Australian Bureau of Statistics

1351.0.55.054

Research Paper

Big Data, Statistical Inference and Official Statistics

www.abs.gov.au

New Issue

Research Paper

Big Data, Statistical Inference and Official Statistics

Siu-Ming Tam and Frederic Clarke

Methodology and Data Management Division

AUSTRALIAN BUREAU OF STATISTICS

EMBARGO: 11.30 AM (CANBERRA TIME) WED 18 MAR 2015

ABS Catalogue no. 1351.0.55.054

Views expressed in this paper are those of the author(s), and do not necessarily represent those of the Australian Bureau of Statistics.
Where quoted, they should be attributed clearly to the author(s).

Produced by the Australian Bureau of Statistics

INQUIRIES

The ABS welcomes comments on the research presented in this paper. For further information, please contact Dr Siu-Ming Tam, Chief Methodologist, Methodology and Data Management Division, on Canberra (02) 6252 7160 or email <Siu-Ming.Tam@abs.gov.au>.

BIG DATA, STATISTICAL INFERENCE AND OFFICIAL STATISTICS

Siu-Ming Tam and Frederic Clarke
Methodology and Data Management Division

EXECUTIVE SUMMARY

Official statisticians have been using a diversity of data sources in the production of official statistics for decades, including "designed" data sources such as censuses and surveys, and "found" data sources such as administrative and transactional data.

As a result of more and more interaction with digital technologies by citizens, and the increasing capability of these technologies to provide digital trails, new sources of data have emerged and are increasingly available to official statisticians. Such sources include data from sensor networks and tracking devices e.g. satellites and mobiles phones, behaviour metrics e.g. search engine queries, and on-line opinion e.g. social media commentaries. The collective term for such data sources is Big Data.

Whilst Big Data have the potential to create a rich, dynamic and focussed picture of Australia for informed decision making, and to improve the efficiency in the production of official statistics, this paper contends that there are a number of issues that an official statistician has to consider before deciding if a particular source from Big Data can be used for the regular production of official statistics.

A principal decision is business need and business benefit. This includes consideration of whether the new data source will improve the offerings of an existing statistical series, or plug statistical data gaps e.g. increasing the frequency of release, improving the richness of details such as small area or small population group statistics, or providing new official statistics that cannot be cost effectively provided using existing data sources. It also includes assessment of the business case in using the new data source, such as whether there will be a reduction of cost in the statistical production or reduction in provider load, and assessment of the quality of statistics produced from Big Data using Data Quality frameworks, against the benefits to be provided from the new source.

Another key decision is the validity of statistical inferences from Big Data. Big Data, depending on the source, suffer from one or more statistical biases, e.g. coverage bias, representational bias or self-selection biases, and measurement errors. Unlike errors due to sampling, the magnitude of these types of error will not be reduced by increasing the size of the data set.

The challenge for official statisticians is to develop a suitable methodology for analysing such data sets so that any conclusions drawn from the analysis are valid statistically. Firstly, official statisticians need a methodology to address any bias from Big Data, and secondly, a methodology in using Big Data to produce fit-for-purpose official statistics.

A Bayesian inference framework is adopted in this paper to assess the conditions under which valid statistical inference can be drawn from Big Data. The conditions are similar to those for making valid statistical inference from survey data: that any underlying process for the inclusion or exclusion of information from the Big Data source is independent of that information *per se*.

By treating Big Data as auxiliary information, and integrating census and survey data – ground truth data – with Big Data, this paper also provides a Bayesian method for using new data sources to produce official statistics. For count data, a dynamic logistic regression model is used. For continuous data, a dynamic linear model is described. The dynamic logistic model is applied to the theoretical analysis of satellite imagery data for the prediction of crop growing areas in Australia.

Other relevant issues for the official statistician to consider when deciding if a particular source from Big Data is to be used for the production of official statistics are: privacy and public trust, data ownership and access, computation efficiency and technology infrastructure.

Until recently, the Australian Bureau of Statistics' (ABS) progress in Big Data domain has been primarily review and monitoring of industry developments while contributing to external strategic and concept development activities. This paper summarises the ABS Big Data Strategy with objectives to build an integrated multifaceted capability for systematically exploiting the potential value of Big Data for official statistics.

This paper also describes the ABS Big Data Flagship Project, which has been established to provide the opportunity for the ABS to gain practical experience in assessing the business, statistical, technical, computational and other issues related to Big Data as outlined earlier in this paper. In addition, ABS participation in national and international activities on Big Data will help it share experience and knowledge, and collaboration with academics will help ABS better acquire the capability addressing business problems using Big Data as a part of the statistical solution.

CONTENTS

BIG DATA, STATISTICAL INFERENCE AND OFFICIAL STATISTICS

Siu-Ming Tam and Frederic Clarke
Methodology and Data Management Division

ABSTRACT

Whilst Big Data have the potential to improve the statistical production and statistical offerings, this paper outlines the issues that need to be considered by the official statistician, before a particular Big Data source can be used for the regular production of official statistics. In addition, the paper outlines Bayesian methods for analysing Satellite imagery data, and also the ABS strategies and initiatives on Big Data.

1. INTRODUCTION

Recent discussions and debates in the public domain about the opportunities presented by Big Data have now permeated into the sphere of official statistics – recent significant events include the discussion of a paper entitled "Big Data and Modernisation of Statistical Systems" by the United Nations Statistical Commission (2014), and the adoption of Scheveningen Memorandum on "Big Data and Official Statistics" by the Heads of European Statistical Offices (Eurostat, 2013). Whilst official statisticians have long been using administrative data and business data – one of the many sources for Big Data – in the production of official statistics, they are generally, and understandably, cautious in fully embracing this practice to other types of Big Data.

Almost always, the public discourse about Big Data is Information and Communication Technologies (ICT)-centric, and is largely preoccupied with the computing infrastructure, systems and techniques needed to effectively and efficiently handle the "volume, velocity and variety" of emerging Big Data sources. Translating this into the context of official statistics, it is about increasing the technological capability of a National Statistical Office (NSO) to capture, store, process and analyse Big Data for statistical production. Such debate raises a number of significant questions for official statistics, which are outlined below in increasing order of importance.

Firstly, is "Big Data technology" sufficiently mature to warrant an investment by the NSO? The widely-used Gartner Hype Cycle (Rivera and van der Meulen, 2013), which assesses the maturity of emerging technologies, places Big Data at the "Peak of Inflated Expectations" in 2013. It is considered unlikely that it will reach the "Plateau of Productivity" associated with mainstream uptake within the next five years.

Secondly, what is the likely benefit of using Big Data for official statistics, beyond that of administrative data and some types of business data? While there is undoubtedly some value in exploratory analysis of novel Big Data sources for opportunistic use, the proposition that a statistical producer should routinely acquire such data sets without an explicit business need, and business case, is tantamount to a Big Data solution in search of a problem. NSOs, faced with increasing budget pressure, are not willing to invest in Big Data unless there is a strong business case for investment.

Finally, how can Big Data be used to provide reliable and defensible statistical outputs? Crawford (2013) argued that "… hidden biases in both the collection and analysis stages present considerable risks, and are as important to the Big Data equation as the numbers themselves." The proposition that bigger datasets are somehow closer to the "truth" is not accepted by statisticians, since the objective "truth" is very much dependent on how representative a particular Big Data source is of the underlying population and the nature of the statistical inference drawn from such data. Other issues concerning the use of Big Data in Official Statistics are outlined in Daas and Puts (2014).

In spite of these issues and concerns, it is our view that Big Data, Semantic Statistics (Clarke and Hamilton, 2014), and Statistical Business Transformation (HLG BAS, 2012; Pink *et al.*, 2009; and Tam and Gross, 2013) are the three most promising initiatives for radically transforming the future business model and information footprint of NSOs. The Big Data challenge for official statisticians is to discover and exploit those non-traditional data sets that can augment or supplant existing sources for the efficient and effective production of 'fit for purpose' official statistics. Indeed, a number of international and national statistical organisations have already started to explore the potential for Big Data (United Nations Statistical Commission, 2014; Eurostat, 2013).

The purposes of this paper are to:

- Highlight some Big Data concepts, and outline concerns about the business value, methodological soundness, and technological feasibility of utilising Big Data for official statistics production;
- Provide a preliminary statistical framework for assessing the validity of making statistical inference for official statistics, and application of this framework in analysing count data from satellite sensing and magnitude data; and
- Present an outline of the statistical activities being undertaken in the ABS to assess the business case for using certain types of Big Data to replace or supplement an existing data source, to create new statistics, or improve the operational efficiency of the Australian Bureau of Statistics (ABS).

2. DEFINITION, USES AND SOURCES OF BIG DATA

What is Big Data?

According to the Big Data Privacy Report (Podesta *et al.*, 2014),

> " … there are many definitions of Big Data, which differ on whether you are a computer scientist, a financial analyst, or an entrepreneur pitching an idea to a venture capitalist."

Wikipedia (2014) defines it as

> " … a blanket term for any collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications."

Big Data is often defined by its characteristics along three dimensions (Daas and Puts, 2014):

- Volume – the number of data records, their attributes and linkages;
- Velocity – how fast data are produced and changed, and the speed at which they must be received, processed and understood; and
- Variety – the diversity of data sources, formats, media and content.

What are the uses of Big Data?

Manyika *et al.* (2011) argued that

> " … there are five broad ways in which using big data can create value. First, big data can unlock significant value by making information transparent and usable at much higher frequency. Second, as organizations create and store more transactional data in digital form, they can collect more accurate and detailed performance information on everything from product inventories to sick days, and therefore expose variability and boost performance. … Third, big data allows ever-narrower segmentation of customers and therefore much more precisely tailored products or services. Fourth, sophisticated analytics can substantially improve decision-making. Finally, big data can be used to improve the development of the next generation of products and services."

