-Centric Engineering, The Alan Turing Institute
[3]Cambridge Centre for Smart Infrastructure and Construction, Department of Engineering, University of Cambridge
[4]Department of Civil and Architectural Engineering, Qatar University
[5]Data Science Institute, Imperial College London

Instrumentation of infrastructure is changing the way engineers design, construct, monitor and maintain structures such as roads, bridges and underground structures. Data gathered from these instruments have changed the hands-on assessment of infrastructure behaviour to include data processing and statistical analysis procedures. Engineers wish to understand the behaviour of the infrastructure and detect changes, e.g. degradation, but now using high frequency data acquired from a sensor network. Presented is a case study that models and analyses in real-time, the dynamic strain data gathered from a railway bridge which has been instrumented with fibre-optic sensor networks. The high frequency of the data combined with the large number of sensors requires methods that efficiently analyse the data. First, automated methods are developed to extract train passage events from the background signal and underlying trends due to environmental effects. Second, a *streaming* statistical model which can be updated efficiently is introduced that predicts strain measurements forward in time. This tool is enhanced to provide anomaly detection capabilities in individual sensors and the entire sensor network. These methods allow for the practical processing and analysis of large data sets. The implementation of these contributions will be essential for demonstrating the value of self-sensing structures.

1. Introduction

The potential of smart infrastructure to make more efficient use of existing and new assets has been estimated to be worth between 2 and 4.8 trillion globally (Bowers et al., 2016). At the centre of this shift toward making assets smarter is the advance and maturity of sensor development and deployment. However, the introduction of vast sensor networks within infrastructure has already begun to inundate owners, engineers and maintainers with large volumes and varied quality, velocity and variety of data. From a civil engineering perspective, instrumenting structures such as bridges has the potential to transform the design, construction, assessment and maintenance lifecycle phases. One of the main challenges lies in the development of innovative methods for managing, processing, analysing and interpreting the data obtained from smart infrastructure assets. A collaboration between engineers at the Centre for Smart Infrastructure and Construction (CSIC) at the University of Cambridge and data-scientists at the Lloyds Register Foundation funded Programme on Data-Centric Engineering (DCE) at the Alan Turing Institute is focused on addressing this challenge. DCE is a synthesis of approaches to studying physical engineering assets which leverages physics-based models which are updated based on measured data from the actual physical asset in operation and statistical (data-driven) models. This approach combines physical prior knowledge with empirical data providing for the physical asset a 'Digital Twin' (Lau et al., In Press). The current study focuses on development of the statistical models. Statistical techniques offer a means of monitoring structural health which does not require knowledge the structures behaviour. Instead, such models can be used to characterise the baseline (undamaged) state of a structure.

From the sensor network, we have long sequences of data which can be regarded as existing in one of two main states: when the bridge is under load (train passage events) and under no load. In reasoning about deterioration one might be interested in how quickly the bridge recovers after a train passage event. We provide a tool to extract the train passage events from such data (see later in Section 3.2). To monitor the bridges' instantaneous health, we need models that handle the high-frequency of the data. These models can then be deployed for anomaly detection, to identify departures

from the recent historical behaviour, as illustrated in Section 3.6. Identifying the timing and frequency of such anomalies will provide another mechanism for reasoning about degradation. We monitor this long-term degradation through the sensor system. The data is the response of the sensor system to stimulus (in this case the passage of a train over the bridge) and not the response of the bridge itself. Thus, through the data, we are reasoning about the recovery of the sensor network, and indirectly the bridge.

While there have been advances in recent years which have studied the application of statistical techniques in SHM, there is still significant scope for improvement and for introducing new concepts. Studies by Gul & Catbas (2009), investigated the use of autoregressive models (AR) in conjunction with an outlier detection algorithm based on the Mahalanobis distance. They validated their techniques based on two simplified laboratory steel beam and steel grid test specimens and under controlled ambient conditions. Rosales & Liyanapathirana (2017) investigated data obtained for a wireless sensor network attached to an experimental test frame. They employed both AR models and AR models with exogenous inputs (ARX) after Lei et al. (2003). Based on a comparison between both techniques, they concluded that the ARX model, while being more computationally costly, provided significant improvement over the AR model in its potential to better localise and quantify damage. Another study conducted by Noman et al. (2012) also utilised AR but applied the technique to a real structure, the Portage Creek Bridge in Victoria, Canada. They were able to use such techniques to conclude that little evidence of long-term deterioration was occurring within the structure. Another approach based on generalised Bayesian dynamic linear models (BDLMs) was proposed by Goulet (2017). Based on simulations, this study developed a framework for constructing, learning and estimating BDLMs whereby hidden effects such as daily and seasonal temperature variations and missing or outlier data could be incorporated.

While several previous studies have investigated various methods for modelling and interpreting data gathered from SHM systems, few have considered this challenge in the context of big data sets obtained from real structures and operating in real time. ‘Self-sensing’ or ‘sensory’ structures are those which contain an integrated sensor system for determining the state of the structure itself (Measures et al., 1992). Based on operational data gathered from a recently constructed ‘self-sensing’ railway bridge, this study proposes several solutions for batch and real-time processing of the data. In particular, the primary research contributions from this paper include:

- Development of a statistical method based on adaptive linear models for analysing and interpreting large and continuously updated data sets in real-time.
- Introduction of a real-time anomaly detection scheme based on individual and network sensor data.

