Statistical Journal of the IAOS 34 (2018) 577–588
DOI 10.3233/SJI-170395
IOS Press

Big Data ethics and selection-bias: An official statistician's perspective

Siu-Ming Tam[a,*] and Jae-Kwang Kim[b]
[a]*Australian Bureau of Statistics, Belconnen, ACT 2617, Australia*
[b]*Iowa State University, Ames, IA 50011, USA*

Abstract. Official statistics are fundamental to democracy. With increasing demands for more relevant, frequent and rich statistical information, and declining resources, National Statistical Offices are continually looking for more cost effective ways in the production of official statistics. With the advent of the Internet of Things, they are increasingly exploring opportunities to harness Big Data as a source for official statistics. Use of Big Data, however, raises a number of ethical and statistical challenges for official statisticians, which are explored in this paper. This paper also proposes methods to adjust for self-selection bias, or coverage bias, normally associated with Big Data, by utilising random samples generally available from National Statistical Offices. We conclude that National Statistical Offices are generally well equipped to address these challenges.

Keywords: Big Data, equity of access, ethics, public good, selection bias

1. Introduction

In his presidential address to the Royal Statistical Society of the United Kingdom in 2008, Professor Tim Holt who was also the first Director of the Office of National Statistics, UK, said:

> "Official statistics are important. They are used to monitor public policies and public services and provide a window on the work of government. They are used to inform decision makers and the public about the *status quo* such as monitoring existing public policies and the current performance of the public service" [1].

His view is also underpinned by one of the United Nations Fundamental Principles of Official Statistics, in which it was stated that:

> "Official statistics provide an indispensable element in the information system of a democratic society, serving the government, the economy and the public with data about the economic, demographic, social and environmental situation" [2].

High quality official statistics rely on the availability of good data sources from which statistics are produced. Many such sources are available for official statistics and, in recent times, official statisticians have started looking at sources beyond the traditional data sources like censuses and surveys and administrative sources for compiling official statistics [3].

Census and survey data in which official statisticians have ultimate control on the what, how, and when to collect have been regarded as the gold standard data sources. Also called "designed" data, censuses and surveys are expensive to conduct and with increasing difficulties in establishing contact with, and declining cooperation from, providers, the representativeness of these sources for the target population which suffer from high non-response rates, is put in doubt.

Administrative sources have also been used by official statisticians over decades to compile official statistics, e.g. birth, death, migration records for vital statistics, custom manifests for trade statistics etc. National Statistical Offices (NSOs) which have good access to registers in countries have developed sophisticated sys-

*Corresponding author: Siu-Ming Tam, Australian Bureau of Statistics, 45 Benjamin Street, Belconnen, ACT 2617, Australia. Tel.: +61 26252 7160; E-mail: siu-ming.tam@abs.gov.au.

tems to exploit these sources for official statistics. For example, Statistics Netherlands have been using registers to replace the tradition censuses to compile population statistics over the past two decades [4].

In recent times, commercial transactions data are being increasingly used for official statistics. Typical examples are the use of scanner data, and web scraping, to get prices to compile the Consumer Price Index [5], and telematics data collected by freight companies to track movement of vehicles for safety and efficiency and to compile freight statistics. An experimental study in the use of telematics data in the Australian Bureau of Statistics (ABS) is given in [6].

The "Internet of Things", i.e. the interconnection via the Internet of computing devices embedded in everyday objects, enabling them to send and receive data, have provided potentially new data sources for official statistics. From the official statistics' perspective, such sources included data from sensors e.g. earth observations data or satellite imagery for crop classification or yield statistics, smart meters for energy use statistics, and mobile phone data for tourism or population statistics. Experimental studies on the use of satellite imagery data for crop classification in the ABS by using State Space Models have been carried out [7–9].

Data from behaviour metrics and online opinions are also increasingly available from technology companies and there are also novel uses of these sources e.g. nowcasting [10] or sentiment data [3].

From the official statistics' perspective, the collection of the above sources is considered as Big Data.

Tam and Clarke [11] outlined the benefits of using Big Data sources to improve the cost effectiveness in the production of official statistics, and also the challenges including the maintenance of trust in official statistics, which requires amongst other things, one's ability to draw reliable statistical inference from such data sources, some of which are well known to suffer from self-selection biases. Elliott and Valliant [12] have outlined two general approaches for correcting such biases, i.e. the use of pseudo weights and super population models which require strong assumptions about the properties of the data. Tam [9] proposed a framework for analysing earth observations data using dynamic super population models for predicting crop classification and yields.

In this paper, we will address two dimensions in the use of Big Data, namely, ethics, and statistical adjustments that may be used to adjust for selection biases, from the point of view of official statistics.

2. Ethics and trust

According to Wikpedia, ethics are:

> "... moral principles that govern one's behaviour or the conduct of an activity"

and the principles underpinning professional ethics include:

> "... honesty, integrity, transparency, accountability, confidentiality, objectivity, and acting lawfully..."

which are also espoused in the UN Fundamental Principles [2], the International Statistical Institute's "Declaration on Professional Ethics" [13], and the Codes of many professional statistical associations, e.g. American Statistical Association [14].

Trust is the currency for official statistics. If official statisticians act unethically, official statistics and the institutions producing these statistics will lose the trust from the users of the statistics. Whilst trust takes years to build, it does not take long to lose it, as well stated in the Dutch proverb: "Vertrouwen komt te voet en vertrekt te paard".

3. Ethical challenges

The ethical challenges faced by official statisticians when using Big Data are:

- The boundary between public good and private good;
- Privacy and confidentiality;
- Transparency;
- Equity of access; and
- Informed use of information.

In addressing these ethical challenges, the official statistician will be guided by such values as professional integrity, rights of society vs rights of data custodians, and rights of individuals.

3.1. Boundary between public good and private good

Unlike censuses and surveys which are created and owned by NSOs, and administrative data owned by government agencies, which are often shared with NSOs, most of the newer Big Data sets are created by commercial organisations who we shall describe as data custodians for the purpose of this paper.