Another potential benefit of Big Data is in providing more regular, and timely information on interesting patterns such early indicators of epidemics, economic upturns or downturns e.g. Google's flu indicators despite its problems, unemployment or housing boom etc., thanks to the lower unit cost of acquiring Big Data sources than the traditional direct data collection methods used by NSOs. An excellent example of this is provided in Choi and Varian (2011) who also coined the term "nowcasting" to describe the process of predicting the present by harnessing information from Google Trends. In a blog to the Washington Post, Mui (2014) argued that the currency of statistics afforded by Big Data – readily available as a by-product of other collections – and how they could be "mined" for interesting patterns, are promising benefit of Big Data over traditional data sources.

On the other hand, Harford (2014) argued, whilst unearthing correlation from Big Data is cheap and easy, correlation, as statisticians have at pains been pointing out, is not the same as causation, and "… a theory-free analysis of mere correlations is inevitable fragile."

From the official statistics perspective, Big Data can be defined as statistical data sources comprising both the traditional sources and new sources that are becoming available from the "web of everything". Whilst the volume and velocity of Big Data are huge and are beyond current data management or processing capabilities, we contend that NSOs do not necessarily need to use the full data set for the production of official statistics, as sampling methods may be applied to provide fit-for-purpose statistics.

Whilst not all Big Data variety are suitable for the production of official statistics, they have the potential to increase the cost efficiency of NSOs, provide new statistical products and services, and increase the frequency in the production of official statistics at little additional cost to NSOs. Big Data may provide an opportunity for NSOs to better fulfil its mission in the provision of official statistics for informed decision making. However, we contend that decisions on which Big Data source to use, including decisions on volume, velocity and variety, have to be assessed against the cost-benefit criteria outlined in the latter part of this paper.

Collectively, the wide variety of extant and emerging Big Data sources of interest to official statistics may be broadly categorised as follows (United Nations Statistical Commission, 2014):

- Sources arising from the administration of Government or private sector programs, e.g. electronic medical records, hospital visits, insurance records, bank records etc.. The source from Government programs has traditionally been referred to as administrative sources by official statisticians;
- Commercial or transactional sources arising from the transaction between two entities, e.g. credit card transactions and online transactions (including from mobile devices);
- Sensor networks sources, e.g. satellite imaging, road sensors and climate sensors;
- Tracking device sources, e.g. tracking data from mobile telephones and the Global Positioning System (GPS);
- Behavioural data sources, e.g. online searches (about a product, a service or any other type of information) and online page views; and
- Opinion data sources, e.g. comments on social media.

Censuses and surveys and the first two sources above, i.e. administrative data, and to a limited extent, business data (e.g. scanner data from supermarkets, motor vehicle sales data etc.) are currently the principal sources for the production of official statistics. Big Data open up opportunities for new data sources for NSOs.

While some of the Big Data sources are identifiable (e.g. satellite sensing data with pixel longitudes and latitudes), many others are not (e.g. prices of on-line goods and services, scanner data, or commercial transactions). Both identifiable and unidentifiable data sources have their respective uses in official statistics. For instance, satellite sensing data can be combined with data provided by farmers in agricultural surveys at the unit record level, whereas on-line prices data can be used in calculating the price relatives for use in Consumer Price Index (CPI) compilations. The challenge for official statisticians is to find effective and valid ways of utilising the Big Data sources, where their use in the regular production of official statistics in justified.

3. BIG DATA AND OFFICIAL STATISTICS

Many NSOs in the world, including the ABS (ABS, 2013), have developed significant and relevant expertise in collecting and processing large amounts of data, and are:

- empowered under its legislation to compel the provision of information by providers for the purposes of producing official statistics;
- an authorised integrator of sensitive data under the statistics legislation;
- given its holdings on statistical benchmarks, uniquely to assess the quality and "representativeness" of Big Data sources;
- able to produce statistics that are of high quality – so that users can be assured that the information they are using is 'fit for purpose'; and
- independent, of high-integrity and impartial. Most NSOs publish the concepts, sources, methods, and results of all collections, and it provides "level playing field" access to all users of official statistics.

Together with the high level of community trust placed on official statistics (see for example ABS, 2010b), these attributes put many NSOs in a good position to experiment with, and explore, the potential use of Big Data.

4. OPPORTUNITIES AND CHALLENGES FOR OFFICIAL STATISTICS

To continue to improve its statistical value proposition, many NSOs strive to reduce the cost of statistical production, improve the timeliness and frequency of its offerings, and create new or richer statistics that meet emerging statistical data needs. As part of its business transformation program to deliver on these aspirations, some NSOs (e.g. the ABS, Statistics Netherlands, the Italian Statistics Office to name a few) are undertaking initiatives to exploit particular Big Data opportunities.

It is our view that a number of applications of Big Data may be identified by drawing parallels with the well-established use in official statistics of administrative data, provided that the sources meet the benefit criteria and statistical validity issues outlined in this paper. These applications include:

- sample frame or register creation – identifying survey population units and/or providing auxiliary information such as stratification variables;
- full data substitution – replacing survey collection;
- partial data substitution for a subgroup of a population – reducing sample size;
- partial data substitution for some required data items – reducing survey instrument length, or enriching the dataset without the need for statistical linking;
- imputation of missing data items – substituting for same or similar unit;
- editing – assisting the detection and treatment of anomalies in survey data;
- linking to other data – creating richer datasets and/or longitudinal perspectives;
- data confrontation – ensuring the validity and consistency of survey data; and
- generating new analytical insights – enhancing the measurement and description of economic, social and environmental phenomena.

While the primary focus is the exploitation of Big Data largely for richer or more timely statistical offerings, Big Data generated in-house can also be used to improve the efficiency of statistical operations of NSOs (Groves and Heeringa, 2006). These include:

- improving the data provider and data consumer experiences;
- improving the operational business efficiencies; and
- monitoring the Web and network security and end-user network experiences.

5. BUSINESS BENEFIT

The decision to use a particular Big Data source in statistical production should be based strictly on business need, and the prospective benefit established on a case-by-case basis – how it might improve end-to-end statistical outcomes in terms of objective costs-benefit criteria. That is, the costs and benefits of using the new data source need to be assessed in terms of factors such as reduction in provider load, sustainability of new source, as well as the accuracy, relevance, consistency, interpretability, and timeliness of those outputs stipulated in Data Quality Frameworks (ABS, 2010a; Brackstone, 1999; OECD, 2011).

As an example, the full data substitution of survey-based with satellite sensing data for producing agricultural statistics – such as land cover and crop yield – can be assessed as follows:

- Costs – What are the likely costs for acquiring, cleaning and preparing the satellite sensing data in a form suitable for further official processing, noting that the computational demands of acquiring, transferring, processing, integrating and analysing large imagery data sets are presently unknown, but are likely to decline over time? What are the costs for the development of a statistical methodology to transform the satellite sensing data into crop yields; and the development of a statistical system to process and dissemination satellite sensing data? What are the equivalent costs for direct data collections, and how they compare with one another?
- Reduction in provider load – How much reduction in provider load would result if direct data collection is replaced by satellite sensing data? How important is it to have this reduction, based on existing provider experience, and prevailing Government policy on reduction of regulatory “red tape”? What is the current degree of cooperation from the farmers and how likely is this degree change for the better, or worse, in the future?
- Sustainability of the statistical outputs – Is the data source available to official statisticians for the regular production of official statistics? How likely will the source be discontinued in the future?
- Accuracy, relevance, consistency, interpretability, and timeliness – How does the new source of data compare with the current source, against the criteria outlined in Data Quality Frameworks (ABS, 2010a; Brackstone, 1999; OECD, 2011). Whilst satellite sensing provide accurate measurements of “reflectance” – measures of the amount of light reflected on objects - there are missing data from Landsat 7 missions (see below), and from cloud covers. Are these issues bigger, or smaller, than missing data issues from direct data collections? In addition, transforming reflectance into crop production statistics require scientific or statistical modelling,

an endeavour not commonly adopted by NSOs and may raise interpretability issues. As satellite sensing data are available once every fortnight, they clearly have a distinct advantage over annual or sub-annual direct data collections in terms of the frequency in the availability of crop yield statistics.

6. VALIDITY OF STATISTICAL INFERENCE

Data sets derived from Big Data sources are not necessarily random samples of the target population. The design-based statistical inferences adopted by most NSOs for estimating finite population parameters such as population means, totals, and quantiles rely on random samples, i.e. the selection mechanism does not depend on the values of the units not selected in the sample (Särndal, Swensson and Wretman, 1977; Kish, 1965); or statistical models to adjust or address the selection bias from non-random samples (Puza and O'Neill, 2006).

As an example, social media services (such as Twitter) are a rich data source for the measurement of public opinion. However, there is little verifiable information about the users of these services, and it is difficult to determine whether the user profiles are "representative" of the population in general. In fact, it is to be expected that some population subgroups will be under-represented in any sample of social media data, due to the differential adoption rate of new technologies. Where the non (self) selection in the social media is dependent on these people's public opinion, estimates of population opinion from such sources, without proper modelling and adjustment, are subject to bias (Smith, 1983).

In general, being custodians of large number and variety of statistical benchmarks, NSOs are uniquely positioned to assess the representativeness of the underlying population of Big Data. In some cases the Big Data might need to be supplemented with survey data to get coverage of un-represented segments of the population. In other cases it may be useful to publish statistics that describe sub-populations. A related issue is that the statistical analysis of large, complex heterogeneous datasets will inevitably yield significantly more spurious model-dependent correlations than would be expected from traditional data sources. This can actually accentuate any modelling bias by reinforcing the selection of the wrong variables, algorithms and metrics of fitness.