2. Self-Sensing Railway Bridge

2.1. Sensor System

Completed in March 2016, a 26.8 metre composite steel-concrete half-through railway bridge located in Staffordshire U.K. was instrumented during its construction with a network of 134 fibre optic strain sensors (FOSS) (see Fig. 1). The FOSS are based on Bragg gratings (fibre Bragg gratings or FBGs) which represent periodic changes in the index of refraction which can be inscribed at discrete points along the length of an optical fibre. As the FOSS cable and inscribed FBG are strained, the initially inscribed Bragg wavelength shifts and can be converted to an equivalent strain via a photo-elastic coefficient. In addition to strain from mechanical effects (i.e. weight of passing trains, etc.), FBGs are sensitive to changes in temperature particularly in how it affects their index of refraction and due to the thermal expansion of the optical fibre itself. Therefore, when evaluating measurements taken by FBGs over periods of time whereby significant temperature changes occur, appropriate temperature compensation techniques must be applied. The use of FBGs in the sensing system was chosen for their improved accuracy, reliability and resistance to corrosion-based deterioration. In addition, up to 20 individual FBG sensors can be inscribed along a single optical fibre thereby greatly reducing wiring lengths and the number of interrogation channels. The FBGs installed as part of the SHM system were manufactured in low bend loss fibre with an additional glass-fibre reinforced polymer coating for added robustness during installation and operation. The FBGs along the optical fibre had inscribed Bragg wavelengths between 1510 nm and 1586 nm with an approximate strain accuracy of ± 4 microstrain.

FBGs were installed and measurements were recorded throughout the construction phase. Critical superstructure elements including the two main I-girders, the midspan cross beams, the midspan section of the reinforced concrete deck and the midspan vertical web stiffeners on the east main girder were instrumented. In addition, three prestressed concrete sleepers were manufactured with several FBGs installed along the top and bottom prestressing strands at the rail seat locations and at their midspan. These self-sensing sleepers were installed at the midspan of the bridge to correspond to the location of the instrumented cross beams. An overview of the monitoring system is presented in Fig. 2.

2.2. Monitoring Programme

The monitoring programme has been divided into two phases: during construction and during operation. Originally, the primary monitoring objectives included, 1) evaluating the robustness of the sensor network during construction, 2) establishing a comprehensive pre-operational performance baseline, and 3) developing analytical tools for long-term assessment, detection of damage (deterioration and/or anomalies) and management of self-sensing bridges. The first two objectives were previously addressed by the authors Butler et al. (2016b).

Figure 1. Installation of fibre optic sensors on bridge

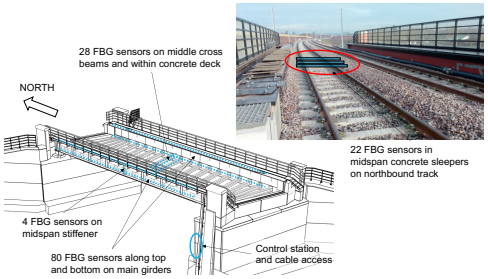

Figure 2. Fibre-optic based monitoring system.

Figure 3. Typical train types.

Operational data on the self-sensing bridge has been recorded since July 2016; several months after the bridge was opened to passenger trains. Since then, strain readings for all of the 134 FBG sensors have been recorded during the passage of over 140 trains. The sensing system is capable of recording data continuously at 250 Hz. The available data, which includes 140 train passage events and a period inactivity around the event, consists of more than 24,000,000 strain readings. Depicted in Fig. 3, two train types typically pass over the bridge, a British Rail Class 350 'Desiro' (4-car formation) and a Class 221 Super Voyager (4- or 5-car formation). These different types of trains cause different responses in the sensor network.

3. Statistical Analysis and Modelling

This section provides a brief description of the sensor data and presents efficient batch methods for extracting train passage events from large datasets.

The extraction of train passage events into a database is a necessary precursor to reasoning about degradation. Studying the historic response of the sensor network when a train passes and its recovery will provide a benchmark to compare against when reasoning about degradation.

An efficient streaming procedure for modelling sensor data while it is being collected at 250Hz is presented. The modelling procedure is used to address ambient (i.e. temperature) variations. Based on this streaming model, a method for tracking long-term deterioration (i.e. damage and anomaly detection) is introduced. The streaming model does not directly measure long-term deterioration, but provides a way of detecting more immediate changes in the sensor network to extract train passage events. Later, in Section 4, we discuss how to use these models to reason about future damage.

The distinction between batch and streaming is as follows. A batch procedure operates on a block of historic data which can be stored in memory and the procedure is able to pass repeatedly over the data. In contrast, a streaming procedure updates when new data arrives and, because of computational constraints, can only access the datum once. Moreover, a streaming procedure needs to handle unknown temporal variation, that is the phenomenon that the future will be different to the present for unknown reasons.

3.1. Sensor Data

The sensor system consists of 134 fibre-optic sensors located at different positions on the bridge. Each fibre-optic sensor records wavelength over time, which measures horizontal strains at discrete locations on the bridge superstructure. As noted earlier, each sensor collects data at a rate of 250Hz. Fig. 4 displays data collected from a single sensor showing two distinct states: the first is the train passage event highlighted in grey; the second is the unloaded state of the bridge. A distinct feature of the data is the banding pattern which arises from the pre-processing algorithm implemented by the fibre optic analyser. Fig. 4 presents all the data, although it seems that fewer than 250 datapoints are shown every second - this is a display artefact in conjunction with the strong banding pattern. The wavelength records can be converted to strain records as follows. Denote the wavelength at time t as λ_t, then the strain at t is

$$\epsilon_t = \frac{1}{1-\rho}\left(\frac{\lambda_t - \lambda_1}{\lambda_1}\right), \quad (1)$$

where $\rho = 0.22$ is the photo-elastic coefficient. More precisely, note that the strain is the change in strain relative to the first reading. In the following sections, the methods and models will use strain.

