Small steps towards Big Data - Some initiatives by the Australian Bureau of Statistics

Journal:	*International Statistical Review*
Manuscript ID:	Draft
Wiley - Manuscript type:	Original Article
Date Submitted by the Author:	n/a
Complete List of Authors:	Tam, Siu-MIng; Australian Bureau of Statistics, Methodology and Data Management Division
Keywords:	Big Data, Quality Framework, Ignorability, Missing Data, Selection Bias, Validity of Statistical Inferences
Abstract:	New sources from Big Data provide an opportunity for official statisticians to deliver a more efficient and effective statistical service. This paper outlines a number of considerations for the official statistician when deciding whether to embrace a particular Big Data source in the regular production of official statistics. The principal considerations are relevance of the source, business benefit of using the source, and the validity of using the source for official statistics, for finite population inferences, or analytic inferences. This paper also provides a Bayesian framework for making 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, where certain sampling and censoring ignorability conditions are fulfilled, inference can be made on the Big Data measurements as if they are acquired from a random sample. This paper describes the ABS Big Data Flagship Project. ABS participation in national and international activities on Big Data will help it share experience and knowledge, and collaboration with academics will also help ABS better acquire the capability addressing business problems using Big Data as a part of the solution.

SCHOLARONE™
Manuscripts

Small Steps Towards Big Data

– Some Initiatives by the Australian Bureau of Statistics

Siu-Ming Tam and Frederic Clarke

Methodology and Data Management Division, Australian Bureau of Statistics, ABS House, 45, Benjamin Way, Belconnen, ACT 2615, Australia

E-mail: Siu-Ming.Tam@abs.gov.au

For Review Only

Abstract

Official statisticians have been dealing with a diversity of data sources for decades. However, new sources from Big Data provide an opportunity for official statisticians to deliver a more efficient and effective statistical service. This paper outlines a number of considerations for the official statistician when deciding whether to embrace a particular Big Data source in the regular production of official statistics. The principal considerations are relevance, business benefit, and the validity of using the source for official statistics, finite population inferences, or analytic inferences.

This paper also provides a Bayesian framework for making 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, where certain sampling and censoring ignorability conditions are fulfilled, inference can be made on the Big Data measurements as if they are acquired from a random sample.

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 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 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 solution.

Keywords: Big Data, Data Quality Framework, Ignorability, Missing Data, Selection Bias, Validity of Statistical Inferences.

For Review Only

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 UN 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 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 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 a not accepted by statisticians, since the objective

For Review Only

"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 ad Puts (2014).

In spite of these issues and concerns, it is our view that Big Data, Semantic Statistics (Clarke and Hamilton, 2013), 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 (UN 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
- 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).

For Review Only

2. Big Data and the ABS

Like other NSOs, the ABS has developed significant and relevant expertise in collecting and processing large amounts of data. In addition, the ABS (ABS, 2013) is:

- 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 ABS legislation;
- given its holdings on statistical benchmarks, uniquely positioned in the Australian context 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. The ABS publishes the concepts, sources, methods, and results of all collections, and it provides "level playing field" access to all users of ABS statistics.

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

3. Sources of Big Data

Census, survey or administrative data, and to a limited extent, business data (e.g. scanner data from supermarkets, motor vehicle sales data etc.) are the principal sources for the production of official statistics. Collectively, the wide variety of extant and emerging Big Data sources may be broadly categorised as follows (UN Statistical Commission, 2013):

For Review Only

- 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.

In short, Big Data sources may be viewed as arising from **b**usiness, **i**ndividual and **g**overnment activities.

The scale of Big Data is often characterised 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.
- Variety – the diversity of data sources, formats, media and content.

For Review Only

4. Opportunities and Challenges for the ABS

To continue to improve its statistical value proposition, the ABS strives to reduce the cost of statistical production, improve the timeliness 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, the ABS is undertaking some small steps 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;

For Review Only

- generating new analytical insights – enhancing the measurement and description of economic, social and environmental phenomena; and
- improving the operational efficiency and effectiveness of the ABS through use of para data created and captured from its statistical operations (Groves and Heeringa, 2006).