As an example, Google Flu Trends – which uses the number of online searches as a measure of the prevalence of flu in the general population – mistakenly estimated that peak flu levels reached 11% of the U.S. public in the 2012 flu season. This was almost double the official estimate of 6% published by public health officials. Google Trends explained the over-estimation by "… heightened media coverage on the severity of the flu season resulted in an extended period in which users were searching for terms we've identified as correlated with flu levels" (Google Trends, 2013). This highlights the importance of assessing under what conditions, and for what applications, the use of Big Data require adjustment or no adjustment, in order to provide statistical estimates that are of the same level of quality as the official statistics regularly published by the NSOs.

6.1 A Theory for Big Data statistical inference

Couper (2013) outlined some significant issues the analyst has to consider when making inferences using Big Data, including coverage bias, selection (representational) bias, measurement bias and response bias. The Section attempts to provide a framework for considering such issues. It is useful to conceptualise the following elements of an inference framework for Big Data:

1. Target population, U (of size u) – the population of interest to the NSOs on which statistical inferences are to be made. In the Twitter example, this may be the population of Australia aged 15 years and above. In the remote sensing example, this may be the agricultural land parcels of Australia;

2. Big Data population, U_B (of size b) – the actual population included in the Big Data. In the Twitter example, this will be the registered Twitter users. For the remote sensing example, this can be the land parcels of Australia. If the coverage of U_B is not the same as the coverage of U, inference based on U_B will suffer from coverage bias. For the rest of this paper, we assume that the coverage of U_B is a subset of U, and conceptualise U_B as a sample (random or otherwise) from U – see point 6 below – with the coverage bias to be addressed through statistical modelling of the missing data process "R" – see point 7 below;

3. Vector of measurements of interest to the NSO, $\mathbf{M}_U$. This could be consumer confidence or crop yields;

4. Vector of proxy measurements available from Big Data, $\mathbf{Z}_B$. This provides the proxy variables, or covariates, to be used to predict $\mathbf{M}_U$. From points 1 and 2 above, we can consider $\mathbf{Z}_B$ as a sample (random or otherwise) of measurements from U to predict $\mathbf{M}_U$. In the Twitter example, $\mathbf{Z}_B$ could be the sentiment data to predict consumer confidence, $\mathbf{M}_U$. In remote sensing example, $\mathbf{Z}_B$ comprise reflectance measurements from selected wavebands captured by remote sensing missions, for discrete pixels of sizes ranging between 10 m^2 to 1 km^2, to predict the annual production of certain types of crops, $\mathbf{M}_U$.

5. A transformation (or measurement) process, "T", is generally required to transform the data, $\mathbf{Z}_B$, to the measurements of interest, $\mathbf{M}_U$. In the remote sensing example, this may be transforming the observed reflectance in selected wavebands captured by the remote sensing mission into the crop types – see Section 6.2 for an example of the "T" process. This is generally a complex scientific or statistical modelling process requiring detailed understanding of the reflectance characteristics of the different ground cover types, which in turn are dependent on the selective spectral absorption characteristics associated with their biophysical and biochemical compositions (Richards, 2013, p. 12).

6. A sampling process – random or otherwise – “I” is used to conceptualise the selection of $\mathrm{U_B}$ from U. In many Big Data examples, “I” is unknown, and requires in-depth contextual knowledge to develop proper statistical models to represent it, if at all. Depending on the type of the Big Data source, this is not generally a straight forward process. However, with remote sensing data through Landsat satellite series (Landsat, 2013), one has the fortunate situation that the coverage of U and $\mathrm{U_B}$ are identical, making the “I” process superfluous in this case;

7. A censoring (missing data) process, “R”, which renders parts of the vector, $\mathbf{Z}_\mathbf{B}$, not available. Where the coverage of $\mathrm{U_B}$ is not the same as U, one could conceptualise a “R” process in play rendering observations in the target population, U, missing. Another instance of missing data will be incomplete observations from $\mathbf{Z}_\mathbf{B}$. In the remote sensing example, missing data could be due to bad weather, or something more systemic – see Section 6.2 below. For this reason, whilst the “I” process can be subsumed in the “R” process, for the purpose of this paper, we conceptualise them as two separate processes. We will use the notation, $\mathbf{Z}_{\mathbf{Bo}}$, to represent the observed covariates or proxy variables from Big Data.

For finite population inferences, we are interested in predicting $g(\mathrm{M_U})$, where $g(\cdot)$ denotes a linear or non-linear function. The data that we have for the inference is $\mathrm{Z_{Bo}}$ (which “survives” the selection and censoring processes I and R).

We denote by $f(\cdot)$ the probability density function (pdf). We assume the pdf, $f(\mathrm{Y_U};\theta)$, indexed by the unknown parameter, θ, with known prior distribution $f(\theta)$, is known. Predicting $\mathbf{M_U}$ from $\mathbf{Z_U}$ will generally be a scientific process – for example, converting remote sensing data, $\mathbf{Z_U}$, into crop yield data, $\mathbf{M_U}$. For the purpose of this paper, we also make the assumption that the pdf $f(\mathrm{M_U}|\mathrm{Z_U};\varphi)$ is known, as is the prior $f(\varphi)$ of the parameter φ. It is further assumed that the parameters, φ and θ, are distinct (Rubin, 1976).

Following Rubin (1976), Little (1982), Little (1983) and Smith (1983), the task of the statistician, using a Bayesian inference framework, is to calculate the posterior distribution $f(\mathrm{M_U}|\mathrm{Z_{Bo}},\mathrm{I},\mathrm{R})$, or simply $[\mathrm{M_U}|\mathrm{Z_{Bo}},\mathrm{I},\mathrm{R}]$ to simplify the notation. Writing $\mathrm{Z}'_\mathrm{U}=(\mathrm{Z}'_\mathrm{Bo},\mathrm{Z}'_\mathrm{C})$ to split up the $\mathrm{Z_U}$ variables of the target population into those from the Big Data and the remainder, we have

$$
\begin{aligned}
&[\mathrm{M_U}|\mathrm{Z_{Bo}},\mathrm{I},\mathrm{R}] \;\propto\; \iiint [\mathrm{M_U},\mathrm{Z_{Bo}},\mathrm{Z_C},\mathrm{I},\mathrm{R},\theta,\varphi]\,d\theta\,d\varphi\,d\mathrm{Z_C}\\
&= \iiint [\mathrm{R}|\mathrm{M_U},\mathrm{Z_{Bo}},\mathrm{Z_C},\mathrm{I},\theta,\varphi]\,[\mathrm{I}|\mathrm{M_U},\mathrm{Z_{Bo}},\mathrm{Z_C},\theta,\varphi]\,[\mathrm{M_U},\mathrm{Z_{Bo}},\mathrm{Z_C},\theta,\varphi]\,d\theta\,d\varphi\,d\mathrm{Z_C}\\
&\propto \iiint [\mathrm{M_U},\mathrm{Z_{Bo}},\mathrm{Z_C},\theta,\varphi]\,d\theta\,d\varphi\,d\mathrm{Z_C} \qquad (1)\\
&= [\mathrm{M_U},\mathrm{Z_{Bo}}]\\
&\propto [\mathrm{M_U}|\mathrm{Z_{Bo}}] \qquad (2)
\end{aligned}
$$

provided that the following ignorability conditions (I) for sampling and censoring are satisfied:

$$\left[R\,|M_U, Z_{Bo}, Z_C, I, \theta, \varphi\right] = \left[R\,|M_U, Z_{Bo}, I\right]$$

and

$$\left[I\,|M_U, Z_{Bo}, Z_C, \theta, \varphi\right] = \left[I\,|Z_{Bo}\right].$$

In other words, subject to the fulfilment of these conditions, the scientific process to translate the Big Data observations Z_{Bo} into the measurements of interest M_U can be performed by disregarding the sampling and censoring processes.

Now

$$\begin{aligned}\left[M_U\,|Z_{Bo}\right] &\propto \iiint \left[M_U, Z_{Bo}, Z_C, \theta, \varphi\right] d\theta\, d\varphi\, dZ_C \\ &= \iiint \left[M_U, Z_U, \theta, \varphi\right] d\theta\, d\varphi\, dZ_C \\ &= \iiint \left[M_U|Z_U, \varphi\right]\left[\varphi\right]\left[Z_U|\theta\right]\left[\theta\right] d\theta\, d\varphi\, dZ_C \\ &= \int E_\varphi\left(M_U|Z_U, \varphi\right)\; E_\theta\left(Z_U|\theta\right) dZ_C \qquad (3)\end{aligned}$$

where $E_\varphi(\cdot)$ and $E_\theta(\cdot)$ denote the expectation with respect to $f(\varphi)$ and $f(\theta)$ respectively.