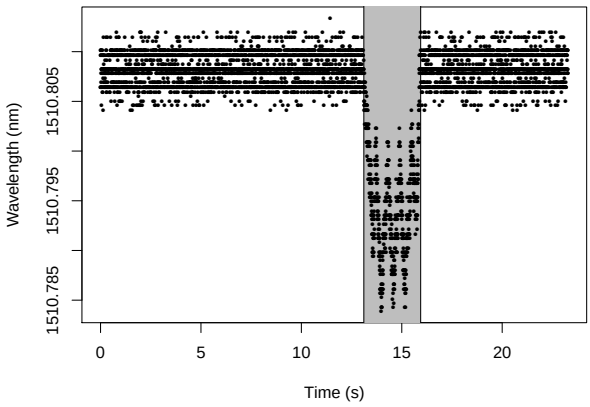

Figure 4. Data from a single sensor include a single train passage event.

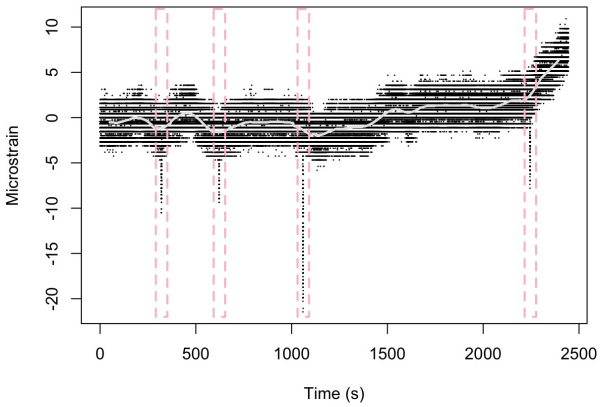

Figure 5. Data (black points) from a single sensor converted to strain. Moving average (grey line) captures the global trend of the data. The train passage events are highlighted in the pink dashed boxes .

3.2. Batch Processing of Large Datasets

This section presents a procedure that extracts the train passage events from large datasets consisting of many sensor records. Figure 5 presents the strain records from a single sensor from the top of the east main girder. Although the record length is only (approximately) 40 minutes, there are 611,108 data points for this single sensor. Considering the entire network of 134 simultaneously recording sensors, this corresponds to over 81 million data points – which certainly represents a big data problem. The four pronounced spike features, highlighted by the pink boxes in Fig. 5, are train passage events. A method that automatically extracts these events, using the data in Fig. 5 as a running example, is now introduced. The pseudocode for this procedure is presented in the Appendix.

As a first step, the main temporal variation in the data, which is likely due to variations in temperature during the data collection

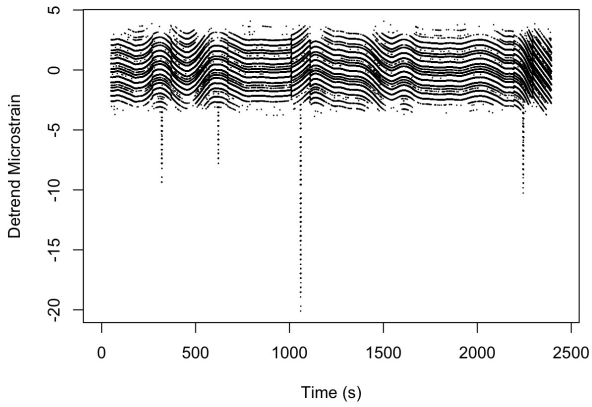

Figure 6. Detrended sensor data.

period, is removed. This temporal variation is estimated using the average of the data in a sliding window using $w = 25,000$ datapoints (100 seconds). The moving average is represented by the blue line in Fig. 5. This moving average is subtracted from the strain data which is then rescaled (see Appendix for details) - the result is presented in Fig. 6. The train passage event times can now be identified by their large variation in comparison to the background data. To quantify the variation, the standard deviation of the detrended data is computed in a batch fashion. This is accomplished by dividing the data into non-overlapping batches of length $v = 500$ datapoints. Then the standard deviation for each batch is computed. A threshold of $\gamma = 1.5$ microstrain is selected, such that a batch standard deviation above this threshold flags a train passage event. This procedure is repeated over all sensors. An alternative approach could be to treat the measurements across all sensors as a multivariate observation. The advantage of the procedure outlined above is its computational speed.

The outcome of this sensor-based procedure is a table of flagged events from each sensor with the number of sensors which suggested it (see Table 1). Notice that some of the train passage events times are within several seconds of each other. This is due to the delayed train response over the distributed sensor network or the peaks produced by the individual axles. This set of times is reduced using the following procedure. Any times that are within 2 seconds of each other are merged (see Table 2 for the result) since it is known that two train passage events cannot occur within this period. This knowlegde is based on the average train speed, train lengths and bridge length. These times are then used to isolate the individual train passage events. For instance, the event at time 321 can be extracted by cutting around the event time from time 321$\pm$ 5 secs.