In spite of having custodianship of these data sets, an interesting question arises as to who has ownership for the data. Are they the data custodians, or the individ-

uals who provided these data in their transaction with these commercial organisations, i.e. data subjects? For example, whether the information provided to a commercial company when creating an account with the company, and the subsequent activities carried out by the account holder and are logged by the companies, belongs to the company or the individual. Clarity on this issue may, however, but not necessarily always, be provided by referring to the terms and conditions of use of the facilities for transacting with the commercial organisations agreed by the data subjects. If the data is considered to be not owned by the data custodian, can the private company provide such data to an NSO for the production of official statistics? In Australia, a Federal Court in 2017 ruled the meta data related to customers using telecommunication services is not personal data and therefore the telecommunication services provider is not obliged to provide the data to the customer [16].

Provided that commercial organisations have ownership of the Big Data, what is the obligation for these organisations to provide the data to NSOs for public good purposes? Given the commercial value of the data sets, what is the boundary between public good and private good?

For decades, NSOs have been getting the cooperation of these organisations to provide information on their operations, e.g. inventories held, sales information, number of employees, industry etc. Are NSOs empowered by statistics legislation to require commercial organisations to provide information on their customers and their activities, and should they?

If they are prepared to release their data to the NSO for the compilation of official statistics, how does the official statistician ensure:

- the statistical products they produce from the Big Data do not directly compete with the commercial products produced by the same Big Data source, thus harming the commercial interests of the Big Data custodians?
- the anonymity of the data custodians is protected, where this is requested?

Provided that there is more than one data custodians to provide the data for the production of a particular field of official statistics, we argue that the second ethical challenge is not new and NSOs have developed and applied sophisticated statistical disclosure avoidance techniques in their data products, which can equally be applied to data provided by Big Data custodians. However, the first challenge is relatively new and requires good judgement in the development of data products by the NSO.

3.2. Privacy and confidentiality

Whilst in ordinary usage, the terms privacy and confidentiality are used interchangeably, they have different statistical meaning and different obligation on an NSO. In general terms, privacy is the right of an individual to control the information related to the person and be freed from intrusion. Confidentiality on the other hand is the obligation on the custodian of the private information to keep it secret and from being disclosed.

In deciding on the information to be collected in a census or survey, the NSO has to balance between respecting one's privacy, and the need for information for society's decision making, and public good. This balance is generally informed by consultation with the relevant stakeholders, including privacy commissioners, affected individuals and users of the statistical information. In the case of Australia, impact assessments on privacy are normally conducted on Australian Bureau of Statistics (ABS) collections as a matter of course, and for high profile collections, they are conducted by independent consultants. This will inform the NSO if the collection processes are consistent with the Information Privacy Principles and whether any privacy impact from the proposed statistical enquiry is within or beyond community expectation. As well, the statistics legislation requires the ABS to table in Parliament all proposed topics to be asked in any compulsory collections, which provides another check on whether the right balance between privacy and public good has been struck.

With the extensive development of statistical disclosure avoidance methods over the past decades, it can be argued that NSOs are well equipped in protecting the confidentiality of individual's private information in its data releases, whether they are in the form of aggregate statistics, or unit record files. As a matter of fact, there is a contemporary view that NSOs are too conservative in their policy on the privacy stance for releasing unit record files, which led to a number of NSOs, including the ABS, looking at beyond safe data protections, e.g. safe projects, safe users, safe setting and safe output [17,18], in more recent data release practices.

In the Big Data space, consideration on the privacy of the information provided by account holders or related to the account holder's activities will be different from that of a statistical collection. Unlike statistical collections which, backed by statistical legislation, oblige respondents to provide the information to the NSO, information from Big Data is either voluntarily

provided by account holders, or a by-product of their activities with the account. However, as mentioned before, there are questions on ownership of this information, whether the data custodians have the authority to release this information for use by others including NSOs, and if the information is released, whether is it done in a way that the privacy of the account holders is protected.

In deciding whether it is ethical to use a particular Big Data source in the production of official statistics, it is prudent for NSOs to consider whether it is legitimate to use the source, undertaking a privacy impact assessment to consider privacy issues arising from its use (e.g. integrating a Big Data source with NSO's census or survey data), and applying statistical disclosure avoidance techniques in its data releases involving the source.

3.3. Transparency

By transparency, we mean openness in the processes and methods used in the collection, processing, compilation and dissemination of the statistics. Transparency is important as it provides the information needed to allow users of official statistics to determine if valid statistical inferences can be made from the statistics, and also if the statistics are fit for the purposes to which the statistics are to be put.

This challenge is generally met by NSOs through the publication of methodologies used in the production of the statistics, including collection instruments, sampling methods, where applicable, non-response follow up or adjustment methods for assessing measures of uncertainty, and data visualisation.

When Big Data are used in the production of official statistics, selection bias correction such that those described in Section 4 below will be required. The transparency challenge for Big Data will be met if the NSO publications on methodologies will be extended to include selection bias correction methods.

3.4. Equity of access

With the advent of the Internet, governments are increasingly adopting a policy of open data and open access – see for example data.gov, data.gov.uk and data.gov.au. Increasingly too, NSOs are also making their data freely available to all, thus removing the financial barrier to access, and making statistical available to, and accessible by, all [23].

Because some statistics produced by the NSO are market sensitive, and owing to the need to ensure that the statistics are not seen to be subject to political interference, many NSOs have a policy of making statistics available to users only after official release. This ensures "a level playing field" for users, and no one will have an advantage, financially or otherwise, from prior access to the information.

In some NSOs, however, limited access to official statistics prior to official release is allowed under "lock up" arrangements, where users are not allowed to communicate with people outside the lock up, or leave it, until official release of the statistics has occurred. The benefit of such arrangement is to allow the users to prepare briefings on the statistics in time to be used after the lock up.

Where the statistics are compiled using Big Data, it is logical for existing policy on equity of access to be extended to these statistics, and where needed, lock ups to be arranged for pre-embargo access to official statistics compiled using Big Data sources.

3.5. Informed use of information

Informed use of the statistics requires, amongst other things, the provision of meta data describing the quality dimensions of the collection in accordance with quality frameworks – see for example, the ABS Data Quality Framework [24]. Quality Declarations can also be used to describe certain class of statistics [25].

Providing this information to facilitate informed use of the statistics is now common practice amongst NSOs and, provided the same practice is extended to Big Data sources, we do not see any new ethical challenges in this area with the use of Big Data sources in producing official statistics.