Several use cases currently under investigation in the ABS are outlined later in the paper. They entail a common set of business, methodological, and technological considerations: business benefit, validity of statistical inference, privacy and public trust, data integrity, data ownership and access, computational efficacy, and technology infrastructure.

While the primary focus is the exploitation of Big Data for statistical value, the application of commercial use cases to the ABS is also of value. These other use cases include:

- Improving ABS data provider and data consumer experiences;
- Improving ABS operational business efficiencies; and
- Monitoring our 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 criteria such as the cost and sustainability of statistical outputs, as well as the accuracy, relevance, consistency, interpretability, and timeliness of those outputs stipulated in Data Quality Frameworks (ABS, 2010a; Brackstone,1999; OECD, 2011).

For Review Only

As an example, the full data substitution of survey-based with satellite imagery data for producing agricultural statistics – such as land cover and crop yield – can be justified if the quality of estimates from satellite imagery data is at least as good as survey-based estimates, and there is an overall net reduction in the end-to-end production cost. However, the computational demands of acquiring, transferring, processing, integrating and analysing large imagery data sets are presently unknown. If the operating costs of the necessary storage, server and network infrastructure is higher than any saving made through eliminating survey collection, then from the production cost point-of-view there is no incentive for replacing survey-based data by remote sensing data.

On the other hand, cost is not always the primary consideration. It could be that the imagery data is considerably more accurate and/or timely, and that may be worth the additional cost. In this case, the business case is based on offering a higher quality product.

In addition, we note that the cost of survey data will increase over time but the cost of Big Data analytics is expected to decrease. Also, imagery data alone will not provide all of the information that is currently collected by Agriculture Surveys in their current form. This suggests that a “blended data” approach is more feasible for many potential applications of Big Data: statistics are derived from heterogeneous data sets that incorporate multiple sources – both traditional and novel – in a way that optimises the end-to-end value across all the benefit criteria.

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 for estimating finite

For Review Only

population parameters such as population means, totals, quantiles etc. rely on random samples, i.e. the selection mechanism does not depend on the values of the units not selected in the sample (Sarndal, Swensson and Wretman, 1977; Kish, 1965; Rubin, 2006); 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 adjustment, will be 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.

For Review Only

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 US public in the 2012 flu season. This was almost double the correct estimate of 6% produced 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.

For Review Only

6.1 A Theory for Big Data Statistical Inference

It is useful to conceptualise the following elements of an inference framework for Big Data:

1. Target population, U, is 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 and above. In the Remote Sensing example, this may be the agricultural land parcels of Australia;
2. Big Data population, U_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 "R" process – see point 7 below;

3. Vector of measurements of interest to the NSO, $\mathrm{M_U}$ – This could be consumer confidence or crop yields;
4. Vector of proxy measurements available from Big Data, $\mathrm{Y_B}$ – This provides the proxy variables, or covariates, to be used to predict $\mathrm{M_U}$. From points 1 and 2 above, we can consider, Y_B as a sample (random or otherwise) from $\mathrm{Y_U}$. In the Twitter example, Y_B could be the sentiment data to predict consumer confidence, $\mathrm{M_U}$. In Remote Sensing example, Y_B may comprise wavelengths from selected wavebands captured by remote sensing missions, for discrete pixels of sizes ranging between 10m^2 to 1Km^2, to predict the annual production of certain types of crops in Australia, $\mathrm{M_U}$.
5. A transformation (or measurement) process, "T", is generally required to transform the data, $\mathrm{Y_U}$, to the measurements of interest, $\mathrm{M_U}$. In the Remote Sensing example, this may be transforming pixels based on the observed wavelengths in selected wavebands captured by the remote sensing mission into crop types - a classification problem. This is a complex scientific process requiring detailed understanding of the reflectance characteristics of the different ground cover, which in turn are dependent on the selective spectral absorption characteristics associated with their biophysical and biochemical compositions (Richards, 2013 p12). For a detailed discussion of the "supervised classification" and "unsupervised classification" techniques see Richards, 2013, p 247 and p 319 respectively).