This is a computationally efficient procedure for extracting train passage events from batch data. An example of an extracted event

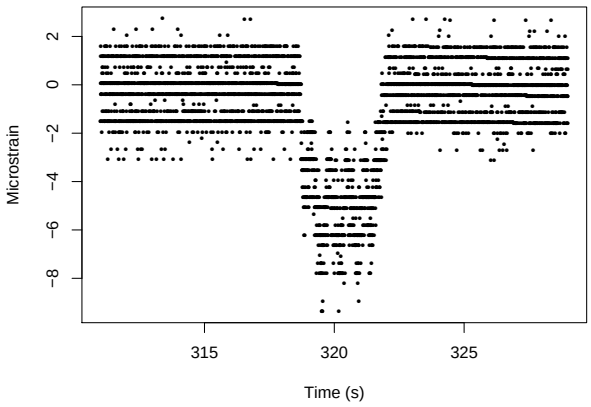

Figure 7. Extracted train passage event from a single sensor.

Table 1. Train passage event times across 134 sensors using extraction method.

Time	320	322	622	624	1060	1062	2244	2246	2248
Count	121	115	117	95	100	103	123	111	98

Table 2. Merged train passage event times across 134 sensors using extraction method

Time	321	623	1061	2246
Count	236	212	203	332

is presented in Fig. 7. In Section 4 we discuss the choice of control parameter values, w, v and γ.

3.3. Statistical Modelling

This section discusses how to sequentially monitor the sensors strain readings individually and collectively using a statistical model. In constructing such a streaming model, considerations need to be taken that address issues of computational efficiency, to handle data arriving at 250Hz, adaptation over time, to account for the temporal variation (i.e. due to temperature effects), and data storage.

Denote the strain record for sensor s at time t as $Y_t^{(s)}$. Further, denote the number of sensors as S. For sensor s, the strain is modelled as

(2)
$$M_s: \quad Y_t^{(s)} = \beta_0 + \sum_{u \in \{1,\dots,S\} \setminus s} \beta_u Y_{t-1}^{(u)} + \omega_t, \quad \omega_t \sim N(0, \sigma^2)$$

for $t = 2, 3, \dots,$ and where $N(\mu, \sigma^2)$ denotes a Normal distribution with mean μ and variance σ^2. The model M_s is a linear model that describes the strain measurements from sensor s at time t as a linear combination of all other sensor measurements at time $t-1$. Notice that this model is a one-step ahead forecast for sensor s, without using sensor s information. Models of the form of M_s are used for each sensor, primarily for computational speed in sequential updating settings. It is shown later that this model describes strain measurements from sensor s without using sensor s data, provides surprisingly accurate predictions.

The unknown parameters of M_s are β_j and σ^2. In batch settings these parameters are typically estimated using maximum likelihood or equivalently a least squares method. Fortunately, the linear structure of these models admits efficient sequential updating and allows the inclusion of a parameter called a forgetting factor which provides temporal adaptation.

3.4. Updating the Model

At a particular time t, only certain information has been revealed, namely $\{Y_\tau^{(s)} : \tau = 1, \dots, t; s = 1, \dots, S\}$. Refitting model M_s when new measurements are received is impractical due to the high data acquisition rate (250 measurements per second). Moreover, it would be undesirable to have a growing window of data due to temporal variation (see Fig. 5) and we seek to avoid using a sliding window. Therefore a recursive method to update the model parameters is used. This updating of linear models is called recursive least squares (see Chapter 9 in Haykin, 2002). This procedure will update the model parameters faster than the acquisition of new data (discussed later in Section 3.8). Further, this streaming regression has fixed computation and memory demand, and requires that no data need be stored.

3.5. Forgetting Factor

To account for the temporal adaptation in the data, a forgetting factor, $\lambda \in (0, 1)$, is introduced into the model M_s, which effectively puts more weight on recent data during the updating procedure. This approach was proposed in Haykin (2002, Chapter 9) and provides both temporal adaptation and efficient updating. For the purpose of exposition, a single fixed λ value is used, although it is possible to tune sequentially (e.g. see Anagnostopoulos et al., 2012).

The concept of the forgetting factor is now illustrated using a simple example. Consider computing the average for a sequence of values $x_1, x_2, \dots$ (for instance, strain measurements). Below are the recursive equations for updating the average value of $x_1, x_2, \dots$ with and without a forgetting factor λ. The average of the data at time t is denoted as $\overline{x}_t$ and $m_0 = n_0 = 0$.

No Forgetting Factor	Forgetting Factor
$m_t = m_{t-1} + x_t$	$m_t = \lambda m_{t-1} + x_t$
$n_t = n_{t-1} + 1$	$n_t = \lambda n_{t-1} + 1$
$\overline{x}_t = \frac{m_t}{n_t}$	$\overline{x}_t = \frac{m_t}{n_t}$

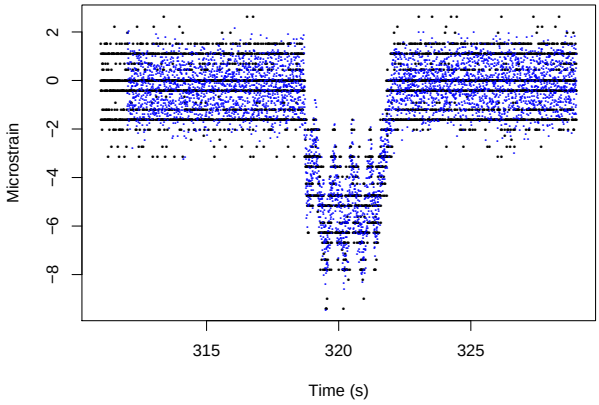

Figure 8. Predicted strain values for sensor 1 using model M_1 with $\lambda = 0.99$. Black points represent the observed measurements and blue points represent the predicted point estimated values from the model.