4. Selection bias correction

The challenges in using Big Data to make valid statistical inference about finite population (and super population) parameters are well known [11,19]. In particular, certain types of data sets from the Internet of Things can be subject to serious selection bias, the use of which will require well designed statistical adjustments for official statistics production.

4.1. Fundamental theorem for estimation error

In a key note speech to the 2016 Royal Statistical

Society conference, Meng [20] gave the following fundamental theorem for estimation error:

$$\hat{\mu}_g - \mu_g = \sqrt{\frac{1-f}{f}}\rho_{I,g}\sigma_g$$

where

$$\mu_g = \frac{\sum_1^N g(y_i)}{N}$$

$$\hat{\mu}_g = \frac{\sum_1^N I_i g(y_i)}{N_B}$$

g is a function of the the N values of the finite population, $y_1, ..., y_N$, $I_i = 1$ if y_i is included in the Big Data set (i.e. observed), or 0 otherwise, for $i = 1, ..., N$, $\rho_{I,g}$ is the correlation between I and $g(Y)$ and $\sigma_g^2 =$ Var$(g(Y))$, where the expectation is over a uniform distribution, unif{1,N}, $f = N_B/N$, and $N_B =$ the size of the Big Data set and is assumed to be fixed throughout this paper.

Assuming the sample of N_B observations of y_i is selected by a probability mechanism, ζ, which in the Big Data case, would generally be unknown to the analyst, then

$$E_\zeta(\hat{\mu}_g - \mu_g)^2 = \frac{1-f}{f}E_\zeta(\rho_{I,g}^2)\sigma_g^2$$

where the Defect Index [18], $E_\zeta(\rho_{I,g}^2)$, is simply $\frac{1}{N-1}$ if ζ is the probability distribution associated with simple random sampling. If the observations are recorded from Big Data, N_B will be very large (but still not equal to N so that $f < 1$), and if the selection bias is ignored, the variance of μ_g will be incorrectly assumed to be $(1-f)\sigma_g^2/N_B$, resulting in very small confidence interval and leading to incorrect inference about μ_g. This is what Meng [20] coined as the Paradox of Big Data i.e. the larger the Big Data size (but with $f < 1$), the more misleading it is for valid statistical inference. For proper inference, Meng [20] showed that the effective sample size, n_{eff}, for a Big Data with size of N_B is approximately $\frac{f}{(1-f)E_\zeta(\rho_{I,g}^2)}$.

Consider the special case of $g(y_i) = y_i$, and $y_i = 0$ or 1, i.e. Y_i is a binary variable. Suppose further

$$p = \Pr(Y_i = 1)$$

$$b = \Pr(I_i = 1|Y_i = 1) - \Pr(I_i = 1|Y_i = 0)$$

$$r = \frac{\Pr(I_i = 1|Y_i = 1)}{\Pr(I_i = 1|Y_i = 0)}.$$

Then it can be shown [21,22] that:

Table 1
Effective sample size to estimate the proportion of English speakers at home, with different values of f and b

		Response bias, b		
Big Data fraction, f	Big Data size	1%	5%	10%
1/10	2,340,189	507	20	5
1/4	5,850,473	3,171	127	32
1/3	7,722,624	5,525	221	55
1/2	11,700,946	12,684	507	127

Table 2
Statistical bias, B, in estimating the proportion of English speakers at home, with different values of f and r

		Response bias, r		
Big Data fraction, f	Big Data size	1.1	1.3	1.5
1/10	2,340,189	2%	4%	7%
1/4	5,850,473	2%	4%	7%
1/3	7,722,624	2%	4%	7%
1/2	11,700,946	2%	4%	7%

Note: +ve sign means over estimation.

$$n_{\mathit{eff}} = \frac{f^2 N}{b^2 p(1-p)(N-1) + f}$$

$$\doteq \frac{f^2}{b^2 p(1-p)},$$

given that N is large and provided that $b > 0$. The bias, B, of the estimator of p derived from Big Data is $B = \frac{-p(1-p)(1-r)}{1-(1-r)p}$.

Note that both B and the approximate n_{eff} are both independent of N_B.

4.2. An example on effective sample sizes and selection bias

To illustrate the power of Meng's Fundamental Theorem, assume that we want to estimate the proportion of Australians who speak English at home from a "Big Data" set which comprises between 10% to 50% of the Australian population (estimated to be over 23 million from the 2016 Census of Population). The proportion derived from the Census was 73%. Tables 1 and 2 provide values of the effective sample size, and the estimation bias, for different value of the Big Data size, b and r respectively.

It can be seen that the inferential value of Big Data is limited by the extent of selection (absolute) bias, b.

Similarly, the bias in estimating the proportion of English speakers at home depends on the relative selection bias, b.

4.3. Selection bias correction for proportions

How do we adjust for selection bias in Big Data?

In general, we can use consider the use of pseudo weights [12]. Let

$$\begin{aligned} w_i^{-1} &= \Pr(I_i = 1) \\ &= E_\varsigma(I_i), \\ \tilde{\mu}_g &= \frac{\sum_1^N I_i w_i g(y_i)}{N}, \end{aligned}$$

Then

$$\begin{aligned} E_\zeta(\tilde{\mu}_g) &= \frac{\sum_1^N E_\varsigma(I_i w_i g(Y_i))}{N} \\ &= \frac{\sum_1^N E_\varsigma(I_i)\{w_i g(Y_i)\}}{N} \\ &= \mu_g. \end{aligned}$$

In the sequel, we will write $E(.)$ instead of $E_\varsigma(.)$ to simplify notations.