For analytic inferences, the interest will be on estimating the parameter φ of the pdf, $f\left(M_U \mid Z_U\,;\varphi\right)$. Now similar to the derivation of (1),

$$\begin{aligned}&\left[\varphi\,|Z_{Bo}, I, R\right] \propto \left[\varphi, Z_{Bo}, I, R\right] \\ &\propto \iint \left[Z_{Bo}, Z_C, I, R, \theta, \varphi\right] d\theta\, dZ_C \\ &= \iint \left[R\,|Z_{Bo}, Z_C, I, \theta, \varphi\right]\left[I\,|Z_{Bo}, Z_C, \theta, \varphi\right]\left[Z_{Bo}, Z_C, \theta, \varphi\right] d\theta\, dZ_C \\ &\propto \iint \left[Z_{Bo}, Z_C, \theta, \varphi\right] d\theta\, dZ_C \qquad (4) \\ &\propto \left[\varphi\,|Z_{Bo}\right] \qquad (5)\end{aligned}$$

provided that the following ignorability conditions (II) are satisfied:

$$\left[R\,|Z_{Bo}, Z_C, I, \theta, \varphi\right] = \left[R\,|Z_{Bo}, I\right]$$

and

$$\left[I\,|Z_{Bo}, Z_C, \theta, \varphi\right] = \left[I\,|Z_{Bo}\right].$$

Likewise,

$$\begin{aligned}\left[\varphi|Z_{Bo}\right] &\propto \iiint \left[M_U, Z_U, \theta, \varphi\right] d\theta\, dZ_C\, dM_U \\ &= \iiint \left[M_U|Z_U, \varphi\right]\left[\varphi\right]\left[Z_U|\theta\right]\left[\theta\right] d\theta\, dZ_C\, dM_U \\ &= \left[\varphi\right]\int \left[M_U|Z_U, \varphi\right] E_\theta\left(Z_U|\theta\right) dZ_C\, dM_U\ . \end{aligned} \qquad (6)$$

The ignorability conditions (I) and (II) are also known as Missing At Random conditions (Rubin, 1976). The inference framework outlined in this paper is similar to the one described in Wikle *et al.* (1998) for predicting temperature data, but is adapted to official statistics, as well as extended to address missing data from the "R" process, and the sampling process, "I".

Whilst the ignorability conditions may not be fulfilled, if NSOs have access to other variables, X_B, such that $\left[I, R \mid M_U, X_B, Z_{Bo}, Z_C, \theta, \varphi\right] = \left[I, R \mid X_B, Z_{Bo}\right]$ is fulfilled, then $\left[M_U \mid X_B, Z_{Bo}, I, R\right] = \left[M_U \mid X_B, Z_{Bo}\right]$. (Tam, 2014)

Often, NSOs will have some information on M_U (say, M_{so}) from sample surveys or administrative data sources, where "so" denotes the observed value from such sources. Conceptually, we assume that a sampling process, $\mathcal{J}$, and a censoring process, $\mathcal{R}$, are at play, where the $\mathcal{J}$ and $\mathcal{R}$ are considered to be independent of the I and R processes.

In this case, inference on M_U should be based on $\left[M_U \mid M_{so}, Z_{Bo}, \mathcal{J}, \mathcal{R}, I, R\right]$.

Now,

$$\begin{aligned}\left[M_U \mid M_{so}, Z_{Bo}, \mathcal{J}, \mathcal{R}, I, R\right] &= \frac{\left[\mathcal{J}, \mathcal{R} \mid M_U, M_{so}, Z_{Bo}, I, R\right]\left[M_U, M_{so}, Z_{Bo}, I, R\right]}{\left[\mathcal{J}, \mathcal{R} \mid M_{so}, Z_{Bo}, I, R\right]\left[M_{so}, Z_{Bo}, I, R\right]} \\ &= \left[M_U \mid M_{so}, Z_{Bo}, I, R\right]\end{aligned}$$

provided that

$$\left[\mathcal{J}, \mathcal{R} \mid M_U, M_{so}, Z_{Bo}, I, R\right] = \left[\mathcal{J}, \mathcal{R} \mid M_{so}, Z_{Bo}, I, R\right]$$

i.e. the $\mathcal{J}, \mathcal{R}$ do not depend on the unobserved and unsampled or unobserved values of M_U – ignorability condition (III). This is a weaker condition than Ignorability Condition I, since design-based sampling conditions are generally used in NSOs, and the incidence of non-response is low for mandatory surveys. This is certainly the case in the ABS.

Example 1

(Puza, 2013) Assume for the simple case where $\mathrm{M_U} = \mathrm{M_B}$, $\mathrm{Z_{Bo}} = \mathrm{Z_B}$, and the sampling and censoring processes – $\mathcal{J}, \mathrm{I}, \mathcal{R}$ and R – are ignorable. Under the models:

$$\mathrm{M_B} \,\big|\, \mathrm{Z_B}, \beta \;\sim\; \mathrm{N}\left(\mathrm{Z_B}\beta, \Sigma\right)$$
$$\beta \,\big|\, \beta_0, \Omega \;\sim\; \mathrm{N}\left(\beta_0, \Omega\right)$$

we have $$\left[\mathrm{M_U} \,\big|\, \mathrm{M_{so}}, \mathrm{Z_B}\right] \;=\; \left[\mathrm{M_C} \,\big|\, \mathrm{M_{so}}, \mathrm{Z_B}\right] \;=\; \mathrm{N}\left(\mu, \psi\right)$$

and $$\left[\beta \,\big|\, \mathrm{M_{so}}, \mathrm{Z_B}\right] \;=\; \mathrm{N}\left(\hat{\beta}, \mathrm{D}\right)$$

where

$$\mu \;=\; \mathrm{Z_r}\,\hat{\beta} + \Sigma_{\mathrm{rso}}\Sigma_{\mathrm{soso}}^{-1}\left(\mathrm{M_{so}} - \mathrm{Z_{so}}\,\hat{\beta}\right),$$

$$\psi \;=\; \Sigma_{\mathrm{rr}} - \Sigma_{\mathrm{rso}}\Sigma_{\mathrm{soso}}^{-1}\Sigma_{\mathrm{rso}} + \left(\mathrm{Z_r} - \Sigma_{\mathrm{rso}}\Sigma_{\mathrm{soso}}^{-1}\mathrm{Z_{so}}\right)\mathrm{D}\left(\mathrm{Z_r} - \Sigma_{\mathrm{rso}}\Sigma_{\mathrm{soso}}^{-1}\mathrm{Z_{so}}\right)',$$

$$\hat{\beta} \;=\; \mathrm{D}\left(\Omega^{-1}\beta_0 + \mathrm{Z'_{so}}\Sigma_{\mathrm{soso}}^{-1}\mathrm{M_{so}}\right),$$

$$\mathrm{D} \;=\; \left(\Omega^{-1} + \mathrm{Z'_{so}}\Sigma_{\mathrm{soso}}^{-1}\mathrm{Z_{so}}\right)^{-1},$$

$$\mathrm{M'_B} \;=\; \left(\mathrm{M'_{so}}, \mathrm{M'_C}\right),$$

$$\mathrm{Z'_B} \;=\; \left(\mathrm{Z'_{so}}, \mathrm{Z'_r}\right),$$

$$\Sigma \;=\; \begin{pmatrix} \Sigma_{\mathrm{soso}} & \Sigma_{\mathrm{sor}} \\ \Sigma_{\mathrm{sor}} & \Sigma_{\mathrm{rr}} \end{pmatrix}.$$

The above theory shows that for proper finite population and analytic inferences, equations (1) and (4) should be used. In general, in most Big Data applications the specification of the censoring model (e.g. how censoring is dependent on the unobserved data in the target population but not in the Big Data population) and the sampling model (e.g. how sampling is dependent on unobserved measurements or proxy measurement) can be subjective and difficult to specify, although we note that there is a vast body of statistical literature to address non-ignorable situations (Puza and O'Neil, 2006; Heckman, 1979; Little, 1982; Little, 1983; Little and Rubin, 2002; Madow, Oklin and Rubin, 1983; Smith, 1983 and Wu, 2010 are just some examples). The challenge for the official statistician is to find and use models that meet the integrity requirements of official statistics. Where information available to the official statistician suggests that the ignorability conditions are fulfilled, then analyses of Big Data can proceed as if it is a random sample from the target population.

6.2 Analysis of satellite sensing data

For remote sensing data, let $\mathbf{M}_U$ represent, for simplicity, the vector of binary variables for all the pixels in U, with a value of 1 assigned for the pixel that contains a certain crop of interest to the NSO, and 0 otherwise. Let $\mathbf{Z}_U$ be a matrix comprising row vectors of remote sensing data (consisting of reflectance measurements from the satellite on-board sensors) and a column vector of "ones".

As the full data $\mathbf{Z}_U$ is available from Landsat (Landsat, 2013), $\mathbf{Z}_B = \mathbf{Z}_U$. That is, there is no sampling involved so the first requirement of ignorability conditions (I) is satisfied. However, when there is missing data, then the second requirement of ignorability conditions (I) needs to be checked. Where missing data is due to random bad weather, it may be safe to assume that the missing data is not associated with the reflectance measurements, and, if so, we may treat the resultant dataset as a random sample. In the case where missing data is due to systemic effects – such as the problems that occurred in May 2003 which caused approximately 22% missingness of the Landsat 7 imagery data that had to be replaced by other data – an assessment is required on whether it is acceptable to assume the observed data set comprises a random sample.

An interest will be to use the reflectance measurements to predict the total yield of the particular crop in question, which requires the specification of a "T" process. A review of the remote sensed information models and crop models for this "T" process is provided in Delécolle *et al.* (1992). Alternatively, a statistical model like a logistic regression model may be used, provided that the NSO has information on $\mathbf{M}_{so}$, available from agricultural censuses or surveys. For the simple case where $\mathbf{M}_U = \mathbf{M}_B$, $\mathbf{Z}_{Bo} = \mathbf{Z}_B$, the sampling and censoring/response processes – $\mathcal{J}, \mathrm{I}, \mathcal{R}$ and R – are ignorable, and letting "a" denote the size of a pixel, then the estimated yield of a certain type of crop will simply be $\mathbf{1}'_C \mathbf{M}_C \, a + \mathbf{1}'_{so} \mathbf{M}_{so} \, a$. The statistical task is to predict $\mathbf{M}_C$, given $\mathbf{M}_{so}$ and $\mathbf{Z}_B$.

So long as the above formulation does not take into account of previously available information, e.g. $\mathbf{Z}_B$ and $\mathbf{M}_{so}$ from earlier time points, the analysis is not optimal. A proper Bayesian analysis must take account of this information. To address this, we introduce the following notations and models.