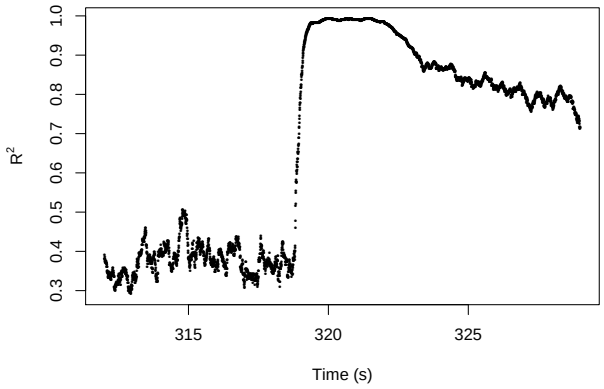

Figure 9. Modified R^2 statistic for model M_s with $\lambda = 0.99$.

Note that $\lambda = 1$ has no temporal adaptation, whereas as $\lambda < 1$ downweights the older data. The n_t is a value that, loosely speaking, describes the number of datapoints used in computing the average, akin to the effective sample size. This simple concept readily transfers to updating the parameter estimates of the linear model M_s.

So far, a statistical tool that can sequentially and adaptively predict the one-step ahead reading for sensor s given the previous tick of data from other sensors has been introduced. That is, the model provides a point estimate of the strain measurement of sensor s at time t. Figure 8 presents these point estimates and the true measurements from a single sensor. Before turning to the construction of an anomaly detection method, which requires a measure of uncertainty of the estimate, a statistic used to quantify the difference between the point estimate and the observed data is introduced.

A modified R^2 statistic, the coefficient of determination, that monitors the goodness-of-fit of a model is computed. This statistic measures how much variation in the data from sensor s is explained by the regression model. This version of the R^2 statistics slightly differs from the coefficient of determination commonly used with linear models, as it incorporates the forgetting factor. The modified R^2 statistic is

$$R^2 = 1 - \frac{\sum_t \lambda^{n-t}(Y_t - \widehat{Y}_t)^2}{\sum_t \lambda^{n-t}(Y_t - \frac{1}{t}\sum_{k=1}^{t} Y_k)^2} \quad (3)$$

where $\widehat{Y}_t$ is the prediction of Y_t.

Fig. 9 presents the modified R^2 statistic computed for the streaming model using $\lambda = 0.99$. Two features are notable in Fig. 9. First, the model provides a reasonably good fit throughout the observation period, indicated by the high R^2 values. Second, during the time of the train passage event, the model becomes increasingly accurate at predicting the sensor measurements (represented by R^2 values close to 1) indicating that the sensor readings move to a state of even higher correlation.

In Fig. 9 the R^2 values after the train passage event do not return to the values prior to the event. There are two plausible explanations for this. First, it takes the sensor network and bridge longer to recover than the observation period. Second, the choice of a fixed forgetting factor, $\lambda = 0.99$, is sub-optimal in respect to the estimation between different regimes. As noted earlier, this can be alleviated by using sequential methods for tuning of the forgetting factor.

3.6. Individual Sensor Changes

Based on the developed statistical model, an anomaly detection method is constructed through characterisation of the models' predictive uncertainty. This involves computing a p-value (or constructing a prediction interval) for the next data point and flagging the data point as unusual if falls below a given threshold. This procedure is first applied to a single sensor then extended to the collective sensor network.

The theory of linear models is used to construct a p-value which extends to the context of the developed streaming regression, provided the forgetting factor λ does not depend on the data. The core result is

$$Y_t^{(s)} \sim N(\widehat{Y}_t^{(s)}, var(\widehat{Y}_t^{(s)})), \quad (4)$$

where $\widehat{Y}_t^{(s)}$ is the one step ahead prediction for sensor s at time $t-1$. From this result, a p-value, p_s, can be computed. This is a measure of how surprising the new data point is with respect to the model. Figure 10 shows the sequence of p-values for a specific sensor. The far left side, highlighted by the grey area in Fig. 10, relates to the initialisation of the model, after which reasonable

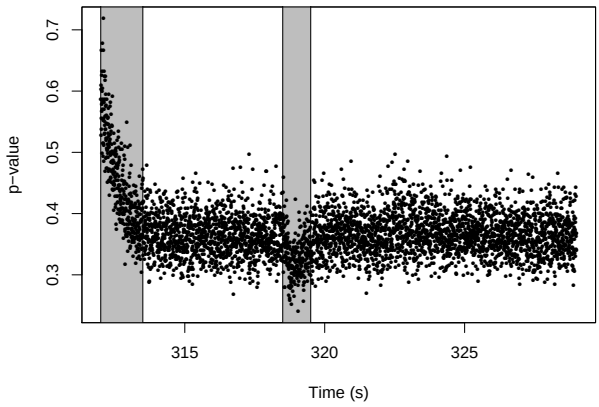

Figure 10. Sequence of p-values from M_s using $\lambda = 0.99$.