In the special case of $g(y_i) = y_i, y_i = 0$ or 1, $\mu_g = p$,

$$\hat{\mu}_g = \hat{p}_B = \frac{\sum_1^{N_B} y_i}{N_B}$$

denotes the estimate of p based on the Big Data set, B etc. as above, from

$$\frac{\Pr(Y=0)}{\Pr(Y=1)} = \frac{\Pr(I=1|Y=1)}{\Pr(I=1|Y=0)} \frac{\Pr(Y=0|I=1)}{\Pr(Y=1|I=1)}$$

or

$$\theta = r\theta_B,$$

where

$$\theta = \frac{\Pr(Y=0)}{\Pr(Y=1)} \text{ and } \theta_B = \frac{\Pr(Y=0|I=1)}{\Pr(Y=1|I=1)}.$$

Noting $p = (1+\theta)^{-1}$ and $\hat{p}_B = (1+\hat{\theta}_B)^{-1}$, where

$$\hat{\theta}_B = \frac{\sum_1^{N_B} I(y_i = 0)}{\sum_1^{N_B} I(y_i = 1)} = \frac{N_B - \sum_1^{N_B} y_i}{\sum_1^{N_B} y_i}$$

is an estimate of θ_B, an estimate of p, $\hat{p}$, can be provided by the following:

$$\left(\frac{1}{\hat{p}} - 1\right) = \hat{r}\left(\frac{1}{\hat{p}_B} - 1\right)$$

or

$$\hat{p} = \frac{\hat{p}_B}{\hat{r} - \hat{p}_B\hat{r} + \hat{p}_B},$$

where

$$\hat{r} = \frac{\left(\frac{1}{\hat{p}} - 1\right)}{\left(\frac{1}{\hat{p}_B} - 1\right)}$$

is an estimate of r.

Puza and O'Neill [21] derived the same result by showing that

$$w_i = \frac{1 - \hat{p}(1-\hat{r})}{\hat{r}}$$

and thus

$$\begin{aligned} \hat{p} &= \frac{\sum_1^N I_i w_i y_i}{N_B} \\ &= \frac{1 - \hat{p}(1-\hat{r})}{\hat{r}} \hat{p}_B \end{aligned}$$

and hence

$$\hat{p} = \frac{\hat{p}_B}{\hat{r} - \hat{p}_B\hat{r} + \hat{p}_B}.$$

To estimate the variance of $\hat{p}$, *Var*$(\hat{p})$, note that using Taylor expansion:

$$\begin{aligned} \hat{p} &\doteq \frac{1}{(1+\theta)} - \frac{1}{(1+\theta)^2}(\hat{\theta} - \theta) \\ &= p - p^2(\hat{\theta} - \theta). \end{aligned}$$

Hence

$$\begin{aligned} \hat{Var}(\hat{p}) &\doteq \hat{p}^4 \hat{Var}(\hat{\theta}) \\ &= \hat{p}^4 \hat{Var}(\hat{r}\hat{\theta}_B) \\ &\doteq \hat{p}^4 \hat{\theta}_B^2 \hat{Var}(\hat{r}), \end{aligned}$$

noting that

$$\begin{aligned} Var(\hat{r}\hat{\theta}_B) &= E(\hat{r}\hat{\theta}_B - r\theta_B)^2 \\ &\doteq E(\hat{r}\theta_B - r\theta_B)^2 \\ &= \theta_B^2 Var(\hat{r}), \end{aligned}$$

as $\hat{\theta}_B \doteq \theta_B$ given $Var(\hat{\theta}_B) \doteq 0$ when N_B is large.

To obtain the approximately unbiased estimate, $\hat{p}$ of p and associated uncertainty, $\hat{SE}(\hat{p})$, the question remains on the estimation of r and its standard error. Using a random sample of the target population, A, available from the NSO and assuming that matching of the units in the random sample with the Big Data set is possible, after which, we can observe, for each of the n units in the random sample, the value of the variables, I_i and Y_i, to indicate if the unit is in the Big Data set ($I = 1$) or not ($I = 0$) and if $Y = 1$ or 0. Using these data, the following Table can be constructed from

Table 3
Counts of units with different values of I and Y from the Simple Random Sample

	$I=1$	$I=0$
$Y=1$	a	b
$Y=0$	c	d

Note: a, b, c, d denote unweighted counts.

which r can be estimated using maximum likelihood by

$$\hat{r} = \frac{\hat{\varphi}_1}{\hat{\varphi}_0} = \left\{ \frac{\frac{a}{a+b}}{\frac{c}{c+d}} \right\} = \left\{ \frac{\frac{a}{n_1}}{\frac{c}{n_0}} \right\}.$$

Using Taylor expansion, assuming the random sample, A, is drawn by simple random sampling without replacement and ignoring the finite population correction, it can be shown (see Appendix 1) the variance of $\hat{r}$ can be estimated by:

$$\hat{Var}(\hat{r}) \doteq \frac{1}{\hat{\varphi}_0^2} \hat{Var}(\hat{\varphi}_1 - \hat{r}\hat{\varphi}_0)$$

where

$$\begin{aligned} \hat{Var}(\hat{\varphi}_1) &\doteq n_1^{-1}\hat{\varphi}_1(1-\hat{\varphi}_1) \\ \hat{Var}(\hat{\varphi}_0) &\doteq n_0^{-1}\hat{\varphi}_0(1-\hat{\varphi}_0) \\ \hat{Cov}(\hat{\varphi}_1, \hat{\varphi}_0) &\doteq 0, \end{aligned}$$

which leads to

$$\begin{aligned} \hat{Var}(\hat{r}) &\doteq \frac{\hat{r}^2}{n_1}\left(\frac{1}{\hat{\varphi}_1} - 1\right) + \frac{\hat{r}^2}{n_0}\left(\frac{1}{\hat{\varphi}_0} - 1\right) \\ &= \frac{\hat{r}^2}{n}\left\{ \frac{1}{\hat{p}_{SRS}}\left(\frac{1}{\hat{\varphi}_1} - 1\right) \right. \\ &\quad \left. + \frac{1}{(1-\hat{p}_{SRS})}\left(\frac{1}{\hat{\varphi}_0} - 1\right) \right\} \end{aligned}$$

where

$$n = n_1 + n_0 \text{ and } \hat{p}_{SRS} = n_1/n.$$

Thus, noting that

$$\hat{\theta}_B = \left(\frac{1}{\hat{p}_B} - 1\right) = \left(\frac{1}{\hat{p}} - 1\right)\frac{1}{\hat{r}}$$

$$S\hat{E}(\hat{p}) \doteq \hat{p}(1-\hat{p}) \sqrt{\begin{array}{l} \frac{1}{n}\left\{ \frac{1}{\hat{p}_{SRS}}\left(\frac{1}{\hat{\varphi}_1} - 1\right) \right. \\ \left. + \frac{1}{(1-\hat{p}_{SRS})}\left(\frac{1}{\hat{\varphi}_0} - 1\right) \right\} \end{array}}.$$

In other words, the correction factor, κ, to be applied to the standard formulae for estimating the variance of $\hat{p}$ is:

$$\kappa = \sqrt{\begin{array}{l} \hat{p}(1-\hat{p})\left\{ \frac{1}{\hat{p}_{SRS}}\left(\frac{1}{\hat{p}_1} - 1\right) \right. \\ \left. + \frac{1}{(1-\hat{p}_{SRS})}\left(\frac{1}{\hat{p}_0} - 1\right) \right\} \end{array}}$$

giving $S\hat{E}(\hat{p}) \doteq \kappa\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.

