The role of Statistics in the era of big data: crucial, critical and under-valued

E Marian Scott

School of Mathematics and Statistics, University of Glasgow

Marian.scott@glasgow.ac.uk

Abstract

What is the role of Statistics in the era of big data, or is Statistics still relevant? I will start this rather personal view with my answer.

Statistics remains highly relevant irrespective of 'bigness' of data, its role remains what is has always been, but is even more important now. As a community, we need to improve our explanations and presentations to make more visible our relevance.

*Keywords***:** data, sampling, variability, inference, uncertainty

Introduction

I am a statistician, but does that mean I cannot be a data scientist or a data analyst and claim expertise and knowledge that are highly relevant in this era of big data? Does a data scientist have different skills? Should I feel threatened professionally, and is my discipline becoming irrelevant? We are all part of the "information ecosystem" (Al Gore, the Assault on Reason, 2017). We have moved from systems of rather clunky data acquisition, taking days, weeks, months to now the internet of things, social media, and the citizen scientist- with data arriving at unprecedented rates, where everyone may have access but may not have the skills needed to make sense of the data. Who checks the quality, who understands

bias and lack of representativity, who appreciates the variability? **Statisticians!** Authors in many disciplines write about the data deluge, and organisations sometimes describe themselves as data rich, information poor, increasingly there is discussion about the data and digital economies- who is well placed (but not uniquely) to swim and not sink in this ocean?. **Statisticians!**

I have not tried to define what is meant by data science, or professionally who is a data scientist or data analyst, though I would say surely Statistics is the science of data, and surely a statistician is a scientist studying data. I have been in situations where describing myself as a statistician is met with a groan, but in the same room and with the same audience describing myself and what I do as Data Science is much more accepted. I do not deny that there are subsets of skills that I do not have that another data scientist (who would not call themselves a statistician) might have (eg database architecture), but we don't expect everyone to be good at everything, in any field of science.

1.1 Are we the first generation to be faced with this revolution, or is this simply natural evolution?

Jerome H. Friedman wrote in 1997 (in The Role of Statistics in the Data Revolution (2001)) "There has been considerable change in the nature of data. In 1977 large and complex data sets were fairly rare, and little need was seen to attempt to analyze those few that did exist. This of course has changed. The computer revolution over the last 20 years has completely altered the economics of data collection. Twenty years ago most data was still collected manually. The cost of collecting it was proportional to the amount collected. This made the cost of collecting large amounts prohibitively expensive. The goal was to carefully design experiments so that maximal information could be obtained with the fewest possible measurements."

This would suggest we are living and working in a naturally and continually evolving system, and we need to evolve too. We need to respond to commentary such as “How statistics lost their power – and why we should fear what comes next. The ability of statistics to accurately represent the world is declining. In its wake, a new age of big data controlled by private companies is taking over – and putting democracy in peril”
https://www.theguardian.com/politics/2017/jan/19/crisis-of-statistics-big-data-democracy

Other headlines also make rather frightening reading including “THE END OF THEORY: THE DATA DELUGE MAKES THE SCIENTIFIC METHOD OBSOLETE”.
We are all familiar with the phrase “All models are wrong, but some are useful." attributed to George Box. Well according to some, “companies like Google, which have grown up in an era of massively abundant data, don't have to settle for wrong models. Indeed, they don't have to settle for models at all”. https://www.wired.com/2008/06/pb-theory/

So it might well seem that our *raison d’etre* is being challenged.

1.2 What are ‘big’ data?

In many quarters, there continues to be debate about what big data are- is size the only attribute that matters? The common definitions we see frequently presented refer to variety, velocity and volume as characteristics of ‘big’ data. For a statistician, size does matter but complexity is also a challenge and it is complexity and size and linkage of different data streams that for me present the real challenge. I use three examples from current research in the Environmental Statistics group at Glasgow to illustrate some of the issues we face.

Examples

The three examples I will describe are environmental which is a scientific field, where some of the benefits of big data are not yet fully realised, but where the technology of earth observation is evolving rapidly. There are a number of publications reflecting on the "data deluge" (Baraniuk, 2011) especially from networks of sensors. Data quality is another key data attribute, one which is of critical importance. (Campbell et al, 2013). The three examples I will use have coloured my views around big and complex data and what the statistical challenges are. All three connect to the growing use of sensors in environmental monitoring and all three research programmes have strong inter-disciplinary connections to Physics and Astronomy, Engineering, Environmental Science, Ecology and Biology.