To simplify notations, we shall drop the subscript "B" in the sequel. We shall also introduce a "t" subscript to denote time. Let

- $\mathbf{M}_t$ denote the $b \times 1$ vector $\mathbf{M}_{tB} = \{m_{t1}, \ldots, m_{tb}\}'$;
- m_{ti} be a Random Variable (to be defined in Example 2 below) for pixel i, where $i = 1, \ldots, b$;
- $\mathbf{Z}_{ti}$ be the $p \times 1$ vector of reflectance and an intercept for pixel i;

- $p = 8$;
- $\boldsymbol{\beta}_t$ be the $p \times 1$ vector of unknown coefficients;
- $\sigma(Z'_{ti}\boldsymbol{\beta}_t) = \left(1 + e^{-Z'_{ti}\boldsymbol{\beta}_t}\right)^{-1}$ be the logistic sigmoid for pixel i;
- $\sigma(Z'_t\boldsymbol{\beta}_t)$ be the $b \times 1$ vector $\{\sigma(Z'_{ti}\boldsymbol{\beta}_t), \ldots, \sigma(Z'_{tb}\boldsymbol{\beta}_t)\}'$;
- $M_{ts} = M_{tso}$ be the $n_t \times 1$ vector of "ones" or "zeros", where "one" is recorded if the pixel grows the crop of interest to the official statistician, and zero otherwise. This information is obtained by "ground truthing" and is also referred to as the training dataset in the Machine Learning literature;
- M_{tC} $(N_t \times 1)$ is defined by the equation $M'_t = \{M'_{tC}, M'_{ts}\}$, where $b = N_t + n_t$;
- Z_t be the $u \times p$ matrix of all reflectance and a column of ones available from satellite sensing data;
- Z_{ts} be the $n_t \times p$ matrix of reflectance and a column of ones corresponding to the training dataset;
- $M_s^{(t)} = \{M_{1s}, \ldots, M_{ts}\}$;
- $Z_s^{(t)} = \{Z_{1s}, \ldots, Z_{ts}\}$; and
- $D^{(t)} = \left\{M_s^{(t)}, Z_s^{(t)}\right\}$.

Example 2

Under the models:

$$m_{ti} \sim \text{independent Binomial Logistic } \left(\sigma(Z'_{ti}\boldsymbol{\beta}_t)\right)$$
$$\boldsymbol{\beta}_t = \boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t \,,\ \ \boldsymbol{\beta}_t \perp Z_t$$
$$\boldsymbol{\varepsilon}_t \sim \text{independent } N(0, \boldsymbol{\Omega}_t)\,,\ \ \boldsymbol{\varepsilon}_t \perp D^{(t)}$$

for $i = 1, \ldots, b$, and $\boldsymbol{\Omega}_t$ known for $t = 1, \ldots, \tau$, then

$$m_{ti} \,\Big|\, D^{(t)} \approx \text{ independent Binomial Logistic } \left(\sigma\left(Z'_{ti}\hat{\boldsymbol{\beta}}_{t|t}\right)\right)$$

for $i = 1, \ldots, b$ and $t = 1, \ldots, \tau$, where $\hat{\boldsymbol{\beta}}_{t|t}$ is the maximum likelihood estimator of the posterior distribution $\left[\boldsymbol{\beta}_t \,\Big|\, D^{(t)}\right]$, and $\boldsymbol{\Sigma}_{t|t-1}$ is the inverse of the negative of the Hessian of the posterior distribution evaluated at $\hat{\boldsymbol{\beta}}_{t|t}$, given by:

$$\hat{\boldsymbol{\beta}}_{t|t} \;=\; \hat{\boldsymbol{\beta}}_{t-1|t-1} + \Sigma_{t|t-1}^{-1}\left\{Z'_{ts}M_{ts} - Z'_{ts}\sigma\left(Z'_t\hat{\boldsymbol{\beta}}_{t|t}\right)\right\} \tag{7}$$

and

$$\boldsymbol{\Sigma}_{t|t-1} \;=\; \boldsymbol{\Sigma}_{t-1|t-1} + \boldsymbol{\Omega}_t \,. \tag{8}$$

In addition,

$$\left[\beta_t \middle| D^{(t)}\right] \approx N\left(\hat{\beta}_{t|t}, \Sigma_{t|t}\right)$$

where $$\Sigma_{t|t} = \left(Z'_{ts} W_{ts}\left(\beta_t\right) Z_{ts} + \Sigma^{-1}_{t|t-1}\right)^{-1}$$

and $\Sigma_{t|t}$ is evaluated at $\beta_t = \hat{\beta}_{t|t}$.

Finally, the posterior mean of $1'_{tC} M_{tC} a + 1'_{ts} M_{ts} a$ is

$$1'_{tC}\sigma\left(Z'_{ti}\hat{\beta}_{t|t}\right)a + 1'_{ts} M_{ts} a,$$

and the posterior variance is $$1'_{tC} W_{tC}\left(\hat{\beta}_{t|t}\right) 1_{tC} a^2,$$

where $W_{tC}\left(\hat{\beta}_{t|t}\right)$ is an $N_t \times N_t$ diagonal matrix, with diagonal elements,

$$\sigma\left(Z'_{ti}\hat{\beta}_{t|t}\right)\left(1-\sigma\left(Z'_{ti}\hat{\beta}_{t|t}\right)\right) \text{ for } i = 1,\ldots,N_t \text{ .}$$

To prove the results, we first let $\left[\beta_t \middle| D^{(t)}\right] = G\left(\beta_t\right)$, say. Using the Taylor series for vectors, it can be shown that

$$\left[\beta_t \middle| D^{(t)}\right] \approx N\left(\hat{\beta}_{t|t}, \Sigma_{t|t}\right)$$

where $$\hat{\beta}_{t|t} = \arg\max_{\beta_t} \ln G\left(\beta_t\right)$$

and $$\Sigma_{t|t} = \left\{-\nabla^2 \ln G\left(\beta_t\right)\right\}^{-1} \text{evaluated at } \hat{\beta}_{t|t},$$

where $\nabla^2 \ln G\left(\beta_t\right)$ denotes the second derivative of $\ln G\left(\beta_t\right)$ with respect to β_t.

From

$$\begin{aligned}
\left[M_t \middle| D^{(t)}\right] &= \left[M_{tc} \middle| D^{(t)}\right] \\
&= \int \left[M_{tc} \middle| D^{(t)}, \beta_t\right]\left[\beta_t \middle| D^{(t)}\right] d\beta_t \\
&= \int \prod_{i=1}^{N_t} \sigma\left(Z'_{ti}\beta_t\right)^{m_{ti}}\left(1-\sigma\left(Z'_{ti}\beta_t\right)^{1-m_{ti}}\right)\left[\beta_t \middle| D^{(t)}\right] d\beta_t \\
&\sim \prod_{i=1}^{N_t} \sigma\left(Z'_{ti}\beta_t\right)^{m_{ti}}\left(1-\sigma\left(Z'_{ti}\beta_t\right)^{1-m_{ti}}\right)
\end{aligned}$$

using the standard form for Laplace approximation for integrals (Tierney *et al.*, 1989), given that $\left[\beta_t \middle| D^{(t)}\right] \approx N\left(\hat{\beta}_{t|t}, \Sigma_{t|t}\right)$.

This proves the first part of the result of Example 2.

Now

$$\begin{aligned}\left[\beta_t \middle| D^{(t)}\right] &= G(\beta_t) \\ &\propto \left[M_{ts} \middle| Z_t, D^{(t-1)}, \beta_t\right]\left[\beta_t \middle| Z_t, D^{(t-1)}\right].\end{aligned}$$

From the model assumptions of β_t, it follows that

$$\begin{aligned}\left[\beta_t \middle| Z_t, D^{(t-1)}\right] &= \left[\beta_t \middle| D^{(t-1)}\right] \\ &= \int \left[\beta_t \middle| D^{(t-1)}, \beta_{t-1}\right]\left[\beta_{t-1} \middle| D^{(t-1)}\right] d\beta_{t-1} \\ &= \int \left[\beta_t \middle| \beta_{t-1}\right]\left[\beta_{t-1} \middle| D^{(t-1)}\right] d\beta_{t-1} \\ &\approx N\left(\hat{\beta}_{t|t-1}, \Sigma_{t|t-1}\right) \qquad (9)\end{aligned}$$

where

$$\begin{aligned}\hat{\beta}_{t|t-1} &= E\left[\beta_t \middle| Z_t, D^{(t-1)}\right] \\ &= E\left[\beta_{t-1} \middle| Z_t, D^{(t-1)}\right] + E\left[\varepsilon_t \middle| Z_t, D^{(t-1)}\right] \\ &= E\left[\beta_{t-1} \middle| Z_t, D^{(t-1)}\right] \\ &= \hat{\beta}_{t-1|t-1} \qquad (10)\end{aligned}$$

and

$$\begin{aligned}\Sigma_{t|t-1} &= V\left[\beta_{t-1} \middle| Z_t, D^{(t-1)}\right] + V\left[\varepsilon_t \middle| Z_t, D^{(t-1)}\right] \\ &= \Sigma_{t-1|t-1} + \Omega_t\ . \qquad (11)\end{aligned}$$

Now (7) follows by noting that

$$\begin{aligned}\nabla \ln G(\beta_t) &= \nabla \ln\left[M_{ts} \middle| Z_t, D^{(t-1)}, \beta_t\right] + \nabla \ln\left[\beta_t \middle| Z_t, D^{(t-1)}\right] \\ &= Z'_{ts}M_{ts} - Z'_{ts}\sigma\left(Z'_t\beta_t\right) - \Sigma^{-1}_{t|t-1}\left(\beta_t - \hat{\beta}_{t-1|t-1}\right) \qquad (12)\end{aligned}$$

with the well-known first term of (12) derived in maximum likelihood estimation for logistic regression models – see for example, Czepiel (2002), and the second term of (12) follows from the multivariate normality of β_t; and (8) follows from noting that

$$\begin{aligned}\Sigma^{-1}_{t|t} &= -\nabla^2 \ln\left[M_{ts} \middle| Z_t, D^{(t-1)}, \beta_t\right] - \nabla^2\left[\beta_t \middle| Z_t, D^{(t-1)}\right] \\ &= Z'_{ts}W_{ts}(\beta_t)Z_{ts} + \Sigma^{-1}_{t|t-1}\end{aligned}$$

where $W_{ts}(\beta_t)$ is an $n_t \times n_t$ diagonal matrix, with diagonal elements

$$\sigma(Z'_{ti}\beta_t)\{1-\sigma(Z'_{ti}\beta_t)\} \text{ , for } i = 1,\ldots,n_t \text{ .}$$

Comment 1. The restrictive assumption that Ω_t is known for all t can be relaxed by, for example, assuming

$$\Omega_t = \Sigma_{t-1|t-1} / \lambda_t$$

for scalars λ_t such that λ_t maximises $\int G(\beta_t)\,d\beta_t$ (McCormick *et al.*, 2012).