4.4. An optimal choice of selection-bias adjusted estimator

Which estimator, $\hat{p}$ or $\hat{p}_{SRS}$, should one use to estimate p? If $\hat{\varphi}_1 \geqslant 1/2$ and $\hat{\varphi}_0 \geqslant 1/2$, from

$$\begin{aligned} \hat{Var}(\hat{p}) &= \frac{1}{n}\hat{p}^2(1-\hat{p})^2\left\{ \frac{1}{\hat{p}_{SRS}}\left(\frac{1}{\hat{\varphi}_1} - 1\right) \right. \\ &\quad \left. + \frac{1}{(1-\hat{p}_{SRS})}\left(\frac{1}{\hat{\varphi}_0} - 1\right) \right\} \\ &\leqslant \frac{1}{n}\hat{p}^2(1-\hat{p})^2\left(\frac{1}{\hat{p}_{SRS}} + \frac{1}{(1-\hat{p}_{SRS})}\right) \\ &= \frac{1}{n}\hat{p}^2(1-\hat{p})^2\left(\frac{1}{\hat{p}_{SRS}(1-\hat{p}_{SRS})}\right) \\ &\cong \frac{1}{n}\hat{p}_{SRS}(1-\hat{p}_{SRS}) \\ &= \hat{Var}(\hat{p}_{SRS}), \end{aligned}$$

i.e. the estimator from the Big Data is preferred, recalling that $\hat{p}_{SRS}$ is an estimator of p.

The strong requirement that $\min\{\hat{\varphi}_0, \hat{\varphi}_1\} \geqslant 1/2$ for the estimator from Big Data to perform better than that from a simple random sample is another demonstration of the corollary of the Fundamental Theorem of Estimation Error, namely, that the effective sample size of Big Data is not as good as one imagines, even with selection bias adjustment. In fact, as pointed out by a referee, unless the requirement $\min\{\hat{\varphi}_0, \hat{\varphi}_1\} \geqslant 1/2$ is met, the random sample plays the main role for estimation, with information being supplemented by the Big Data source, rather than the other way around.

However, noting that $\hat{Cov}(\hat{p}, \hat{p}_{SRS}) \cong 0$ (Appendix 2), and from the Gauss-Markov theorem, the best linear combination, in terms of minimal variance, of $\hat{p}$ and $\hat{p}_{SRS}$ is give by $\hat{p}$, where

$$\hat{p} = \frac{\hat{p}\hat{Var}(\hat{p})^{-1} + \hat{p}_{SRS}\hat{Var}(\hat{p}_{SRS})^{-1}}{\hat{Var}(\hat{p})^{-1} + \hat{Var}(\hat{p}_{SRS})^{-1}}$$

with

$$\hat{Var}(\hat{\hat{p}}) = \frac{1}{\hat{Var}(\hat{p})^{-1} + \hat{Var}(\hat{p}_{SRS})^{-1}}$$

being smaller than $\hat{Var}(\hat{p})$ or $\hat{Var}(\hat{p}_{SRS})$.

In other words, one can always get a better estimator by borrowing strength from both the biased-adjusted estimator from Big Data, and the estimator from the random sample.

4.5. An alternative method for selection bias correction

Alternatively, if matching of the Big Data units to the random sample is not possible, but there is auxiliary information available from both the Big Data set and the random sample, say $x_i = k$, where $k = 1, ..., K$, and provided that:

$$\begin{aligned}&\Pr(I_i = 1|Y_i = 1, X_i = k)\\ &= \Pr(I_i = 1|X_i = k)\\ &= \chi_k\end{aligned}$$

where χ_k denotes the propensity of unit i with auxiliary value x_i to be included in the Big Data set, then this propensity can be approximated by dividing the number of Big Data units that have a value of k by the estimated number of units in the target population that have the same value of k. Mathematically this can be denoted by:

$$\hat{\chi}_k = \frac{\sum\limits_{i\in B} I(x_i = k)}{\sum\limits_{i\in A} w_{Ai} I(x_i = k)}$$

where w_{Ai} denotes the weight of the ith unit in the random sample, and recalling B and A denotes the Big Data set and random sample respectively. In this case, a different estimate of p is possible, namely,

$$\hat{p} = \frac{\sum\limits_{i}^{N_B} y_i/\hat{\chi}_i}{N}$$

where $\hat{\chi}_i = \hat{\chi}_k$, if $x_i = k$.

Assuming the sample is drawn using simple random sampling without replacement, then

$$\begin{aligned}\hat{\chi}_k &= \frac{\sum\limits_{i\in B} I(x_i = k)}{\sum\limits_{i\in A} w_{Ai} I(x_i = k)}\\ &= \frac{n}{N}\frac{N_{B_k}}{n_k}\\ &= w_k^{-1}\end{aligned}$$

where

$$N_{B_k} = \sum_{i\in B} I(x_i = k)$$

$$n_k = \sum_{i\in A} I(x_i = k)$$

$$w_{Ai} = \frac{N}{n}.$$

Let

$$N_{B_k 1} = \sum_{i\in B} y_i I(x_i = k)$$

be the number of y_i's with $y_i = 1$, and $x_i = k$, then $\hat{p}$ may be rewritten as

$$\begin{aligned}\hat{p} &= \frac{\sum\limits_{j}^{K} w_j N_{B_j 1}}{N}\\ &= \sum_{j}^{K} \frac{n_j}{n}\frac{N_{B_j 1}}{N_{B_j}}.\end{aligned}$$