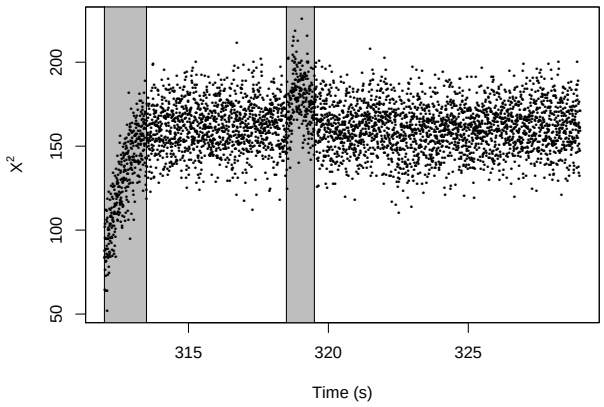

Figure 11. X^2 scores from 80 sensors located in the girders.

parameter estimates are obtained. Further, the decrease in p-values, highlighted by the centre grey area in Fig. 10, indicates the response of the sensor to a passing train.

To construct an anomaly detection method, the p-value, p_s, is compared with a threshold α, such that if $p_s < \alpha$ an anomaly is flagged. The choice of α is determined by the required detection sensitivity of the system. Any such statistical procedure will make mistakes by chance, and α is set to balance this false positive rate with the amount of detections that are of practical interest. Since hundreds of tests per second are being performed, the threshold α should be selected to be very small. Performing multiple tests will inevitably lead to a high false signal rate.

A flagged anomaly is the departure from the new observation from the postulated model M_s, which is based on the sensor measurements. This flagged anomaly could relate to the failure of the sensor i.e. debonding of the sensor from the structure, and/or damage of the bridge at the sensor location. Therefore, this anomaly detection method can be used to indicate locations on the superstructure where individual sensors may be faulty and/or elements of the structure are damaged.

3.7. Sensor Network Changes

In Section 3.6, an anomaly detection method for individual sensors was introduced, where a sequence of p-values for each sensor was produced. This anomaly detection method is only capable of flagging anomalies in individual sensors. From each individual sensor model, the p-values can be efficiently computed, and indeed models for all sensors can be computed at the same rate (or faster) as the data acquisition rate (i.e. 250Hz). To monitor the global response of the bridge, the p-values from all sensors are combined to provide an overall view of the collective sensor system response. Combination of p-values is well studied in statistics and Fisher's method (Fisher, 1925) is a popular approach. Under the null hypothesis that each model M_s is the true data generating model, Fisher's method uses the statistic

$$X^2 = -2\sum_{s=1}^{S} \log(p_s). \tag{5}$$

This X^2 follows a χ^2_{2S} distribution under the null hypothesis.

Fisher's method is motivated to combine independent p-values. While this is not true in this sensor network setting, the modifications required for dependent p-values requires knowledge of the dependence structure.

Figure 11 shows the X^2 scores for the 80 FBG sensors installed along the top and bottom flanges of the east and west main girders, for the data extracted in Section 3.2. These scores would again require comparison with a threshold to flag anomalies. In examining the X^2 values in Fig. 11 the train passage event seen (highlighted by the centre grey box). As discussed in Section

3.6 for the individual sensor p-values, the far left side relates to the initialisation of the model. The unloaded periods of the data are χ^2_{2S} distributed as indicated by Fisher's method. This X^2 statistic provides a collective summary of the response from the entire sensor network. As with the individual p-values, an anomaly detection method would require the X^2 score to be compared with a threshold, ζ, such that if $X^2 > \zeta$, an anomaly would be flagged. A flagged anomaly in this case would suggest a collective change in all the sensors. Such an anomaly detection method can used to indicate problems with the entire sensor network and/or the whole bridge.

3.8. Speed of Computation

The previous sections introduced a streaming linear model which sequentially and adaptively updates its parameter estimates, enhanced with a detection framework to flag changes in individual sensors and the entire sensor network.

To be practically useful, the model updating process and p-value computation must take less time than the arrival between two consecutive strain records. Our computations, on an offsite 1.7GHz laptop computer, show that updating a single model with a new observation and computing the p-value takes on average 2.3×10^{-4} seconds (with a standard deviation of 4×10^{-6}), which is faster than the 250Hz data rate. Parallelisation of the entire updating procedures across each model is possible and the additional effort of computing the X^2 statistic is negligible. The statistical software `R` was used in this study to perform the computations. Implementations of these methods in other programming languages such as C++ would lead to a significant increase in computation speed.

4. Results Discussion

The previous sections have presented a number of useful processing methods and statistical models for extracting critical structural response information provided by a fibre optic sensor network installed on a newly constructed railway bridge. Based on the developed models, a framework for tracking statistically significant changes in the individual sensors as well as across entire groups of sensors was presented. While the dataset considered in this study represents a relatively small sample of the total amount of data currently being generated from the self-sensing bridge, the statistical techniques developed are directly applicable to any size of dataset. This is important for the long-term monitoring of structures in which many months and years of sensor data may be required to be analysed and assessed. The following sections discuss additional considerations for statistical modelling, provide insight into how these techniques may be used in long-term SHM, and discuss the applicability of the developed techniques to other structure types.