Space weather

The first example is based on a current PhD project looking at space weather (Shu, 2017). This is a collaboration between statisticians, physicists and engineers. Space weather refers to electromagnetic disturbances in the near-Earth environment. Severe disturbances can have significant impact on national infrastructure and as such they are recognised on many national risk registers (often as one of the top ten risks). We have both space based (from satellites- the Cluster mission (http://sci.esa.int/cluster/)) and ground based (from fixed monitoring stations) observations of magnetic field fluxes and also complex physical models. The ultimate goal is to develop statistical tools to warn of severe disturbances in a timely manner to allow preventive action to be taken on the ground. The characteristics of this example would be data observed in space and time (but not in any sense over a regular grid), due to energy requirements satellites may be placed only in specific orbital configurations, so if we try and design a network we are constrained by physics. There are 4 dimensions (space and time) inextricably linked, there is a high frequency of data observations (eg every 4 secs), the solar wind is naturally varying, and interacting with the earth's magnetic field. These data need to be transmitted to ground stations and there are power constraints- in terms of the energy of transmission. The Cluster mission launched 4

satellites, there are plans to design and launch many more small satellites, can statisticians help design these missions?

Air pollution

The second example concerns the monitoring of air quality. Like the space weather example, new sensor technology has the potential to revolutionise our understanding of air quality and how it changes in space and time. New sensors (small enough to be worn by an individual) means that we can now track the path that our cyclist or pedestrian wanders through the city. Our mobile phones and other smart devices provide instruments for monitoring and alerting the citizen on high levels. However, while such sensors may be inexpensive, they may not be precise, there are also "*gold standard*" monitors (typically small numbers) to calibrate against. Such "*gold standard*" monitors are typically fixed in space, with all the statistical issues arising from fusing together data streams with different integration periods and different scales of spatial support (Huang, 2016). Naturally such data are linked to public health where there is much debate concerning the scale of health impacts from air pollution (see this very interesting piece from David Speigelhalter, https://wintoncentre.maths.cam.ac.uk/news/does-air-pollution-kill-40000-people-each-year-uk).

Lake water quality

My third example is drawn from a current NERC funded project Globolakes (http://www.globolakes.ac.uk/) and two associated PhD projects, (Gong, 2017 and Wilkie 2017). In this project, time series of satellite remote sensing images of 1000 lakes globally are being studied to identify how lake water quality is changing (and in related work) what might be driving the changes. Lake images range from a few thousand pixels to millions of pixels, with non-random missingness patterns (eg cloud cover in certain seasons) and with corresponding instrumental uncertainties presenting challenges to how we undertake our

statistical analysis. A natural way to think about the data structure is to recognise the hierarchy of spatial and temporal information at the pixel level within the lake, but then the spatial and temporal common patterns across lakes. In some but not all lakes, we also have 'ground truthing' usually from a very small number of locations within a lake where we have in-situ and more traditional water chemistry results.

In all three examples, we are searching for common patterns and trends, through extending and developing formal statistical tools including spatio-temporal modelling, missing data imputation, functional data analysis and some though not all are set in a Bayesian context.

Having painted this backdrop, I will now explain where I see my role realising that my text is highly personal and reflects my experiences in a variety of applied projects.

2. Data, information, meaning, communication (DIMC)

The world of data is changing with technological advances in data acquisition which needs to be matched to innovations in data management, storage and retrieval and ultimately in data analytics which are the processes and skills required in acquiring information from data. The many forms of data, their complexity and variations present challenges from data generating systems, to algorithms and inference about patterns through modelling leading ultimately to visualisation and communication. Statistics is key in this transformational sequence.

2.1 Data: the data generating process relies on our understanding of the system (the context) we are interested in. We frequently use statistical sampling ideas and surveys, or designed experiments (where we can) to generate data for our scientific enquiry. We need to understand the properties of the observation system we are using, specifically the technology being used (and crucially the uncertainties of measurement). Statisticians (and

therefore statistics) are crucial in the design of the sampling frame, and of the experiment. We recognise the intrinsic variability that exists in our observation systems, our statistical models capture this variation, and the uncertainties intrinsic in measurement. We relish messy data, by which I mean data that show variation, missingness, unusual and unexpected behaviours. We care about data quality.

