Statistics and Probability Letters 136 (2018) 58–62

Contents lists available at ScienceDirect

Statistics and Probability Letters

journal homepage: www.elsevier.com/locate/stapro

The role of statistics in data-centric engineering

F. Din-Houn Lau [a,b,*], Niall M. Adams [a,d], Mark A. Girolami [a,b], Liam J. Butler [c], Mohammed Z.E.B. Elshafie [c,e]

[a] Department of Mathematics, Imperial College London, United Kingdom
[b] The Lloyds Register Foundation Programme on Data-Centric Engineering, The Alan Turing Institute, United Kingdom
[c] Cambridge Centre for Smart Infrastructure and Construction, Department of Engineering, University of Cambridge, United Kingdom
[d] Data Science Institute, Imperial College London, United Kingdom
[e] Department of Civil and Architectural Engineering, Qatar University, Qatar

ARTICLE INFO

Article history:
Available online 26 February 2018

Keywords:
Big data
Data-centric engineering
Digital twin
Fibre-optic sensor
Instrumented infrastructure
Statistics

ABSTRACT

We explore the role of statistics for Big Data analysis arising from the emerging field of Data-Centric Engineering. Using examples related to sensor-instrumented bridges, we highlight a number of issues and challenges. These are broadly categorised as relating to uncertainty, latent-structure modelling, and the synthesis of statistical models and abstract physical models.

1. Introduction

Statistics is the science of transforming data into useful information. Raw data are seldom useful of itself, however, through data analysis, statistical summaries and models, informative conclusions can be reached. A key role of the statistician is determining which statistical tools to employ and drawing statistically sound conclusions. This has always been the case. In today's era of "Big Data", statisticians are confronted with an increasing variety of data from a growing number of applications, which has provided new challenges about the development, deployment and interpretation of statistical tools.

The term Big Data seems to implicitly make the promise: *The more data I have, the better, more accurate the results or the stronger conclusion*. A primary problem with Big Data, which are frequently collected automatically, is that little consideration is given to either biases in the collection process or the presence of a useful statistical signal. From a statistical standpoint, Big Data are collected without detailed thought about the nature of the statistical object which is the target for inference. This is unsurprising since the data collection is rarely the result of a design process.

Relatively cheap consumer hardware, and software systems, such as Hadoop (see for example White, 2012) and its successors, are partially responsible for the emergence of Big Data. There are many challenging computer science problems related to such infrastructure, as indeed there are computational challenges related to deployment of statistical methods at scale. Here we elect to reason about issues other than computation.

The main obstacle for statistical inference when analysing Big Data for a specific purpose is that the data are usually being co-opted to a purpose other than that for which it was collected. This obstacle is not new, and was considered a strength in data mining (see Hand et al., 2000, for a discussion).

[*] Corresponding author at: Department of Mathematics, Imperial College London, United Kingdom.
E-mail address: dhl@imperial.ac.uk (F.D.-H. Lau).

https://doi.org/10.1016/j.spl.2018.02.035

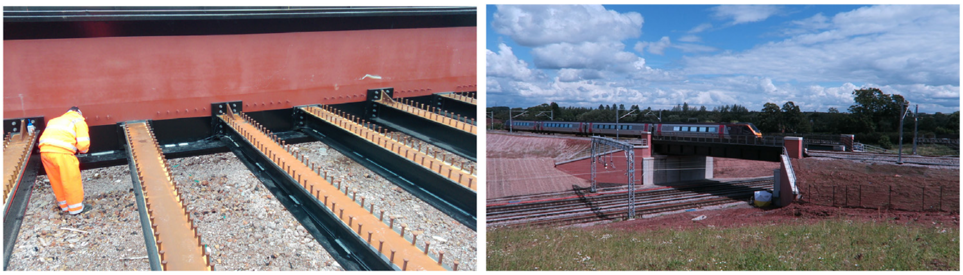

Fig. 1. Instrumented railway bridge. Left: Installation of sensor cables onto the steel cross beams. Right: Operational instrumented railway bridge.