Comment 2. Whilst (7) is of interest in its own right, it is not useful to calculate $\hat{\beta}_{t|t}$, given non-linearity of the equation. Instead, a Newton-Raphson method to calculate $\hat{\beta}_{t|t}$ is proposed, as follows:

$$\hat{\beta}_{t|t}^{(n+1)} = \hat{\beta}_{t|t}^{(n)} + \Sigma_{t|t}^{(n)} \nabla \ln G\left(\hat{\beta}_{t|t}^{(n)}\right)$$

where we set the starting value of $\beta_{t|t}^{(1)}$ to be $\hat{\beta}_{t-1|t-1}$.

Comment 3. The results in this Section can be readily extended to multinominal logistic regression models (see Czepiel, 2002).

6.3 Analysis of continuous Big Data

In Big Data scenarios where M_U can be modelled as a continuous variable, and where the data, M_{ts}, are regularly observed, Tam (1987) extended the model $\beta \mid \beta_0, \Omega \sim N(\beta_0, \Omega)$ in Example 1 to $\beta_t - \beta_{t-1} \mid \Omega \sim N(0, \Omega_t)$, where t denotes time. Noting the above models form a State-Space model, and Kalman (1960) provided the best linear unbiased predictor for $\hat{\beta}_t$ given $D^{(t)}$. In this Section, we extend Puza (2013) to provide the predictive distribution, $\left[M_{tC} \mid D^{(t)}\right]$.

Example 3

Using the notation of Section 6.2, assume $M_{tU} = M_{tB}$, $Z_{tBo} = Z_{tB}$ and the sampling and censoring processes – $\mathcal{J}, I, \mathcal{R}$ and R – are ignorable. Under the models:

$$M_{tB} \mid Z_{tB}, \beta_t \sim N(Z_{tB}\beta_t, \Sigma_t)$$
$$\beta_t = \beta_{t-1} + \varepsilon_t \text{ , } \beta_t \perp Z_t$$
$$\varepsilon_t \sim \text{independent } N(0, \Omega_t) \text{, } \varepsilon_t \perp D^{(t)}$$

for $i = 1,\ldots,b$ and $t = 1,\ldots\tau$, and where Ω_t is assumed to be known for every t, to simplify the illustration, then

$$\left[M_{tC} \mid D^{(t)}\right] = N(\mu_t, \Psi_t)$$

where

$$
\begin{aligned}
\boldsymbol{\mu}_{t} &= \mathbf{Z}_{r}\hat{\boldsymbol{\beta}}_{t|t} + \boldsymbol{\Sigma}_{trs}\boldsymbol{\Sigma}_{tss}^{-1}\left(\mathbf{M}_{ts} - \mathbf{Z}_{ts}\hat{\boldsymbol{\beta}}_{t|t}\right), \\
\boldsymbol{\Psi}_{t} &= \boldsymbol{\Sigma}_{trr} - \boldsymbol{\Sigma}_{trs}\boldsymbol{\Sigma}_{tss}^{-1}\boldsymbol{\Sigma}_{tsr} + \left(\mathbf{Z}_{tr} - \boldsymbol{\Sigma}_{trs}\boldsymbol{\Sigma}_{tss}^{-1}\mathbf{Z}_{ts}\right)\boldsymbol{\Sigma}_{t|t}\left(\mathbf{Z}_{tr} - \boldsymbol{\Sigma}_{trs}\boldsymbol{\Sigma}_{tss}^{-1}\mathbf{Z}_{ts}\right)', \\
\hat{\boldsymbol{\beta}}_{t|t} &= \boldsymbol{\Sigma}_{t|t}\left(\boldsymbol{\Sigma}_{t|t-1}^{-1}\hat{\boldsymbol{\beta}}_{t-1|t-1} + \mathbf{Z}'_{ts}\boldsymbol{\Sigma}_{tss}^{-1}\mathbf{M}_{ts}\right), \qquad (13) \\
\boldsymbol{\Sigma}_{t|t} &= \left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{Z}'_{ts}\boldsymbol{\Sigma}_{tss}^{-1}\mathbf{Z}_{ts}\right)^{-1}.
\end{aligned}
$$

In addition, the posterior mean of $\mathbf{1}'_{tC}\mathbf{M}_{tC} + \mathbf{1}'_{ts}\mathbf{M}_{ts}$ is $\mathbf{1}'_{tC}\boldsymbol{\mu}_{t} + \mathbf{1}'_{s}\mathbf{M}_{ts}$, and the posterior variance is $\mathbf{1}'_{tC}\boldsymbol{\Psi}_{t}\mathbf{1}_{tC}$.

Proof of these results follows from observing that

- $\boldsymbol{\beta}_{t} \,\middle|\, \hat{\boldsymbol{\beta}}_{t-1|t-1}, \boldsymbol{\Sigma}_{t|t-1}$ is $N\left(\hat{\boldsymbol{\beta}}_{t-1|t-1}, \boldsymbol{\Sigma}_{t|t-1}\right)$ from (9), (10) and (11), and by replacing $\boldsymbol{\beta}_{0}$ by $\hat{\boldsymbol{\beta}}_{t-1|t-1}$ and $\boldsymbol{\Omega}$ by $\boldsymbol{\Sigma}_{t|t-1}$ in the results of Example 1; and likewise,
- $\hat{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}_{t|t}$ and $\mathbf{D}_{t} = \boldsymbol{\Sigma}_{t|t}$.

Now rewriting

$$
\begin{aligned}
&\boldsymbol{\Sigma}_{t|t}\left(\boldsymbol{\Sigma}_{t|t-1}^{-1}\hat{\boldsymbol{\beta}}_{t-1|t-1} + \mathbf{Z}'_{ts}\boldsymbol{\Sigma}_{tss}^{-1}\mathbf{M}_{ts}\right) \\
&= \boldsymbol{\Sigma}_{t|t}\left\{\left(\boldsymbol{\Sigma}_{t|t-1}^{-1} + \mathbf{Z}'_{ts}\boldsymbol{\Sigma}_{tss}^{-1}\mathbf{Z}_{ts}\right)\hat{\boldsymbol{\beta}}_{t-1|t-1} + \mathbf{Z}'_{ts}\boldsymbol{\Sigma}_{tss}^{-1}\mathbf{M}_{ts} - \mathbf{Z}'_{ts}\boldsymbol{\Sigma}_{tss}^{-1}\mathbf{Z}_{ts}\hat{\boldsymbol{\beta}}_{t-1|t-1}\right\},
\end{aligned}
$$

(13) can be rewritten in the more familiar form:

$$
\hat{\boldsymbol{\beta}}_{t|t} = \hat{\boldsymbol{\beta}}_{t-1|t-1} + \boldsymbol{\Sigma}_{t|t}\mathbf{Z}'_{ts}\boldsymbol{\Sigma}_{tss}^{-1}\left(\mathbf{M}_{ts} - \mathbf{Z}_{ts}\hat{\boldsymbol{\beta}}_{t-1|t-1}\right).
$$

The results extended those derived in Tam (1987) and Tam (1988, Chapter 5) which treated $\boldsymbol{\Sigma}_{t}$ as a diagonal matrix.

7. PRIVACY AND PUBLIC TRUST

The privacy landscape is fundamentally changed by the emergence of Big Data. There is an obvious contention between the systematic exploitation of Big Data sources for better decision-making across government, and the acknowledged need to establish and maintain public trust in the use of personal information by government agencies. NSOs' operations are governed by, and their authority to undertake collections enshrined in, statistical legislation. This sets the ground rules for how such data sets can be acquired, combined, protected, shared, exposed, analysed and retained. The legislation and associated policy framework is designed to promote trust and privacy and Big Data sources will further test our decision-making in adherence to the framework.

A significant unresolved issue is the threat of disclosure through data accumulation. Every individual is a unique mosaic of publicly visible characteristics and private information. In a data rich world, distinct pieces of data that may not pose a privacy risk when released independently are likely to reveal personal information when they are combined – a situation referred to as the "mosaic effect". The use of Big Data greatly amplifies the mosaic effect because large rich data sets typically contain many visible characteristics, and so individually or in composition may enable spontaneous recognition of individuals and the consequential disclosure of their private information. This will be a significant issue when disseminating microdata sets from Big Data sources.

8. DATA OWNERSHIP AND ACCESS

Data ownership and access is a key issue for NSOs and one where there is a generally lack of legislation and a supporting framework. The challenge is to unlock public good from privately collected data whilst protecting the commercial interests of the data custodians.