For Review Only

6. A sampling process – random or otherwise – "I" is used to conceptualise the selection (or inclusion) of $\mathrm{Y_B}$ from $\mathrm{Y_U}$. 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. With remote sensing data through Landsat (Landsat, 2013), one has the fortunate situation that the coverage of U and 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, Y_U, not available. Where the coverage of U_B is not the same as U, one could conceptualise a "R" process in play rendering observations in the target population, but not in the Big Data population, missing. 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.

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

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

For Review Only

Following Rubin (1976), Little (1982), Little (1983), and Smith (1983), the task of the statistician, using a Bayesian inference framework, is to predict the posterior distribution $\mathrm{f}(\mathrm{M_U}|\mathrm{Y_B},\mathrm{I},\mathrm{R})$, or simply $[\mathrm{M_U}|\mathrm{Y_B},\mathrm{I},\mathrm{R}]$ to simplify the notation. Writing $\mathrm{Y'_U} = (\mathrm{Y'_B},\mathrm{Y'_C})$ to split up the Y_U variables of the target population into those from the Big Data and the remainder, we have

$$[\mathrm{M_U}|\mathrm{Y_B},\mathrm{I},\mathrm{R}] \propto \iiint [\mathrm{M_U},\mathrm{Y_B},\mathrm{Y_C},\mathrm{I},\mathrm{R},\theta,\varphi]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C}$$

$$= \iiint [\mathrm{R}|\mathrm{M_U},\mathrm{Y_B},\mathrm{Y_C},\mathrm{I},\theta,\varphi]\,[\mathrm{I}|\mathrm{M_U},\mathrm{Y_B},\mathrm{Y_C},\theta,\varphi][\mathrm{M_U},\mathrm{Y_B},\mathrm{Y_C},\theta,\varphi]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C} \quad (1)$$

$$\propto \iiint [\mathrm{M_U},\mathrm{Y_B},\mathrm{Y_C},\theta,\varphi]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C}$$

$$= [\mathrm{M_U},\mathrm{Y_B}]$$

$$\propto [\mathrm{M_U}|\mathrm{Y_B}], \quad (2)$$

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

$$[\mathrm{R}|\mathrm{M_U},\mathrm{Y_B},\mathrm{Y_C},\mathrm{I},\theta,\varphi] = [\mathrm{R}|\mathrm{Y_B}]$$

and

$$[\mathrm{I}|\mathrm{M_U},\mathrm{Y_B},\mathrm{Y_C},\theta,\varphi] = [\mathrm{I}|\mathrm{Y_B}]$$

are satisfied. In other words, subject to the fulfilment of these conditions, the scientific process to translate the Big Data observations Y into the measurements of interest M can be performed by disregarding the sampling and censoring processes.

Now

$$[\mathrm{M_U}|\mathrm{Y_B}] \propto \iiint [\mathrm{M_U},\mathrm{Y_B},\mathrm{Y_C},\theta,\varphi]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C}$$

$$= \iiint [\mathrm{M_U},\mathrm{Y_U},\theta,\varphi]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C}$$

For Review Only

$$= \iiint [\mathrm{M_U}|\mathrm{Y_U}, \varphi][\varphi][\mathrm{Y_U}|\theta][\theta]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C}$$

$$= \int \mathrm{E}_{\varphi}(\mathrm{M_U}|\,\mathrm{Y_U})\mathrm{E}_{\theta}(\mathrm{Y_U})\mathrm{dY_C}, \qquad (3)$$

where $\mathrm{E}_{\varphi}[\cdot]$ and $\mathrm{E}_{\theta}[\cdot]$ denote the expectation with respect to $\mathrm{f}(\varphi)$ and $\mathrm{f}(\theta)$ repectively.