4.1. Statistical Modelling

The statistical model and the anomaly detection methods provide the basis of a monitoring system capable of providing real-time updates on the bridges' sensor network health. There are a number of control parameters in the methods proposed throughout this work. In the processing of a large dataset (Section 3.2), the parameters are the width of the sliding window, w, the length of the non-overlapping batches, v datapoints, and the standard deviation threshold, γ. The values used in Section 3.2 lead to the extraction of all the train passage events from big datasets. For other applications, e.g. where the event signal is not distinct, the parameter values may need to be tuned. The forgetting factor, λ, used in the statistical model (Section 3.3) is another parameter which needs to be chosen. This parameter can be tuned using past data, however, the corresponding theoretical results, used to construct the anomaly detection method, no long hold. Moreover, this tuning would require further computational effort. The anomaly detection methods outlined in Sections 3.6 and 3.7 both require a threshold to be set in some fashion. For instance, for the collective sensor network summary discussed in Section 3.7, a cumulative sum (CUSUM) chart (Page, 1954) can be used to monitor the X^2 scores and signal a change. The CUSUM charts can be adapted to detect a change, e.g. a shift of the mean, in the distribution of a sequence of X^2 scores. The CUSUM chart methodology lends itself particularly well to this problem since it is a sequential method with quick updates.

4.2. Long-Term SHM

Traditionally, bridge condition monitoring and assessment is performed on the basis of visual condition surveys to provide a condition rating for the bridge. Self-sensing bridges allow for a data-driven approach to condition monitoring where assessment of a bridges' health can be based on the data gathered via the sensing system. The monitoring of the sensor network, discussed in Section 3.7, can be used to study groups of sensors e.g. west versus east main girders in order to assess their long-term load sharing ratio. These types of statistical modelling and anomaly detection methods may be used form the basis of a structural health monitoring system. For instance, if the X^2 scores which characterise the response of the sensor network (see Section 3.7), deviate from its known null distribution while the bridge is unloaded, then some global structural change may have occurred. Another way to monitor the structural health of the bridge is to compare similar train passage events (extracted by the method outlined Section 3.2) for changes. For similar trains, i.e. same number of carriages, similar mass etc, it would be expected that the bridge and sensor response be almost identical. Therefore, significant changes in the bridge/sensor response may indicate alteration in the structure.

Applying statistical techniques to long periods of data allows for the ability to characterise the effects of environmental factors (e.g. temperature and humidity) on the structural response of the bridge. These characterisations will enable more accurate models to be developed which will provide better measures of how the bridge deteriorates with time. If real-time processing of certain sensor datasets is not critical, implementation of other statistical techniques which are capable of damage identification and localisation are also possible. Ideas for addressing these issues, from a statistical standpoint, are discussed in Lau et al. (In Press) and form the basis of the authors' future work in this area.

4.3. Other Sensor and Structure Types

The discussion up to this point has focused on extracting information from and applying statistical techniques to strain data. However, it is worth noting that the strain data itself may be pre-processed in order to calculate other measures important for assessment of structural condition. These measures could include beam curvature, stresses, and neutral axis location, all of which could be modelled, tracked over time and used as indicators of long-term deterioration. Therefore, the techniques presented above may also be directly applied to other structural performance measures including strain. Data collected from other sensor types installed on a structure which continuously measure displacement, acceleration,

temperature, etc. can also be readily assessed and analysed using the proposed statistical methods. Steel-composite bridges are not the only structures which have been instrumented with permanent monitoring systems, for instance, another railway bridge composed of prestressed concrete girders and a composite concrete deck slab has also been instrumented with an integrated fibre optic sensor network and is currently being studied by the authors (Butler et al., 2016a). In addition, a variety of other structures reported in the literature including high-rise buildings (Glisic et al., 2005), tunnel linings and reinforced concrete foundation piles (Kechavarzi et al., 2016), have all implemented continuously recorded sensing systems. A variety of self-sensing structures, in which large sets of continuously collected data are required to be efficiently and expediently processed and analysed can leverage the statistical techniques presented herein.

5. Conclusion

This paper presents a big data case study in which a self-sensing railway bridge which has been instrumented with 134 discrete fibre optic sensors which record strain simultaneously at 250 Hz. A subset of the overall bridge monitoring dataset has been used in order to develop statistical tools which can process and analyse the data while being continuously updated. A new processing method for extracting useful operational information from long periods of sensor records was first presented. This fast, batch method extracts the individual train passage events within the large datasets and decouples the underlying background strain changes due to environmental effects such as temperature change. The extraction method will be of great practical use to engineers and operators who are tasked with quickly processing large sensor datasets generated from self-sensing bridges.

A recursive statistical model was then introduced that is able to update faster than the incoming recorded data, adapt to account for the temporal variation (i.e. due to environmental effects) in the data through implementation of a forgetting factor and accurately describe the data over time while requiring only minimal data storage. Based on these adaptive models, anomaly detection methods were developed and are capable of monitoring sensors individually (based on p-values) and across the entire sensor network (based on X^2 statistic). The X^2 scores which characterise the response of a group of sensors (and a component of the bridge) can be tracked over time for any deviations from their baseline distribution in order to provide an indication of deterioration and/or damage.

The statistical tools developed as part of this study will form the core component of a long-term SHM strategy in which considerations such as weekly, seasonal and yearly environmental trends can be characterised and used to update and create more robust prediction models. In addition, techniques developed in this study may be directly implemented in the analysis of other measured quantities (e.g. displacement, acceleration, temperature, etc.) and structure types (e.g. high-rise buildings, tunnels, etc.). This combination of a real-world self-sensing bridge case study and an innovative statistical framework for efficiently analysing the large monitoring datasets provides a valuable demonstrator for smart infrastructure systems.