Data-Centric Engineering[1] (DCE) is a synthesis of approaches to studying physical engineering assets which leverages mathematical physics-based models which are updated and refined based on measured data from the actual physical asset in operation and statistical (data-driven) models, combining physical prior knowledge with empirical data providing for the physical asset a *Digital Twin*.

The synthesis of the mathematical model with the data is unclear, and represents a core challenge for statistics and Big Data. The mathematical model embodies the researcher's prior knowledge on the engineering problem. This is prior elicitation at a meta-level, which goes well beyond the received version of the Bayesian paradigm. A possible way to combine is to embed the mathematical model in a statistical procedure, for example, in Azzimonti et al. (2015, 2014) where the regularisation penalty is constrained by knowledge embodied in a mathematical model. Alternatively, the data can be used to estimate unknown parameters in physical models, for example, estimating Young's Modulus in a beam equation — see Eq. (1). The ambition of DCE is to develop methodology that falls somewhere in between these two poles and essentially provides a single representation of physics and data.

We discuss the role of statisticians with Big Data, looking at DCE as an example. In DCE, engineers familiar with a specific tool kit, such as finite element analysis (FEA, see for example Ern and Guermond, 2013) and discrete event simulation, are facing new challenges as sensor data describing physical systems become abundant. To statisticians these problems have a familiar flavour. At the crux of DCE is the dichotomy between the presence of a signal arising from a sensor system contrasted with the availability of mathematical abstractions from an idealised construct. We illustrate these issues using data and concepts related to an instrumented bridge. Particularly, we explore issues of uncertainty, the relationship between physical models and sensor systems, and common-sense interpretation. The latter relates to the concern of engineers to provide simple, interpretable descriptions of anomalies arising from the interaction between the physical system and the sensor system.

2. Instrumented infrastructure and data

Fig. 1 displays a concrete and steel railway bridge constructed in 2016. During construction, the bridge was fitted with a fibre-optic sensor network system (see Butler et al., 2016 for details about the sensor installation procedure and locations). Initially, this system was installed to monitor the construction process. Now that the bridge is operational, the ambition of engineers is to use the sensor data system to address various issues, such as (i) understanding the behaviour of the bridge, and (ii) detecting anomalies, short and long term (such as degradation), in bridge behaviour as a step towards what engineers call "Condition-Based Maintenance" of physical assets.

Underpinning this ambition is the fundamental question of how to process, analyse and visualise the sensor data. Questions of this type have always been of interest to engineers, as they are concerned with the structural health of their infrastructure. The newly available sensor data acquired through instrumentation have provided different opportunities for behaviour assessment. Obviously, statistics provides the basic tools for such problems.

These instrumented infrastructure challenges typically constitute Big Data problems. For example, Network Rail[2] owns approximately 28,000 railway bridges all with the potential for instrumentation. If each bridge is instrumented with 100 sensors, each collecting data at 100 Hz, then in 1 day approximately 240,000 million records are produced. Needless to say, the storage, curation and processing of this amount of data are problematic.

Much modern infrastructure is being similarly equipped to address issues of monitoring. A recent example is the Queensferry Crossing (previously known as the Forth Replacement Crossing Bridge[3]) equipped a sensor system used to monitor the structural health of the bridge. It is in this context that we will explore the role of statistics in Big Data.

The related engineering field of structural health monitoring (SHM) (e.g. see Brownjohn, 2007 for a comprehensive overview and references therein) is concerned with developing systems that inform operators about the real-time condition

[1] https://www.turing.ac.uk/research_projects/programme-data-centric-engineering/.

[2] Network Rail owns and operates most of the railway infrastructure in England. https://www.networkrail.co.uk.

[3] https://www.transport.gov.scot/projects/forth-replacement-crossing/.