For analytic inferences, the interest will be on estimating the parameter ψ of the pdf, $\mathrm{f}(\mathrm{M_U};\ \psi)$. Where $\mathrm{E}(M_U) = Z_U\beta, \text{and } Z_U$ is a matrix of covariates, and the variance-covariance matrix of $\mathrm{M_U}$ is Ω, then $\psi = (\beta, \Omega)$. Now similar to the derivation of (1),

$$[\psi|\mathrm{Y_B}, \mathrm{I}, \mathrm{R}] \propto [\psi, \mathrm{Y_B}, \mathrm{I}, \mathrm{R}]$$

$$\propto \iiint [\,\mathrm{Y_B}, \mathrm{Y_C}\,, \mathrm{I}, \mathrm{R}, \theta, \varphi, \psi]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C}$$

$$= \iiint [\mathrm{R}|\mathrm{Y_B}, \mathrm{Y_C}\,, \mathrm{I}, \theta, \varphi, \psi]\,[\mathrm{I}|\mathrm{Y_B}, \mathrm{Y_C}\,, \theta, \varphi, \psi][\,\mathrm{Y_B}, \mathrm{Y_C}\,, \theta, \varphi, \psi]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C} \qquad (4)$$

$$\propto \iiint [\,\mathrm{Y_B}, \mathrm{Y_{C,}}\, \theta, \varphi, \psi]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C}$$

$$\propto [\psi|\mathrm{Y_B}], \qquad (5)$$

provided that the following ignorability conditions (II):

$$[\mathrm{R}|\mathrm{Y_B}, \mathrm{Y_{C,}}\mathrm{I}, \theta, \varphi, \psi] = [\mathrm{R}|\mathrm{Y_B}]$$

and

$$\left[\mathrm{I}\middle|\mathrm{Y_B}, \mathrm{Y_{C,}}\, \theta, \varphi, \psi\right] = [I|Y_B]$$

are satisfied. Likewise, where θ, φ, ψ are distinct (Rubin, 1976)

$$[\psi|\mathrm{Y_B}] \propto \int \iiint [\mathrm{M_U}, \mathrm{Y_{U,}}\, \theta, \varphi, \psi]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C}\mathrm{dM_U}$$

$$= [\psi] \int \iiint \left[\mathrm{M_U}\middle|\mathrm{Y_{U,}}\, \varphi\right][\varphi][\,\mathrm{Y_U}\,|\theta][\theta]\mathrm{d}\theta\mathrm{d}\varphi\mathrm{dY_C}\mathrm{dM_U}$$

For Review Only

$$= [\psi] \iint \mathrm{E}_{\varphi}(\mathrm{M}_{\mathrm{U}} | \mathrm{Y}_{\mathrm{U}}) \mathrm{E}_{\theta}(\mathrm{Y}_{\mathrm{U}}) \mathrm{dY}_{\mathrm{C}} \mathrm{dM}_{\mathrm{U}}$$

$$\propto [\psi] \int [\mathrm{M}_{\mathrm{U}} | \mathrm{Y}_{\mathrm{B}}] \mathrm{dM}_{\mathrm{U}}. \quad (6)$$

The Ignorability Conditions (I) and (II) are also known as Missing At Random conditions (Rubin, 1986). 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".

In reality, given the volume of Big Data, it is unlikely that the data set, Y_B, is used in full for official statistics production, without some form of sampling. Traditional design-based sampling methods (Kish, 1965; Sarndal et. al. 1977) fulfil the Ignorability Conditions specified above, and the sampling mechanism can therefore be ignored. If Y_s is a random sample of Y_B, denoting by Y_r, the complement of Y_B to make up Y_s, and assuming that we can observe Y_s fully, valid statistical inference can be made by "integrating out" Y_r as described below.

For descriptive inferences,

$$[\mathrm{M}_{\mathrm{U}} | \mathrm{Y}_{\mathrm{s}}, \mathrm{I}, \mathrm{R}] \propto \int \iiint [\mathrm{M}_{\mathrm{U}}, \mathrm{Y}_{\mathrm{B}}, \mathrm{Y}_{\mathrm{C}}, \mathrm{I}, \mathrm{R}, \theta, \varphi] \mathrm{d}\theta \mathrm{d}\varphi \mathrm{dY}_{\mathrm{C}} \, \mathrm{dY}_{\mathrm{r}}$$

$$\propto \int [\mathrm{M}_{\mathrm{U}}, \mathrm{Y}_{\mathrm{B}}] \mathrm{dY}_{\mathrm{r}} \quad (7)$$

$$\propto [\mathrm{M}_{\mathrm{U}} | \mathrm{Y}_{\mathrm{s}}]$$

with (7) following from (1) and subject to the fulfilment of the Ignorability Conditions (I) for the "I" and "R" processes. From (3), it follows that:

$$[\mathrm{M}_{\mathrm{U}} | \mathrm{Y}_{\mathrm{s}}] \propto \int \int \mathrm{E}_{\varphi}(\mathrm{M}_{\mathrm{U}} | \mathrm{Y}_{\mathrm{U}}) \mathrm{E}_{\theta}(\mathrm{Y}_{\mathrm{U}}) \mathrm{dY}_{\mathrm{C}} \mathrm{dY}_{\mathrm{r}}.$$