6. Acknowledgements

The authors would like to acknowledge the Lloyd's Register Foundation, EPSRC and Innovate UK for funding this research through the Programme on Data-Centric Engineering at the Alan Turing Institute and through the Centre for Smart Infrastructure and Construction (CSIC) Innovation and Knowledge Centre. Research related to installation of the sensor system was carried out under EPSRC grant no. EP/L010917/1. Data related to this publication is available at the University of Cambridge data repository.

REFERENCES

ANAGNOSTOPOULOS, C., TASOULIS, D. K., ADAMS, N. M., PAVLIDIS, N. G. & HAND, D. J. (2012). Online linear and quadratic discriminant analysis with adaptive forgetting for streaming classification. *Statistical Analysis and Data Mining* **5**, 139–166.

BOWERS, K., BUSCHER, V., DENTTEN, R., EDWARDS, M., ENGLAND, J., ENZER, M., PARLIKAD, A. K. & SCHOOLING, J. (2016). Smart infrastructure: Getting more from strategic assets. *Centre for Smart Infrastructure and Construction* .

BUTLER, L. J., GIBBONS, N., HE, P., MIDDLETON, C. & ELSHAFIE, M. Z. (2016a). Evaluating the early-age behaviour of full-scale prestressed concrete beams using distributed and discrete fibre optic sensors. *Construction and Building Materials* **126**, 894 – 912.

BUTLER, L. J., GIBBONS, N., MIDDLETON, C. & ELSHAFIE, M. Z. E. B. (2016b). 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* , 1767–1775.

FISHER, R. (1925). *Statistical Methods For Research Workers*. Oliver and Boyd (Edinburgh).

GLISIC, B., INAUDI, D., LAU, J. M., MOK, Y. C. & NG, C. T. (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.*

GOULET, J.-A. (2017). Bayesian dynamic linear models for structural health monitoring. *Structural Control and Health Monitoring* , e2035–n/aE2035 stc.2035.

GUL, M. & CATBAS, F. N. (2009). Statistical pattern recognition for structural health monitoring using time series modeling: Theory and experimental verifications. *Mechanical Systems and Signal Processing* **23**, 2192–2204.

HAYKIN, S. S. (2002). *Adaptive Filter Theory*. Prentice-Hall information and system sciences series. Prentice Hall.

KECHAVARZI, C., SOGA, K., DEBATTISTA, N., PELECANOS, L., ELSHAFIE, M. Z. E. B. & MAIR, R. J. (2016). *Distributed fibre optic strain sensing for monitoring civil infrastructure - a practical guide*. Institution of Civil Engineers Publishing, Thomas Telford Ltd.

LAU, F. D.-H., ADAMS, N. M., GIROLAMI, M. A., BUTLER, L. J. & ELSHAFIE, M. Z. E. B. (In Press). The role of statistics in data-centric engineering. *Statistics & Probability Letters* .

LEI, Y., KIREMIDJIAN, A., NAIR, K., LYNCH, J., LAW, K., KENNY, T., CARRYER, E. & KOTTAPALLI, A. (2003). Statistical damage detection using time series analysis on a structural health monitoring benchmark problem. In *Proceedings of the 9th International Conference on Applications of Statistics and Probability in Civil Engineering*.

MEASURES, R. M., LEBLANC, M., LIU, K., FERGUSON, S., VALIS, T., HOGG, D., TURNER, R. & MCEWEN, K. (1992). Fiber optic sensors for smart structures. *Optics and Lasers in Engineering* **16**, 127–152.

NOMAN, A. S., DEEBA, F. & BAGCHI, A. (2012). Health monitoring of structures using statistical pattern recognition techniques. *Journal of Performance of Constructed Facilities* **27**, 575–584.

PAGE, E. S. (1954). Continuous inspection schemes. *Biometrika* **41**, 100–115.

ROSALES, M. J. & LIYANAPATHIRANA, R. (2017). Data driven innovations in structural health monitoring. *Journal of Physics: Conference Series* **842**, 012012.

Appendix

Data: Denote the strain measurement from sensor s at time t as $Y_t^{(s)}$ for $s = 1, \ldots, S$ and $t = 1, \ldots, T$.
Input: Moving average length w; Batch length v datapoints; Standard deviation threshold γ.
Output: Sensor s's train passage event times, U_s, for $s = 1, \ldots, S$.
for $s = 1, 2, \ldots, S$ **do**
Compute moving averages

(6)
$$Z_b = \frac{1}{2k+1} \sum_{j=-k}^{k} Y_{b+j+1}^{(s)} \quad \text{for } b = k, k+1, \ldots, T-k-1$$

Detrend the series

(7) $$\widetilde{Y}_b^{(s)} = Y_b^{(s)} - Z_b \quad \text{for } b = k, k+1, \ldots, T-k-1$$

Scale the detrended series

(8) $$C_b^{(s)} = \frac{\widetilde{Y}_b^{(s)}}{\sqrt{\frac{1}{2l}\sum_{\tau=b-l}^{b+l}(\widetilde{Y}_\tau^{(s)})^2}}$$

Compute the standard deviation in batches of length v

(9) $$\sigma_j = \sqrt{\frac{1}{v-1} \sum_{i=(j-1)v+1}^{jv} (C_i^{(s)})^2}$$

Identify large σ_j

(10) $$U_s = \{vj : \sigma_j > \gamma\}$$

end

Algorithm 1: Train Passage Event Extraction.