Then

$$\begin{aligned}Var(\hat{p}) \doteq &\sum_{j}^{K}\left(\frac{N_{B_j 1}}{N_{B_j}}\right)^2 Var\left(\frac{n_j}{n}\right)\\ &+\sum_{j=1}^{K}\sum_{l\neq j}^{K}\frac{N_{B_j 1}}{N_{B_j}}\frac{N_{B_l 1}}{N_{B_l}} Cov\left(\frac{n_j}{n}, \frac{n_l}{n}\right)\end{aligned}$$

given $Var\left(\frac{N_{B_j 1}}{N_{B_j}}\right) \doteq 0$ for large $N_{B_j 1}$ and N_{B_j}, $j = 1, ..., K$. It can be shown (Appendix 3) that an approximately unbiased estimator of $Var(\hat{p})$ is given by:

$$\begin{aligned}\hat{Var}(\hat{p}) &= \frac{1}{n-1}\left\{\sum_{j}^{K}\frac{n_j}{n}\left(\frac{N_{B_j 1}}{N_{B_j}}\right)^2 - \left(\sum_{j}^{K}\frac{n_j}{n}\frac{N_{B_j 1}}{N_{B_j}}\right)^2\right\}\\ &= \frac{1}{n-1}\left\{\sum_{j}^{K}\frac{n_j}{n}\left(\frac{N_{B_j 1}}{N_{B_j}}\right)^2 - \hat{p}^2\right\}.\end{aligned}$$

It is easy to see that the above method can be extended from binary to multi-nominal variables, which we shall not further discuss in this paper.

4.6. Relaxing the assumption of a constant r

In deriving the estimator of p in Section 4.3, we made the assumption of a constant r, that is,

$$r_i = \frac{\Pr(I_i = 1|Y_i = 1)}{\Pr(I_i = 1|Y_i = 0)} = r,$$

for $i = 1, ..., N$. Where this assumption cannot be made, but when auxiliary information x_i is available

for all the units, we can extend the idea of Section 4.5, when matching of the random sample to the Big Data set is possible, by using

$$\hat{p} = \frac{\sum_{i}^{N_B} y_i/\hat{\lambda}_i}{N}$$

where $\hat{\lambda}_i = \hat{\lambda}_k$, if $x_i = k$, and

$$\hat{\lambda}_k = \frac{\sum_{i \in A} y_i * I(x_i = k) * I_i}{\sum_{i \in A} y_i * I(x_i = k)},$$

for simple random sampling.

Let $n_{B_k 1} = \sum_{i \in A} y_i * I(x_i = k) * I_i$, $n_{k1} = \sum_{i \in A} y_i * I(x_i = k)$ then $\hat{\lambda}_k = n_{B_k}/n_k$,

$$\hat{p} = \frac{\sum_{j}^{K} w_j N_{B_j 1}}{N} = \sum_{j}^{K} \frac{n_{j1}}{n_{B_j 1}} \frac{N_{B_j 1}}{N}$$

and

$$Var(\hat{p}) \doteq \sum_{j}^{K} \left(\frac{N_{B_j 1}}{N}\right)^2 Var\left(\frac{n_{j1}}{n_{B_j 1}}\right) + \sum_{j=1}^{K} \sum_{l \neq j}^{K} \frac{N_{B_j 1}}{N} \frac{N_{B_l 1}}{N} Cov\left(\frac{n_{j1}}{n_{B_j 1}}, \frac{n_{l1}}{n_{B_l 1}}\right)$$

given $Var\left(\frac{N_{B_j 1}}{N}\right) \doteq 0$ for large $N_{B_j 1}$, $j = 1, ..., K$. Using Taylor expansion, $\frac{n_j}{n_{B_j}} = \frac{1}{\hat{\lambda}_j} = \frac{1}{\lambda_j} - \frac{1}{\lambda_j^2}(\hat{\lambda}_j - \lambda_j)$, it can be shown, using arguments similar to those of Appendix 2, that:

$$Var\left(\frac{n_{j1}}{n_{B_j 1}}\right) \doteq \frac{1}{\lambda_j^4} \frac{\lambda_j(1-\lambda_j)}{n_{j1}}$$

and

$$Cov\left(\frac{n_{j1}}{n_{B_j 1}}, \frac{n_{l1}}{n_{B_l 1}}\right) \doteq 0.$$

Hence

$$Var(\hat{p}) \doteq \sum_{j}^{K} \frac{1}{n_{B_j 1}} \left(\frac{N_{B_j 1}}{N}\right)^2 \frac{(1-\lambda_j)}{\lambda_j^2}.$$

5. Conclusion

In this paper, we have argued that there are no new ethical challenges in relation to equity of access and informed use of statistics compiled using Big Data sources.

However, there are new ethical challenges in determining whether the commercial information held by companies can be used by NSOs because of data ownership and the need to adhere to information privacy principles. If the information can be provided to NSOs for official statistics production, and provided that there are more than one data custodian, statistical disclosure avoidance techniques may be applied to protect the confidentiality of the information provided by the data custodians.

As well because of the self-selection bias of many Big Data sets, the inferential value of Big Data where such bias exists, can be substantially reduced. This paper also shows that, in the case of binary variables, the bias of the estimate remains constant and does not reduce even with increasing the size of the Big Data set. Using random samples of the target population available from the survey operations of a NSO, this paper also outlines methods for adjusting the self-selection bias to estimate proportions, depending on whether data matching is possible or if auxiliary information is available, and assessing the uncertainties of the resulting estimates.

Acknowledgments

The views expressed in this paper are those of the authors and do not necessarily represent the views of the Australian Bureau of Statistics. The research of the second author was partially supported by a grant from the US National Science Foundation (MMS – 1733572).

An earlier version of this paper was presented to the 61st World Statistical Congress held in 2017. We would like to thank Rory Tarnow-Mordi for motivating the derivation of the formulae for $V\hat{a}r(\hat{r})$.

We would also like to thank the referees for their valuable comments on an earlier version of this paper.

References

[1] Holt, T. (2007). Official statistics, public policy and public trust. *Journal of Royal Statistical Society*, A171, 1-20. Presidential Address.

[2] United Nations (2013). Fundamental principles of official statistics. https://unstats.un.org/unsd/dnss/gp/fp-english.pdf.

[3] Daas, P, Puts, M, Buelens, B, van den Hurk, P. (2015). Big Data as a Source for Official Statistics : Journal of Official Statistics, 31, 249-262.