Likewise, for analytic inferences,

For Review Only

$$[\psi|Y_s, I, R] \propto \int \iiint [\, Y_B, Y_C\, , I, R, \theta, \varphi, \psi] d\theta d\varphi dY_C\, dY_r$$

$$\propto \int [\psi, Y_B] dY_r$$

$$\propto [\psi|Y_s]$$

provided that the Ignorability Conditions (II) hold. Furthermore, from (6), we have:

$$[\psi|Y_B] \propto [\psi] \iint [M_U|Y_B] dM_U\, dY_r.$$

6.2 Application of the theory to certain Big Data sets

The theory in Section 6.1 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, 1989; 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. We will illustrate the above theory using some applications of Big Data reported by NSOs.

In the case of the remote sensing example, M_U represents crop yields in the Australian continent, and Y_U the remote sensing data covering Australia. An interest

For Review Only

will be to use remote sensing data to predict crop yields, 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 Delecolle et. al. (1992). As the full data $\mathrm{Y_U}$ is available from Landsat (Landsat, 2013), $\mathrm{Y_B} = \mathrm{Y_U}$. 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 measure of interest (e.g. crop yields), and 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.

Thijssen and Daas (2013) found that the sentiments of the Twitter and Facebook, Y_s, are strongly correlated with consumer confidence, M_U, published by Statistics Netherlands. For this example, the "T" process is to transform the sentiment data into consumer confidence estimates. Compiling sentiment data from social media is a significant topic of research on its own - see Pang et.al. (2002) for the methodology using machine learning techniques. The target population for consumer confidence is consumers, covering in theory all age groups. A study by Beevolve of 36 million Twitter user profiles suggests that over 75% of users are within the 15-25 age group (Beevolve, 2012), and that the data is skewed towards the younger demographic "… since a lot of teenagers are comfortable disclosing their age on social networks as compared to their peers …" The statistician therefore

For Review Only

needs to determine if presence or non-presence in Twitter is dependent on sentiment. If it is, then proper analyses will require the modelling of the "R" and "I" processes, which can be challenging. With the strong correlation between sentiment and consumer confidence data reported by Statistics Netherlands, it appears that the "R" and "I" processes are ignorable. For sentiment data to replace consumer confidence data from an NSO, however, one also needs to be assured that the ignorability assumption continues to hold into the future.

Makita *et al* (2013) reported the use of mobile phone location and demographic data to estimate the day time and night time (de facto) population by "grid squares" and compared the accuracy of the estimates with those from the Statistic Bureau of Japan (SBJ). The mobile phone data is collected by a major mobile phone operator, NTT DOCOMO, which has a 40% market share of mobile phone services in Japan. In this Big Data application, the inference problem is to predict the population count, M_U, from a "sample" based on 40% of Japan's mobile phone location and demographic data, Y_s. The target population is population of all age groups in the grid squares. The Big Data population group is 15 –79, so modelling of the "R" process is required if one needs to estimate the target population in the grid squares – and this modelling, for populations under the age of 15, or over 79, is a challenge. In addition, as NTT DOCOMO represents 40% of the market share of mobile phone services, modelling of the "I" process needs to be considered, which can be ignored if use of NTT DOCOMO mobile phone services by a user is not dependent on where the user is located – which may not be true for those living in the remote part of a country which requires access to satellite, rather than 3G or 4G phones. Assessment of the accuracy of night time population estimates based on

For Review Only

mobile phone data with SBJ estimates suggests the “I” tends to be more ignorable for larger grids than smaller grids.