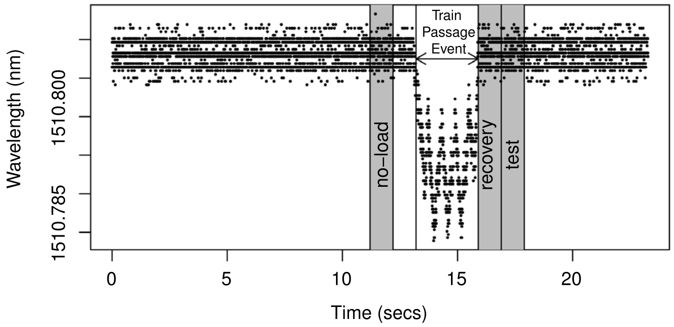

Fig. 2. Records from a single sensor over 23 s.

of their structures. Typically, in SHM, the main focus is on identifying features in the data that characterise the damaged and undamaged structure (referred to as feature extraction). However, little attention has been given to using statistical methods and models in the SHM literature. This is the difference between DCE and SHM.

2.1. Sensor data

For our bridge case study, the sensor system consists of 134 fibre-optic sensors located at different positions on the bridge. Each fibre-optic sensor records shifts in wavelength over time, which can be converted to measure strain at various locations along the bridge. Fig. 2 displays data collected from a single sensor when a 4-car passenger train passes over the bridge. There are some striking features. First, there are only 64 unique values among the 5819 sensor records. Second, there is a clear banding structure in the data. Third, there are clearly two regimes, which correspond to background (no load on the bridge) and a train passage event. Similar structure is apparent in all sensors over this observation period.

There are obvious statistical issues related to the discretisation and banding present in the data. A fundamental issue comes from the realisation that we are reasoning about the *sensor system* response to stimulus, and *not* the bridge response. This is apparent in the left part of Fig. 2, in which the bridge is under no load. The sensor system is still producing a signal, which a statistician would regard as background noise. A mathematical model of the bridge would not typically capture this source of randomness over the same period.

3. Mathematical models

Commonly, engineers use mathematical models that describe the physics of the engineering problem. These are most frequently used at the design stage, and often embodied in finite-element analysis and other methods. The problem of relating sensor data to such models remains open and challenging, and this presents a new opportunity to combine statistics, Big Data, and abstract physical models.

Mathematical models that exists describe the vertical deflections of a beam under load. Simplified models (e.g. Uzzal et al., 2012; Thambiratnam and Zhuge, 1996; Iwnicki, 2006, Chapter 6), usually based on an Euler–Bernoulli beam equation, can be used to approximate a single span girder bridge. As an illustration, the differential equation introduced in Iwnicki (2006, Chapter 6), models the deflection $w(x, t)$ at location $x \in \mathbb{R}$ at time t as

$$EI\frac{\partial^4 w(x,t)}{\partial x^4} + \rho A\frac{\partial^2 w(x,t)}{\partial t^2} = q(x,t), \tag{1}$$

where EI is the bending stiffness of the beam, ρ is the density, A is the cross-sectional area and $q(x, t)$ is the load on the beam. The beam is supported at the ends only giving the boundary conditions $w(x, t) = \frac{\partial^2 w(x,t)}{\partial t^2} = 0$ for $x = 0$ and $x = L$ where L is the length of the beam. Whilst under specific conditions this beam model has an analytic solution, it is a gross simplification of the bridge response. The beam is assumed to be uniform in density and material. More realistic models exist, such as the plate equations (e.g. see Hughes et al., 1977), and approaches based on stochastic PDEs (e.g. see Zibdeh, 1995) have also been proposed.

Realistic mathematical models such as these seldom admit an analytic solution. Typically, FEA is employed to obtain a numerical solution. Data are used to calibrate FE models; to align the physical quantities in the differential equations with the data. Further, the data are treated as gospel without error.

Another limitation is that these models do not directly account for slow degradation of the structure, which is of particular interest. This is also the case for more complicated models that take into account the different materials and the complex train–bridge interaction.

More relevant here is that none of these mathematical models take into account the sensor system response. As noted earlier, these sensor records relate to a sensor system which introduces errors and structures (such as banding), which are not modelled in equations such as (1). These structures are intrinsic to the sensor system itself, not the bridge.