Are these still relevant for "big data"? There have been views expressed such as "if organisations aren't already thinking about phasing out sampling and other "artefacts" of past best practices, they're behind the curve. Data science is inherently diminished if you continue to make the compromise of sampling when you could actually process all of the data, Sampling is an artefact of past best practices; it's time has passed." http://www.computerweekly.com/feature/Big-data-analytics-and-the-end-of-sampling-as-we-know-it

I note this quote refers to data science, but is that Statistics and would I agree that sampling is an artefact of the past? My own view is that in many cases, sampling is actually even more relevant, but that often we need to recognise that randomness may not be the primary goal, and that we need to be careful of selection bias, especially as increasingly in the environmental context (and in other areas too), there is a move towards citizen science where many of the newer data generation technologies are being rolled out through social media, smart phones etc. With the internet, a massive amount of information on species abundance can be collected by citizen science.

Do I agree that sampling is an artefact of the past? A simple answer is No, a more nuanced answer would be that sampling is still relevant, but that as statisticians, we need to look beyond randomness and recognise that in some situations, we have preferential sampling, that there will most likely be systematic biases and that our modelling (next step) needs to develop to accommodate this. An approach that is more frequently being discussed in the context of new data streams is Citizen science, which is not new, and for many years has

been commonplace in biodiversity monitoring, where volunteer based surveys have been widely used. However with the internet, a massive amount of new information on species abundance can be collected by citizen science programs. In these cases, newer forms of data streams (often unstructured) are being generated through social media, smart phones etc. "However, these data are often difficult to use directly in statistical inference, as their collection is generally opportunistic, and the distribution of the sampling effort is often not known." (Giraud et al, 2016). Similar statements could be made in other contexts, but the statement says difficult, not impossible.

2.2 Information: Once we are content with our data generation process and the data quality, the next step is statistical modelling and inference. There are many goals in scientific enquiry, simply put we can capture an overarching goal to be "to understand better the world we live in, its complexities and interactions, to demonstrate change and to understand what is driving that change". As statisticians, we use models for explanation and prediction. Our models may be chosen empirically based on exploratory analysis of the data, they can be a fusion of deterministic and stochastic models, where the stochastic element reflects variation in the population and uncertainty in our determinations. We have a wealth of modelling approaches to choose from, and an ability to create bespoke models for the context we are working in. We are all familiar and well steeped in the "all models are wrong, some models are useful", we are also commonly tasked on an everyday basis with thinking hard about the comparison of models with data or models with other models and in quantifying uncertainty.

What does a data scientist need? According to some, "6 types of analyses- Descriptive, Exploratory, Inferential, Predictive, Causal and Mechanistic".

https://datascientistinsights.com/2013/01/29/six-types-of-analyses-every-data-scientist-should-know/

Would we not suggest, that statisticians have these 6 types of analyses?

In this new data world, we are challenged by thinking about and designing fast and efficient algorithms (or re-formulating our models to be less dependent on matrix algebra), by working in high dimensions (a problem frequently solved by adopting dimension reduction techniques like principal components) and more recently by formulating our modelling in a functional data frame (especially when dealing with large numbers of time series of observations as we have in Globolakes).

"The statistics profession has reached a tipping point. The need for valid statistical tools is greater than ever; data sets are massive, often measuring hundreds of thousands of measurements for a single subject. The field is ready for a revolution, one driven by clear, objective benchmarks by which tools can be evaluated." (van der Laan and Rose, 2010*)* Though now 7 years old, the sentiment of this article remains true, we continually need this revolution (or at least evolution).

2.3 Meaning: the meaning (and sense) of what we discover from the exploratory and inferential steps of modelling comes in relating our findings to the scientific context of our enquiry. Statisticians (and especially applied statisticians as I am) work best in teams, we appreciate and relish inter-disciplinary work, it motivates and encourages to keep moving forward. Our results may raise additional questions, further work, contradict current understanding and spark new thoughts to be examined. Statistics and statisticians are often involved in projects and research involving other scientists and disciplines, so the meaning is related to the scientific context. Meaning also requires a critical eye, and a sense check- which is even more important than ever as the data complexity grows- statisticians are trained from birth to have such an eye.