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, where warranted, 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. The ABS adheres to relevant legislation (e.g. the Census and Statistics Act and the Privacy Act) in setting 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 in the intelligence community 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 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.

For Review Only

8. Data Ownership and Access

Data ownership and access is a key issue for the ABS and one where there is a 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.

The Big Data opportunity for the ABS spans a vast array of direct sources beyond the government sector, such as commercial transactions (e.g. scanner data from supermarkets), sensors (e.g. satellite imagery), tracking devices (e.g. location-sensitive services associated with mobile devices), and administrative by-product (e.g. electronic health records). It also includes data derived from the results of analysis on public and private data sets, such as behavioural profiling (e.g. activity analysis of online purchases or credit card transactions), opinion mining (e.g. topic or sentiment analysis of Twitter feeds and on-line searches), and social network analysis (e.g. link and influence analysis of Facebook data).

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

For Review Only

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 ABS 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

For Review Only

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

For Review Only

hardware. A number of these technologies have value when applied to 'traditional' processing and analysis.

11. ABS Activities on Big Data

A number of activities 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 Big Data Flagship Project

The ABS Big Data Flagship Project – an initiative led by the 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

For Review Only

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 Open 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.

For Review Only

11.2. Participation in the Australian Public Service (APS) Data Analytics Initiatives

The APS Data Analytics Centre of Excellence (APS DACoE) was formed in late 2013 as a response to a recommendation in the Australian Public Service Information and Communications Technology Strategy 2012-2015. Its objective is 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.

Currently, the Centre is finalising 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.

11.3 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

For Review Only

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 imagery 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, mathematics, computational specialists and computer scientists, is funded by the Australian Research Council for a total of A$20m over 7 years . As an industry partner, the ABS was successful 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.

For Review Only

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, 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 the Big Data source.

This paper also provides a 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, 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 addressing 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).

For Review Only

Acknowledgement

The authors would like to thank Brian Studman for his comments on an earlier version of this paper. Views expressed in this paper are those of the authors and do not necessarily reflect the views of the ABS.

References

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

Australian Bureau of Statistics (2010b). Measuring trust in official statistics: the Australian Experience. The OECD Statistics Newsletter 50, 9 – 11.

Australian Bureau of Statistics (2013). Big data and official statistics. ABS Annual Report, 2012-13. 27 – 31.

Australian Bureau of Statistics (2014). Big data strategy. Unpublished report.

Beevolve (2012). An exhaustive study of Twitter users across the world. http://www.beevolve.com/twitter-statistics/

Brackstone, G. (1999). Managing data quality in a statistical agency. Survey Methodology 25, 139 – 149.

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/.

Daas, P. and Puts, M. (2014) Big data as a source of statistical information. The Survey Statistician 69, 22- 31. http://isi.cbs.nl/iass/N69.pdf

Department of Finance and Regulation (2013). Big Data Strategy – Issues paper. http://www.finance.gov.au/files/2013/03/Big-Data-Strategy-Issues-Paper1.pdf

Delecolle, 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), p 145 -161.

Eurostat (2013). Scheveningen Memorandum on Big Data and Official Statistics. http://epp.eurostat.ec.europa.eu/portal/page/portal/pgp_ess/0_DOCS/estat/SCHEVENINGEN_MEMORANDUM%20Final%20version.pdf

Fellegi, I. P. and Sunter, A. B. (1969). A theory of record linkage. Journal of the American Statistical Association 64, 1183 – 1210.

Groves, R. and Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs.

For Review Only

Hamilton, A. and Clark, F. (2014). From metadata to meaning: Semantic statistics in the ABS. Unpublished ABS manuscript.

Heckman, J. (1979). Sample selection bias as a specification error. Econometrica 47(1), 53 – 68.

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

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

Little, R.J.A. (1982). Models for non-response in sample surveys. Journal of the American Statistical Association 77, 237 – 250.

Little, R.J.A. (1983). Superpopulation models for non-response. In Incomplete Data in Sample Surveys, Ed. Madow, W., Oklin, I. and Rubin, D. Vol 2, 337 – 413.

Little, R.J.A. and Rubin, D. (2002). Statistical analysis with missing data, 2nd Edition. New York: Wiley.