4. Statistical models

There are a variety of direct opportunities for statistical modelling in the Big Data aspect of instrumented infrastructure. The deployment of such methods would naturally require characterisation of error structures, such as the discretisation observed in Fig. 2. For example, sequential anomaly detection procedures could be adapted for, and deployed on each sensor separately (see Gandy and Lau, 2013 for a CUSUM approach with FDR control), or preferably on the whole sensor system. Such procedures, building up from basic sequential change-point detection procedures, like CUSUM (Page, 1954), face a number of challenges for deployment in this "continuous monitoring" context (see Bodenham and Adams, 2016 for a detailed discussion).

A more interesting challenge relates to reasoning about the collective *time* and *space* response of the sensor system. Individually, the areas of time series and spatial statistics are mature in terms of their applications and theory. Spatio-temporal models, particularly in medical statistics, are well-studied. The ambition of DCE is not satisfied by such tools. We seek to reason about the *collective* response of the sensor system, and hence indirectly the bridge, in time and space. This is analogous to monitoring human health, where we reason about the overall health of the organism, which is a collection of spatially distributed interacting systems.

Short term monitoring of individual sensors, as discussed above, requires relatively straightforward modification of existing tools. A concern with such monitoring is distinguishing, for example, sensor change from bridge change. Longer term monitoring is more ambitious. Approaches for SHM exhibit a natural preference for physical testing of laboratory models, or non-destructive experimentation with the structure itself. Long term, we wish to reason about the slow degradation of the structure without using such direct techniques. This requires a different set of tools, and provides new challenges. One aspect relates to the ambition of curating, managing and using data in a manner that is suitable for an extended period, say 50 or more years in the future.

The statistical methodology issues relate to the analysis of such data. In our example, longer-term monitoring would relate to the bridge's return to a no-load state, after a train passage event. Fig. 2, displays, with grey bars, a no-load, *recovery*, and *test* period. During the test period, we would seek to determine when the bridge has returned to the no-load state that preceded the train passage event. The interval between the end of the train-passage event, and the return to no-load, is the recovery period.

The properties of this recovery period, particularly the duration, may give an indication of the degradation of the bridge. A simple way to reason about degradation is to examine only the sequence of recovery times, seeking evidence for either abrupt or gradual change. More complicated approaches would use the recovery period data itself, to characterise details of structural changes. A similar methodology could be used for the collection of all train passage events. In conducting such analysis over a long time frame, other phenomena would need to be accounted for, such as temperature. Of course, as in the discussion of the previous paragraph, such analysis should operate over the collective response of the sensor system, not individual sensors.

Notice that we are not directly interested in reasoning about the *failure* of bridges or other structures, but their degradation that will take place over decades, that may lead to a reduction in their functionality and reliability. Catastrophic failure of a structure takes place when either the structures is subjected to a large unexpected stimulus such as a train-structure collision or an earthquake event or as a result of slow-term degradation. We only have a chance of monitoring the latter through the sensor system.

5. Discussion

In the context of Big Data arising from DCE, there are various opportunities for novel statistical work. These are familiar statistical tropes, but require reformulation and up-scaling.

The process that generates the data is complicated, as illustrated in Fig. 2. The method of generating a single sensor value involves sophisticated physics and optimisation procedures. Developing probabilistic models that properly describe the data generation process is an excellent avenue for research. At a more ambitious level, the probabilistic model should capture temporal and spatial phenomena.

The development of appropriate statistical models for capturing the sensor system response over space and time that also incorporate the noise models described above, will provide the core element of the DCE Big Data toolkit. Such models will provide the basis of both short- and long-term monitoring tools. A challenge here relates to the problem of monitoring infrastructure over a very long time horizon, perhaps more than a hundred years.

Finally, the synthesis of abstract physical models, such as Eq. (1), with statistical methods (e.g. see Chkrebtii et al., 2016) for Big Data, will provide an unprecedented clarity of understanding the behaviour of instrumented infrastructure.