2.4 Communication:

Finally, the last but certainly not least challenge is communication. Visualisation and info graphics are very much to the fore of current work, whether in newspapers (eg

https://flowingdata.com/tag/new-york-times/) or in Offices of National Statistics (ONS). In the UK, ONS infographics should be "Informative, effective, functional, honest and elegant" (https://gss.civilservice.gov.uk/wp-content/uploads/2012/12/infographics-guidelines.pdf). "*The graph is king*".

"Visualization is the means by which humans understand complex analytics and is often the most crucial and overlooked step in the analytics process. As you increase the complexity of your data, the complexity of your final model increases as well, making effective communication and visualization of data even more difficult and critical to end users. Data visualization is the key to actionable insights." HTTP://DATACONOMY.COM/2017/05/BIG-DATA-DATA-VISUALIZATION/

Part of the challenge remains our language and visualisations (one picture worth a thousand words) helps overcome some of the language hurdles, however, the audience need to be able to make sense of the sometimes rather subtle messages. We can and must do better!

3. Data governance, privacy, confidentiality and ownership

Finally, never far from our thoughts, there are many changes and challenges around data governance. There is an ever increasing emphasis on the discussion in our society (both scientific and civil) around data. Never have data been such a commodity, and with its value (financial and non-economic) (the data economy distinct but related to the digital economy) being debated. This is another key role for the statistical community to be engaged in, especially with our long history and skills in considering how to protect and anonymise personal data records.

4. Conclusions and final thoughts

Statistics has never been more relevant and challenged- from the discussions around the reproducibility of science (and of course the p-value arguments (Wasserstein and Lazar, 2016)), to the changing nature of a sample and sampling (with possible systemic biases), to technical challenges about the computational limitations of some models and to the messiness of data. Variation has not suddenly disappeared in the floods of data, our goals of explanation and prediction have not suddenly changed, and our desire to capture our uncertainty and communicate it in a way that is meaningful is ever more important.

Statisticians are delivering new methodologies and tools, in practical real-world systems dealing with large scale data challenges in areas as diverse as environmental monitoring, systems biology, neuro-imaging, and urban systems to mention only a few. We are neither irrelevant nor behind the times, and we will continue to work with our colleagues who are creating the new data streams to learn from data.

Acknowledgements

To my collaborators and colleagues (including Duncan Lee, Claire Miller, Ruth O'Donnell, Peter Craigmile, Andrew Tyler, Lyndsay Fletcher, Matteo Ceriotti, Stefan Reis), thank you for all the discussions around our shared projects. Thanks also to my students, Guowen Huang, Mengyi Gong, Qingying Shu and Craig Wilkie for all their hard work.

References.

Baraniuk R (2011). More Is Less: Signal Processing and the Data Deluge. Science 331.

Campbell, J, Rustad L, Porter J, Taylor J, Dereszynski E, Shanley J, Gries C, Henshaw D, Martin M, Sheldon W and. Boose E (2013). Quantity is Nothing without Quality: Automated QA/QC for Streaming Environmental Sensor Data. BioScience*. 63(7)*

Friedman J (2001). The Role of Statistics in the Data Revolution. International Statistical Review Vol. 69, pp. 5-10

Giraud C, Calenge C, Coron C, and Julliard C (2016). Capitalizing on Opportunistic Data for Monitoring Relative Abundances of Species. Biometrics 72, 649–658

Huang G (2016). Quantification of Air Quality in Space and Time and Its Effects on Health PhD thesis, University of Glasgow

Gong M (2017). Statistical Methods for Sparse Image Times Series of Remote-sensing Lake Environmental Measurements. PhD thesis, University of Glasgow

Gore A, (2017) The Assault on Reason, Penguin Press, New York.

Shu Q (2017). Statistical modelling of the near-Earth magnetic field in space weather. PhD thesis, University of Glasgow

Wilkie C (2017). Nonparametric statistical downscaling for the fusion of in-lake and remote sensing data. PhD thesis, University of Glasgow

Van der Laan M, Rose S (2010). Statistics ready for a revolution. Amstat news, Sept 2010.

Wasserstein R, Lazar N (2016). The ASA's Statement on *p*-Values: Context, Process, and Purpose. The American Statistician, 70(2).