Madow, W, Oklin. I, and Rubin, D. (1983). Incomplete data in panel surveys. New York: Academic Press.

Makita, N., Kimura, T., Kobayashi, M., and Oyabu Y. (2013). Can mobile phone network data be used to estimate small area population? A comparison from Japan. Journal of the International Association of Official Statistics, 29, 223 – 232.

OECD (2011). Quality dimensions, core values for OCED statistics and procedures for planning and evaluating statistical activities. http://www.oecd.org/std/21687665.pdf

For Review Only

Pang, B., Lee, L. and Vaithyanathan S. (2002). Thumbs up? Sentiment classification using machine learning techniques. Proceedings of the Conference on Empirical Methods in Natural Language Processing. 79 – 86.

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, 125 - 133.

Puza, B. and O'Neill, T. (2006). Selection bias in binary data from volunteer surveys. Mathematical Scientists 31, 85 – 94.

Richards, J. (2013). Remote Sensing Digital Image Analysis. 5th Edition. Berlin Heidelberg: Springer-Verlag.

Rivera, J. and van der Meulen, R. (2013). Gartner Hype Curve. http://www.gartner.com/newsroom/id/2575515

Rubin (1976). Inference and missing data. Biometrika 63, 581 – 592.

Sarndal, C.E., Swensson, B. and Wretman, J. (1977). Model-assisted survey sampling. New York: Springer-Verlag.

Smith, T.M.F. (1983). On the validity of inference from non-random samples. Journal of the Royal Statistical Society A146, 394 – 403.

Tam, S.M. and Gross B. (2013). Discussion. Journal of Official Statistics 29, 209 – 211.

Thijssen, T. and Daas, P. (2013). Big data insights from social media. Paper presented to the Conference of Director Generals of National Statistical Institutes. The Hague, Netherlands.

UN Statistical Commission (2014). Big data and modernisation of statistical systems.

For Review Only

http://unstats.un.org/unsd/statcom/doc14/2014-11-BigData-E.pdf

Winkle, C.K., Berliner, M. and Cressie, N. (1998). Hierarchical Bayesian space-time models. Environmental and Ecological Statistics 5, 117-154.

Wu, L. (2010). Mixed effects models for complex data. Boca Raton: Chapman and Hall.

For Review Only

Resume

Les offices statistiques mobilisent depuis longtemps des types de sources de données très variés. Aujourd'hui, un nouveau type de données appelé Données Massives leur offre la possibilité de produire une information de manière plus efficace et performante. Ce papier présente un certain nombre de considérations à prendre en compte par un office statistique au moment de décider si des Données Massives peuvent être intégrées à la production courante des statistiques officielles. Les éléments les plus importants à prendre en compte sont la pertinence de la source de données, le bénéfice à l'utiliser et sa fiabilité pour des statistiques officielles, soit pour les inférences en population finie, soit pour les inférences analytiques.

Ce papier fournit également un cadre d'inférence bayésien pour Données Massives basé sur les processus conceptualisés de transformation, d'échantillonnage et d'anonymisation appliqués à la mesure des Données Massives. L'inférence propre nécessitera la modélisation de ces trois processus, ce qui peut s'avérer complexe, à défaut impossible. Cependant, sous certaines conditions, l'inférence peut être réalisée sur Données Massives comme si elles étaient issues d'un échantillon aléatoire.

Jusqu'à récemment l'investissement du Bureau Australien de la Statistique (ABS) dans les Données Massives a consisté à suivre les développements industriels dans le domaine tout en réflechissant aux enjeux conceptuels et stratégiques. Ce papier décrit le projet phare de l'ABS sur les Données Massives qui a été mis en place pour offrir à l'institut une expérience pratique lui permettant d'évaluer les diffcultées liées à l'utilisation des Données Massives (problèmes administratifs et légaux, statistiques, techniques ou informatiques). La participation de l'ABS aux projets nationaux et internationaux sur les Données Massives permettra à l'institut de partager expérience et savoir dans ce domaine. La collaboration avec les universitaires aidera également l'ABS à améliorer sa capacité à surmonter les défis soulevés par l'utilisation des Données Massives.

For Review Only