[4] Nordholt, E. (2005). The Dutch virtual Census 2001: A new approach by combining different sources – IOS Press Statistical Journal of the United Nations Economic Commission for Europe, 22, 25-37.

[5] Australian Bureau of Statistics (2016).
[6] Husek, N. (2017). Telematics data for official statistics – an experience with Big Data. Submitted for publication.
[7] Marley, J, Defina, R, Traeger, K, Elazar, D, Amarasinghe, A, Biggs, G, Tam. S-M. (2016). Investigative Pilot Report (unpublished).
[8] Tam, S-M. (1987). Analysis of a repeated survey using a dynamic linear model. *International Statistical Review*, 55, 63-73.
[9] Tam, S-M. (2015). A Statistical Framework for Analysing Big Data. *Survey Statistician*, 72, 36-51.
[10] Choi, H, Varian, R. (2009). Predicting initial claims for unemployment insurance using Google Trends. Google Technical Report. https://static.googleusercontent.com/media/research.google.com/en/archive/papers/initialclaimsUS.pdf.
[11] Tam, S-M, Clarke, F. (2015). Big Data, official statistics and some experience of the Australian Bureau of Statistics. *International Statistical Review*, 83, 436-448.
[12] Elliott, M, Valliant, R. (2017). Inference for non-probability samples. *Statistical Science*, 32, 249-264.
[13] International Statistical Institute (1986). ISI declaration of professional ethics. ISI Declaration-isi-web.org.
[14] American Statistical Association (2016). Ethical guidelines for statistical practice. Ethical Guidelines for Statistical Practice.
[15] Royal Statistical Society (1993). Code of coduct. http://www.rss.org.uk/favicon.ico.
[16] Sydney Morning Hearlad (2017). Federal Court rejects application for Telstra to supply 'personal' metadata. http://www.smh.com.au/technology/technology-news/federal-court-rejects-application-for-telstra-to-supply-personal-metadata-20170120-gtvc85.html.
[17] Felix, R. (2013). International access to restricted data – a principle based approach. *Journal of the International Association of Official Statistics*, 29, 289-300. International access to restricted data: A principles-based standards approach – IOS Press.
[18] Tam, S-M, Farley-Larmour, K, Gare, M. (2010). Supporting research and protecting confidentiality – ABS microdata access: current strategies and future directions. *Journal of the International Association of Official Statistics*, 26, 65-74. Supporting research and protecting confidentiality. ABS microdata access: Current strategies and future directions – IOS Press.
[19] Couper, M. (2013). Is the sky falling? *Survey Research Methods*, 7, 145-156. Is the Sky Falling? New Technology, Changing Media, and the Future of Surveys | Couper | Survey Research Methods.
[20] Meng, X. (2016). Statistical paradises and paradoxes in Big Data. Talk to the 2016 Royal Statistical Society Conference. https://www.youtube.com/subscribe_embed?usegapi=1&card=1&channelid=UC83oOOF9lg-g1XMT_UK1tUw&origin=https%3A%2F%2Fapis.google.com&gsrc=3p&jsh=m%3B%2F_%2Fscs%2Fabc-static%2F_%2Fjs%2Fk%3Dgapi.gapi.en.ellQXbSf-LI.O%2Fm%3D__features__%2Fam%3DAAg%2Frt%3Dj%2Fd%3D1%2Frs%3DAHpOoo9jm0At0b0B7I7G3MSvlepU00mZfA.
[21] Puza, B, O'Neill, T. (2006). Selection bias in binary data from voluntary surveys. *Mathematical Scientist*, 31, 85-94. selection bias in binary data from voluntary surveys – Google Scholar.
[22] Raghunathan, T. (2015). Statistical challenges in combining information from big and small data sources. Paper presented to the Expert Panel meeting at the National Academy of Science. https://deepblue.lib.umich.edu/bitstream/handle/2027.42/120417/NAS-Paper.pdf?sequence=1&isAllowed=y.
[23] Tam, S-M. (2008). Informing the nation – open access to statistical information in Australia. *Journal of the International Association of Official Statistics*, 24, 145-153. Informing the nation – open access to statistical information in Australia – IOS Press.
[24] Australian Bureau of Statistics (2009). ABS data quality framework. 1520.0 – ABS Data Quality Framework, May 2009.
[25] Tam, S-M, Kraayenbrink, R. (2006). Data communication – emerging international trends and practice of the Australian Bureau of Statistics. *Journal of the United Nations Economic Commission for Europe*, 23, 229-247. Data communication – Emerging international trends and practices of the Australian Bureau of Statistics – IOS Press.

Appendix 1

Let $r = \varphi_1/\varphi_0$. Using Taylor expansion, we have

$$\hat{r} = r + \frac{1}{\varphi_0}(\hat{\varphi}_1 - r\hat{\varphi}_0)$$

and

$$\hat{\varphi}_1 = \frac{a}{n_1} \doteq \varphi_1 + \frac{1}{E(n_1)}(a - \varphi_1 n_1)$$

where $\varphi_1 = E(a)/E(n_1) = A/N_1$ say and $E(n_1) = nN_1/N$, recalling N is the size of the target population. Under simple random sampling without replacement and ignoring n/N, we have

$$Var(a) = n\frac{A}{N}\left(1 - \frac{A}{N}\right)$$

$$Var(n_1) = n\frac{N_1}{N}\left(1 - \frac{N_1}{N}\right)$$

and

$$Cov(a, n_1) = n\frac{A}{N}\left(1 - \frac{N_1}{N}\right).$$

Thus

$$Var(a - \varphi_1 n_1) = Var(a) + \varphi_1^2 Var(n_1)$$

$$-2\varphi_1 Cov(a, n_1) = n\frac{A}{N}(1 - \varphi_1)$$

and

$$Var(\hat{\varphi}_1) \doteq \{E(n_1)\}^{-1}\varphi_1(1 - \varphi_1).$$

Similarly, $V\hat{a}r(\hat{\varphi}_0) \doteq \{E(n_0)\}^{-1}\hat{\varphi}_0(1 - \hat{\varphi}_0)$ follows the same arguments. Finally, proof of