In many cases, commercial value is placed on primary and derived non-government data sets by their owners, since either the provision of such data is the basis of their business, or its possession is a significant element of competitive advantage. This raises the issue of how the NSO might acquire commercially valuable or sensitive data for statistical production, particularly if the statistics compete directly with information products created by the data owner or they compromise its market position. This issue is made more complex by the fact that there may be several parties with some form of commercial right in relation to a data set, either through ownership, possession or licensing arrangements.

Much Web content is also unstructured and ungoverned – the metadata describing its usage and provenance (origin, derivation, history, custody, and context) are either incomplete or incongruous. Indeed, the long-term reliability of Big Data sources may be an issue for ongoing statistical production. Reputable statistics for policy making and service evaluation are generally required for extended periods of time, often many years. However, large data sets from dynamic networks are volatile – the data sources may change in character or disappear over time. This transience of data streams and sources undermines the reliability of statistical production and publication of meaningful time series.

9. COMPUTATIONAL EFFICACY

The exploitation of Big Data will have a significant impact on the ICT resource demands of data acquisition, storage, processing, integration, and analysis. Existing computational models for the most common statistical problems in the typical NSO scale very poorly for the number, diversity and volatility of data elements, attributes and linkages associated with Big Data sources.

In particular, traditional relational database approaches are not sufficiently flexible for handling dynamic multiply-structured data sets in a computationally efficient way, and the execution of complex statistical algorithms at the scale of Big Data problems is likely to exceed the memory and processor resources of existing platforms. For example, probabilistic data linking under the Fellegi-Sunter model (Fellegi and Sunter, 1969) is generally treated as a constrained Maximum Likelihood problem using simplex-based algorithms. The complexity of this problem is at least $O(N^3)$, which cannot be solved with existing computing resources when the size of the data set N is at the scale of Big Data.

One possible approach is to outsource the analytics to the data owner. Statistics New Zealand is looking to do this with scanner data, as the data owner has the necessary computing infrastructure and performing the analysis where the data is stored is cheaper and easier. An added and important benefit of this approach is that the data owner does not need to share the underlying data, which may be very sensitive. A joint effort by methodologists and technologists is needed to develop techniques for reducing data volume and complexity while preserving statistical validity, and for improving algorithmic tractability and efficiency. This will involve explicitly recasting existing problems into a form that is better suited for distributed computing approaches, making greater use of approximate techniques, and favouring heuristic predictive models in the appropriate circumstances.

10. TECHNOLOGY INFRASTRUCTURE

Big Data technology has emerged from the extreme scale of Internet processing and progressively been applied to a growing range of business domains in the last decade. Industry supported open source technology developments have rapidly matured to the point where 'enterprise class' processing – in conjunction with traditional processing technologies – provides a stronger integrated set of technology options. Stand-alone and 'point' Big Data solutions are diminishing as they are integrated into wider solution architectures. Most established technology suppliers now include Big Data technology as part of their product portfolio. Big Data infrastructure and tools are evolving and there will continue to be proprietary and point solutions.

Big Data processing also requires new types of data representation (semantic data, graph database), inference (AI-based analytical techniques in conjunction with robust statistical analysis), visualisation (for complex network relationships), analytical languages (such as R and SAS), and the use of scale-out commodity hardware. A number of these technologies have value when applied to 'traditional' processing and analysis.

11. ABS INITIATIVES ON BIG DATA

As Australia's centralised national statistical agency, the ABS provides official statistics on a wide range of social, demographic, economic and environmental issues to encourage informed decision making, research and discussion within governments and the community. The principal legislation determining the functions and responsibilities of the ABS are the *Australian Bureau of Statistics Act 1975* and the *Census and Statistics Act 1905*.

A number of initiatives are being progressed to build future capability in the exploitation of Big Data sources and to position the ABS nationally and internationally as a leading agency in advanced data analytics.

11.1 A Big Data strategy

To position the ABS to harness Big Data opportunities, a Big Data Strategy paper (ABS, 2014) has been developed and approved by ABS senior management. The objectives of the Strategy are to build an integrated multifaceted capability for systematically exploiting the potential value of Big Data for official statistics.

This capability comprises:

- A skilled workforce able to interpret information needs and communicate the insights gleaned from rich data;
- Advanced methods, tools and infrastructure to represent, store, manipulate, integrate and analyse large, complex data sets;
- A diverse pool of government, private and open data sources available for statistical purposes;
- Safe and appropriate public access to microdata sets and statistical solutions derived from an array of data sources; and
- Strong multidisciplinary partnerships across government, industry, academia and the statistical community.

11.2 Big Data Flagship Project

The ABS Big Data Flagship Project – an initiative led by ABS methodologists – is intended to coordinate research and development (R&D) effort that will build a sound methodological foundation for the mainstream use of Big Data in statistical production and analysis. The desired outcomes of the project are to:

- Promote a greater understanding of Big Data concepts, opportunities, practicalities and challenges within the ABS;

- Encourage methodological rigour in the use of different sources of Big Data for statistical production;
- Build a seminal capability in exploring, combining, visualising and analysing large, complex and volatile data sets;
- Cultivate strong links to networks of Big Data experts in government, industry, academia, and the international statistical community; and
- Enhance national and international standing for the ABS in Big Data inference.

The project has scheduled the following work packages:

- Environmental Scanning and Opportunity Analysis – survey the operational environment for Big Data sources of potential use in statistical production, and to identify business problems and 'pain-points' that can be addressed through non-traditional data sources and analytical methods;
- Remote Sensing for Agricultural Statistics – investigate the use of satellite sensor data for the production of agricultural statistics such as land use, crop type and crop yield;
- Mobile Device Location Data for Population Mobility – investigate the use of mobile device location-based services and/or global positioning for measuring population mobility;
- Predictive Modelling of Unemployment – investigate the application of machine learning to the construction of predictive small-area models of unemployment from linked survey and administrative data;
- Visualisation for Exploratory Data Analysis – investigate advanced visualisation techniques for the exploratory analysis of complex multidimensional data sets;
- Analysis of Multiple Connections in Linked Data – investigate Linked Data techniques for analysing multiply connected data entities at different levels of granularity;
- Predictive Modelling of Survey Non-Response – investigate the application of machine learning to the construction of predictive small-domain models of non-response behaviour using para data from past surveys; and
- Automated Content Analysis of Complex Administrative Data – investigate techniques for the automated extraction and resolution of concepts, entities and facts from multi-structured content in administrative data sets.

11.3 Participation in the Australian Public Service (APS) data analytics initiatives

The ABS is a member of the Leadership Group of the APS Data Analytics Centre of Excellence (APS DACoE), which was formed in late 2013, to build collaborative capability across Government in the use of advanced data analytics by:

- sharing technical and business knowledge, tools and techniques, skills development and standards for operating such as protocols for privacy and information management practices;
- exploring and identifying opportunities to add business value through the use of analytics, considering: developments in information and knowledge management practices; industry developments in analytics technology, infrastructure and software; accreditation and professional development of analytics professionals for public-sector employment; and
- identifying and providing advice to the Chief Information Officers Committee on common issues and concerns affecting the analytics capability; barriers to the effective use of Big Data; Big Data pilot projects; other actions as outlined in the APS Big Data Strategy.

The DACoE has developed a best practice guide for Big Data/Big Analytics, which provides a whole-of-Government strategy on the use and implementation of Big Data amongst Australian Government agencies. It is currently compiling an inventory of business problems across government and the analytical methods and data sets that are being employed to solve them. The DACoE is also seeking to shape public sector engagement, recruitment and retention practices for data analysis professionals.

11.4 Collaboration with research community

ABS is establishing a collaboration network with leading Australian researchers in the field of data analytics to advance the research objectives of the Big Data Flagship Project. In particular, the project will draw on the expertise of the Image Processing and Remote Sensing Group at the Canberra campus of the University of New South Wales and the Advanced Analytics Institute at University Technology Sydney for areas such as satellite sensing and predictive modelling.

ABS is also an industry partner of a Centre of Excellence for Mathematical and Statistical Frontiers of Big Data, Big Models and New Insights, headed by the eminent mathematical statistician, Professor Peter Hall, of University of Melbourne. The Centre, comprising a multi-disciplinary team of statisticians, mathematicians, computational specialists and computer scientists, is funded by the Australian Research Council for a total of A$20 million over seven years . As an industry partner, the ABS was able to influence the Centre's research program to include research themes such as data fusion and integration, which are of significant interest to the ABS.

12. CONCLUDING REMARKS

Official statisticians have been dealing with a diversity of data sources for decades. Whilst new sources from Big Data provide an opportunity for official statisticians to deliver a more efficient and effective statistical service, in deciding whether to embrace a particular Big Data source, we argue that there are a number of threshold considerations, namely, business need, business benefit, and the validity of using the source for official statistics for finite population inferences, or analytic inferences. The Data Quality Framework is useful in assessing the quality of the Big Data sources, and for assessing fitness of purpose of use of Big Data.

This paper also provides a Bayesian framework for Big Data inferences, based on conceptualised transformation, sampling and censoring processes applied to the Big Data measurements. Proper inference will require modelling of all three processes, which can be very complex, if at all possible. However, in situations where ignorability conditions are fulfilled, inference can be made on the Big Data measurements as if they are acquired from a random sample.

Until recently, ABS' progress in Big Data domain has been primarily review and monitoring of industry developments while contributing to external strategic and concept development activities. The ABS Big Data Flagship Project provides the opportunity to gain practical experience in assessing the business, statistical, technical, computational and other issues outlined in this paper. ABS participation in national and international activities on Big Data will also help it share experience and knowledge, and collaboration with academics will help ABS better acquire the capability to address business problems using Big Data as a part of the solution. Finally, these and related initiatives have been summarised in an ABS Big Data Strategy paper (ABS, 2014).

ACKNOWLEDGEMENTS

The authors would like to thank Brian Studman and Peter Radisich for their helpful comments on an earlier version of this paper.