Acknowledgements

This work was supported by The Alan Turing Institute under the EPSRC grant EP/N510129/1 and the Turing-Lloyd's Register Foundation Programme for Data-Centric Engineering. The authors would also like to acknowledge EPSRC and Innovate UK (grant no. 920035) for funding this research 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/N021614. Data related to this publication are available at the University of Cambridge data repository.

The authors would like to thank the two referees for their insightful comments.

References

Azzimonti, L., Nobile, F., Sangalli, L.M., Secchi, P., 2014. Mixed finite elements for spatial regression with PDE Penalization. SIAM/ASA J. Uncertain. Quantif. 2 (1), 305–335. http://dx.doi.org/10.1137/130925426.

Azzimonti, L., Sangalli, L.M., Secchi, P., Domanin, M., Nobile, F., 2015. Blood flow velocity field estimation via spatial regression with PDE Penalization. J. Amer. Statist. Assoc. 110 (511), 1057–1071. http://dx.doi.org/10.1080/01621459.2014.946036.

Bodenham, D.A., Adams, N.M., 2016. Continuous monitoring for changepoints in data streams using adaptive estimation. Stat. Comput. 27, 1–14. http://dx.doi.org/10.1007/s11222-016-9684-8.

Brownjohn, J., 2007. Structural health monitoring of civil infrastructure. Phil. Trans. R. Soc. A 365 (1851), 589–622. http://dx.doi.org/10.1098/rsta.2006.1925.

Butler, L., Gibbons, N., Middleton, C., Elshafie, M., 2016. Integrated fibre-optic sensor networks as tools for monitoring strain development in bridges during construction. In: The 19th Congress of IABSE Proceedings, pp. 1767–1775. http://dx.doi.org/10.17863/CAM.5948.

Chkrebtii, O.A., Campbell, D.A., Calderhead, B., Girolami, M.A., 2016. Bayesian solution uncertainty quantification for differential equations. Bayesian Anal. 11 (4), 1239–1267. http://dx.doi.org/10.1214/16-BA1017.

Ern, A., Guermond, J., 2013. Theory and Practice of Finite Elements. Springer.

Gandy, A., Lau, F. D.-H., 2013. Non-restarting cumulative sum charts and control of the false discovery rate. Biometrika 100 (1), 261–268. http://dx.doi.org/10.1093/biomet/ass066.

Hand, D.J., Blunt, G., Kelly, M.G., Adams, N.M., 2000. Data mining for fun and profit. Statist. Sci. 15 (2), 111–131. http://dx.doi.org/10.1214/ss/1009212753.

Hughes, T.J.R., Taylor, R.L., Kanoknukulchai, W., 1977. A simple and efficient finite element for plate bending. Internat. J. Numer. Methods Engrg. 11 (10), 1529–1543. http://dx.doi.org/10.1002/nme.1620111005.

Iwnicki, S., 2006. Handbook of Railway Vehicle Dynamics. CRC Press.

Page, E.S., 1954. Continuous inspection schemes. Biometrika 41 (1–2), 100–115. http://dx.doi.org/10.2307/2333009.

Thambiratnam, D., Zhuge, Y., 1996. Dynamic analysis of beams on an elastic foundation subjected to moving loads. J. Sound Vib. 198 (2), 149–169. http://dx.doi.org/10.1006/jsvi.1996.0562.

Uzzal, R.U.A., Bhat, R.B., Ahmed, W., 2012. Dynamic response of a beam subjected to moving load and moving mass supported by Pasternak foundation. Shock Vib. 19 (2), 205–220. http://dx.doi.org/10.3233/SAV-2011-0624.

White, T., 2012. Hadoop: the definitive guide. O'Reilly.

Zibdeh, H., 1995. Stochastic vibration of an elastic beam due to random moving loads and deterministic axial forces. Eng. Struct. 17 (7), 530–535. http://dx.doi.org/10.1016/0141-0296(95)00051-8.