$$Cov(\hat{\varphi}_1, \hat{\varphi}_0) \cong 0$$

follows by noting

$$Cov(\hat{\varphi}_1, \hat{\varphi}_0) \doteq \frac{1}{E(n_1)}\frac{1}{E(n_0)}$$

$$
\begin{aligned}
&Cov(a-\varphi_1 n_1, c-\varphi_0 n_0)\\
&Cov(a,c) = Cov(a,\varphi_0 n_0) = Cov(\varphi_1 n_1, c)\\
&\quad = Cov(\varphi_1 n_1, \varphi_0 n_0) = -n_A \left(\frac{A}{N}\right)\left(\frac{C}{N}\right),
\end{aligned}
$$

where

$$C = E(c).$$

Appendix 2

Using Taylor expansion, we have

$$
\begin{aligned}
\hat{p} &\doteq p - p^2(\hat{\theta}-\theta)\\
&= p - p^2(\hat{r}\hat{\theta}_B - \theta)\\
\hat{r} &\doteq r + \frac{1}{\varphi_0}(\hat{\varphi}_1 - r\hat{\varphi}_0)
\end{aligned}
$$

and

$$\hat{\varphi}_1 = \frac{a}{n_1} \doteq \varphi_1 + \frac{1}{E(n_1)}(a-\varphi_1 n_1)$$

Then

$$
\begin{aligned}
Cov(\hat{p}, \hat{p}_{SRS}) &= -p^2 Cov(\hat{r}\hat{\theta}_B, \hat{p}_{SRS})\\
&\doteq -p^2\theta_B Cov(\hat{r}, \hat{p}_{SRS})\\
&\doteq -p^2\theta_B Cov\left(\frac{\hat{\varphi}_1 - r\hat{\varphi}_0}{\varphi_0}, \hat{p}_{SRS}\right)\\
&= 0
\end{aligned}
$$

given that $\hat{\theta}_B \doteq \theta_B$ as the size of N_B is large, and noting

$$
\begin{aligned}
&Cov(\hat{\varphi}_1, \hat{p}_{SRS}) \doteq \frac{1}{nE(n_1)} Cov(a-\varphi_1 n_1, n_1)\\
&= \frac{1}{nE(n_1)}\left(1-\frac{N_1}{N}\right)\left(\frac{A}{N} - \varphi_1\frac{N_1}{N}\right)\\
&= 0,
\end{aligned}
$$

recalling $\varphi_1 = A/N_1$, and using

$$Var(n_1) = n\frac{N_1}{N}\left(1-\frac{N_1}{N}\right)$$

and

$$Cov(a, n_1) = n\frac{A}{N}\left(1-\frac{N_1}{N}\right),$$

from Appendix 2, Similarly $Cov(\hat{\varphi}_0, \hat{p}_{SRS}) = 0$.

Appendix 3

Let $N_j = E(n_j)$, then

$$
\begin{aligned}
&Var\left(\frac{n_j}{n}\right) = \frac{1}{n}\frac{N_j}{N}\left(1-\frac{N_j}{N}\right)\\
&Cov\left(\frac{n_j}{n}, \frac{n_j}{n}\right) = -\frac{1}{n}\frac{N_j}{N}\frac{N_l}{N}.\\
&Var(\hat{p}) \doteq \sum_j^K \left(\frac{N_{B_j 1}}{N_{B_j}}\right)^2 Var\left(\frac{n_j}{n}\right)\\
&+\sum_{j=1}^K\sum_{l\neq j}^K \frac{N_{B_j 1}}{N_{B_j}}\frac{N_{B_l 1}}{N_{B_l}} Cov\left(\frac{n_j}{n}, \frac{n_l}{n}\right)\\
&= \frac{1}{n}\sum_j^K \frac{N_j}{N}\left(1-\frac{N_j}{N}\right)\left(\frac{N_{B_j 1}}{N_{B_j}}\right)^2\\
&-\frac{1}{n}\sum_{j=1}^K\sum_{l\neq j}^K \frac{N_j}{N}\frac{N_l}{N}\frac{N_{B_j 1}}{N_{B_j}}\frac{N_{B_l 1}}{N_{B_l}}\\
&= \frac{1}{n}\sum_j^K \frac{N_j}{N}\left(\frac{N_{B_j 1}}{N_{B_j}}\right)^2 - \frac{1}{n}\left\{\sum_j^K \frac{N_j}{N}\frac{N_{B_j 1}}{N_{B_j}}\right\}^2\\
&= \frac{1}{n}\left\{\sum_j^K P_j P_{B1|j}^2 - \left(\sum_j^K P_j P_{B1|j}\right)^2\right\},
\end{aligned}
$$

where $P_j = \frac{N_j}{N}$ and $P_{B1|j} = \frac{N_{B_j 1}}{N_{B_j}}$. Let $\hat{P}_j = \frac{n_j}{n}$, noting that

$$
\begin{aligned}
&E\left\{\left(\sum_j^K \hat{P}_j P_{B1|j}\right)^2\right\} = \left(\sum_j^K P_j P_{B1|j}\right)^2\\
&+ Var\left(\sum_j^K \hat{P}_j P_{B1|j}\right) = \left(\sum_j^K P_j P_{B1|j}\right)^2\\
&+ Var(\hat{p}).
\end{aligned}
$$

an approximately unbiased estimator of $Var(\hat{p})$ is given by:

$$
\begin{aligned}
&V\hat{a}r(\hat{p}) = \frac{1}{n-1}\\
&\left\{\sum_j^K \hat{P}_j P_{B1|j}^2 - \left(\sum_j^K \hat{P}_j P_{B1|j}\right)^2\right\}
\end{aligned}
$$

noting

$$
\begin{aligned}
E\left\{\hat{Var}(\hat{p})\right\} &= \frac{1}{n-1} \\
&\left\{\sum_{j}^{K} P_j P_{B1|j}^2 - E\left(\sum_{j}^{K} \hat{P}_j P_{B1|j}\right)^2\right\} \\
&= \frac{1}{n-1}\left\{\sum_{j}^{K} P_j P_{B1|j}^2 - \left(\sum_{j}^{K} P_j P_{B1|j}\right)^2\right. \\
&\quad \left. - Var\left(\sum_{j}^{K} \hat{P}_j P_{B1|j}\right)\right\} \\
&\doteq Var(\hat{p}).
\end{aligned}
$$