REFERENCES

Australian Bureau of Statistics (2010a) *The ABS Data Quality Framework*. <https://www.nss.gov.au/dataquality/aboutqualityframework.jsp>

—— (2010b) "Measuring Trust in Official Statistics: The Australian Experience", *OECD Statistics Newsletter*, 50, pp. 9–11. <http://www.oecd.org/std/46561991.pdf>

—— (2013) "Big Data and Official Statistics", *ABS Annual Report, 2012–13*, cat. no. 1001.0, ABS, Canberra, pp. 27–31. <http://www.ausstats.abs.gov.au/ausstats/abs@.nsf/mf/1001.0>

—— (2014) *Big Data Strategy*, unpublished report, ABS, Canberra.

Brackstone, G. (1999) "Managing Data Quality in a Statistical Agency", *Survey Methodology*, 25(2), pp. 139–149.

Choi, H. and Varian, H. (2011) *Predicting the Present with Google Trends*. <http://people.ischool.berkeley.edu/~hal/Papers/2011/ptp.pdf>

Clarke, F. and Hamilton, A. (2014) *From Metadata to Meaning: Semantic Statistics in the ABS*, unpublished manuscript, Australian Bureau of Statistics, Canberra.

Couper, M. (2013) "Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys", *Survey Research Methods*, 7(3), pp. 145–156. <https://ojs.ub.uni-konstanz.de/srm/article/view/5751/5289>

Crawford, K. (2013) "The Hidden Biases in Big Data", *Harvard Business Review Blog*. <http://blogs.hbr.org/2013/04/the-hidden-biases-in-big-data/>

Czepiel, S. (2002) *Maximum Likelihood Estimation of Logistic Regression Models: Theory and Implementation*. <http://czep.net/stat/mlelr.pdf>

Daas, P.J.H. and Puts, M.J.H. (2014) "Big Data as a Source of Statistical Information", *The Survey Statistician*, 69, pp. 22–31. <http://isi.cbs.nl/iass/N69.pdf>

Delécolle, R.; Maas, S.J.; Guerif, M. and Baret, F. (1992) "Remote Sensing and Crop Production Models: Present Trends", *ISPRS Journal of Photogrammetry and Remote Sensing*, 47(2–3), pp. 145–161.

Department of Finance and Deregulation (2013) *Australian Public Service Information and Communications Technology Strategy 2012– 2015*, DFD, Canberra.

Eurostat (2013) *Scheveningen Memorandum on Big Data and Official Statistics*. <http://ec.europa.eu/eurostat/documents/42577/43315/Scheveningen-memorandum-27-09-13>

Fellegi, I.P. and Sunter, A.B. (1969) "A Theory for Record Linkage", *Journal of the American Statistical Association*, 64(328), pp. 1183–1210.

Groves, R.M. and Heeringa, S.G. (2006) "Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs", *Journal of the Royal Statistical Society (Series A): Statistics in Society*, 169(3), pp. 439–457.

Harford, T. (2014) "Big Data: Are We Making a Big Mistake?", *Financial Times*. <http://www.ft.com/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html#axzz36aRk9emv>

Heckman, J.J. (1979) "Sample Selection Bias as a Specification Error", *Econometrica*, 47(1), pp. 153–161.

HLG BAS (2015) *High-Level Group for the Modernisation of Statistical Production and Services*, United Nations Economic Commission for Europe. <http://www1.unece.org/stat/platform/display/hlgbas/High-Level+Group+for+the+Modernisation+of+Statistical+Production+and+Services>

Kalman, R.E. (1960) "A New Approach to Linear Filtering and Prediction Problems", *Transactions of the ASME – Journal of Basic Engineering*, 82(Series D), pp. 35–45.

Kish, L. (1965) *Survey Sampling*, Wiley: New York.

Landsat (2013) *Landsat Project Description*, United States Geological Service. <http://landsat.usgs.gov/about_project_descriptions.php>

Little, R.J.A. (1982) "Models for Nonresponse in Sample Surveys", *Journal of the American Statistical Association*, 77(378), pp. 237–250.

Little, R.J.A. (1983) "Superpopulation Models for Nonresponse – The Ignorable Case", in W.G. Madow, I. Olkin and D.B. Rubin (eds), *Incomplete Data in Sample Surveys, Volume 2: Theory and Bibliographies*, New York, Academic Press, pp. 337–413.

Little, R.J.A. and Rubin, D.B. (2002) *Statistical Analysis with Missing Data, Second Edition*, Wiley, New York.

Madow, W.G.; Oklin, I. and Rubin, D.B. (1983) *Incomplete Data in Panel Surveys*, Academic Press, New York.

Manyika, J.; Chui, M.; Brown, B.; Bughin, J.; Dobbs, R.; Roxburgh, C. and Byers, A.H. (2011) *Big Data: The Next Frontier for Innovation, Competition, and Productivity*, Report, McKinsey Global Institute, McKinsey & Company. <http://www.mckinsey.com/insights/business_technology/big_data_the_next_frontier_for_innovation>

McCormick, T.H.; Raftery, A.E.; Madigan, D. and Burd, R.S. (2012) "Dynamic Logistic Regression and Dynamic Model Averaging for Binary Classification", *Biometrics*, 68(1), pp. 23–30.

Mui, Y.Q. (2014) "The Weird Google Searches about Unemployment and What They Say about the Economy", *The Washington Post*. <http://www.washingtonpost.com/blogs/wonkblog/wp/2014/05/30/the-weird-google-searches-of-the-unemployed-and-what-they-say-about-the-economy/>

OECD (2011) *Quality Dimensions, Core Values for OECD Statistics and Procedures for Planning and Evaluating Statistical Activities*. <http://www.oecd.org/std/21687665.pdf>

Pink, B.; Borowik J. and Lee G. (2009) "The Case for an International Statistical Innovation Program – Transforming National and International Statistics Systems", *Statistical Journal of the International Association for Official Statistics*, 26(3–4), pp. 125– 133.

Podesta, J.; Pritzker, P.; Moniz, E.J.; Holdren, J. and Zients, J. (2014) *Big Data: Seizing Opportunities, Preserving Values*, Executive Office of the President, Washington. <http://www.whitehouse.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf>

Puza, B. (2013) *Bayesian Statistics*, unpublished manuscript.

Puza, B. and O'Neill, T. (2006) "Selection Bias in Binary Data from Volunteer Surveys", *The Mathematical Scientist*, 31, pp. 85–94.

Richards, J.A. (2013) *Remote Sensing Digital Image Analysis – An Introduction, Fifth Edition*, Springer-Verlag, Berlin.

Rivera, J. and van der Meulen, R. (2013) *Gartner's Hype Cycle for Emerging Technologies*, Gartner Inc.. <http://www.gartner.com/newsroom/id/2575515>

Rubin, D.B. (1976) "Inference and Missing Data", *Biometrika*, 63(3), pp. 581–592.

Särndal, C.E.; Swensson, B. and Wretman, J. (1977) *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Smith, T.M.F. (1983) "On the Validity of Inferences from Non-Random Samples", *Journal of the Royal Statistical Society (Series A): Statistics in Society*, 146(4), pp. 394–403.

Tam, S.M. (1987) "Analysis of Repeated Surveys Using a Dynamic Linear Model", *International Statistical Review*, 55(1), pp. 67–73.

Tam, S.M. (1988) *Estimation in Finite Population Sampling – Optimality and Robustness*, unpublished PhD thesis, Australian National University. <https://digitalcollections.anu.edu.au/handle/1885/10575>

Tam, S.M. (2014) *Taming Non-Response in Finite Population Sampling with Auxiliary Information*, manuscript in preparation.

Tam, S.M. and Gross B. (2013) "Discussion of 'A Potential Framework for Integration of Architecture and Methodology to Improve Statistical Production Systems' by J.L. Eltinge, P.P. Biemer and A. Holmberg", *Journal of Official Statistics*, 29(1), pp. 209–211.

Tierney, L.; Kass, R.E. and Kadane, J.B. (1989) "Fully Exponential Laplace Approximations to Expectations and Variances of Nonpositive Functions", *Journal of the American Statistical Association*, 84(407), pp. 710–716.

United Nations Statistical Commission (2014) *Big Data and Modernisation of Statistical Systems*, report prepared for the Forty-Fifth Session of the Statistical Commission, New York, 4–7 March 2014. <http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf>

Wikipedia (2014) *Big Data*. <http://en.wikipedia.org/wiki/Big_data>

Wikle, C.K.; Berliner, L.M. and Cressie, N. (1998) "Hierarchical Bayesian Space-Time Models", *Environmental and Ecological Statistics*, 5(2), pp. 117–154.

Wu, L. (2010) *Mixed Effects Models for Complex Data*, Chapman and Hall/CRC Press, Boca Raton.

All URLs last viewed on Tuesday 3 March 2015

FOR MORE INFORMATION . . .

INTERNET	**www.abs.gov.au** The ABS website is the best place for data from our publications and information about the ABS.
LIBRARY	A range of ABS publications are available from public and tertiary libraries Australia wide. Contact your nearest library to determine whether it has the ABS statistics you require, or visit our website for a list of libraries.

INFORMATION AND REFERRAL SERVICE

Our consultants can help you access the full range of information published by the ABS that is available free of charge from our website, or purchase a hard copy publication. Information tailored to your needs can also be requested as a 'user pays' service. Specialists are on hand to help you with analytical or methodological advice.

PHONE	1300 135 070
EMAIL	client.services@abs.gov.au
FAX	1300 135 211
POST	Client Services, ABS, GPO Box 796, Sydney NSW 2001

FREE ACCESS TO STATISTICS

All statistics on the ABS website can be downloaded free of charge.

WEB ADDRESS	www.abs.gov.au

Produced by the Australian Bureau of Statistics