The Survey Statistician, 2020, Vol. 81, 21-35.

New and Emerging Methods

New Data Sources for Official Statistics – A Game Changer for Survey Statisticians?

Siu-Ming Tam[1] and Anders Holmberg[2]

[1] Australian Bureau of Statistics, University of Wollongong, Australia, stattam@gmail.com
[2] Australian Bureau of Statistics, anders.holmberg@abs.gov.au

Abstract

Faced with declining budgets, rising data collection costs and increasing demand for richer, more detailed and frequent statistics, National Statistical Offices are increasingly looking at using new data sources for the production of official statistics. However, as the inferential value of new data sources is limited by issues such as coverage bias and measurement errors, it is paramount that methods are developed and used to address those issues. In this article, we summarize methods which, given underlying data structures, are advocated in the literature to address under-coverage and measurement error. Finally, the article also proposes 10 "rules" for engaging with new data sources for the production of official statistics.

Keywords: big data, estimation, integrating data, secondary data sources, prediction methods, survey quality.

1 Introduction

Responding to the new Australian Bureau of Statistics (ABS) strategic directions on data leadership, the methodologists in the organisation nominated new data algorithms (also known as machine learning (ML) and artificial intelligence (AI)), new data sources and data integration as the new statistical frontiers, and pledged to prioritise research in these areas.

In his President's Invited paper delivered at the 62nd World Statistics Congress in August 2019 hosted jointly by the Malaysian Statistics Office and the International Statistical Institute, Professor Bradley Efron gave an interesting lecture on Data Science. Our take home messages from the lecture are that, as demonstrated in the examples shown in his talk, ML methods generally perform better than "classical" statistical methods in terms of prediction, but do not generally do well in attribution, i.e. understanding the relationship between response variables and the "features" (also known as auxiliary variables in classical statistics). According to his discussant, Professor Noel Cressie, Professor Efron's talk highlighted the distinction between science (which is to find out the truth) and

engineering (which is to make things work). Classical statistics have been developed to deal with the former, whilst ML/AI are developed to deal with the latter.

There is already an increasing trend in National Statistical Offices (NSO) to take up ML/AI in official statistics in, for example, predicting the occupancy status of dwellings for Census operations (Dzhumasheva, 2019), predicting the best time to contact respondents (Wang (2019)), coding operations (Gweon et al. (2018)), editing and imputation (Richman et al. (2002), Jentoft and Zhang (2019a), Ruiz (2018)), replacing a survey question by predictive modelling using registry data (Burger et al. (2019)) etc.. We further note, in some other NSOs, ML and AI, in combination with new data sources, are used to supplement or replace traditional data sources in the production of official statistics e.g. use of support vector machines and satellite imagery data to predict crop classification (Handbook, 2017).

Data integration, which links two or more data sets together that have overlapping population units, creates new data sets that will have more public value than either of its component data sets. For example, by linking migrants' data from immigration records with census data over time, the analyst can look at the settlement outcomes of different migration cohorts and develop better targeted policies for migrants. The relevance of data integration in this paper is that it is a process that creates new data sources for official statistics. However, as there will undoubtedly be linkage errors in fusing two or more data sets together, many integrated data sets will have characteristics similar to non-probability samples and raise challenges for statistical inference. Lothian et al. (2019) outlined the opportunities and challenges in linking data sets across time, space and sources, and proposed a schema for linking traditional and non-traditional data sets.

In this paper, we will outline some inference challenges and proposed solutions in the literature that come with new data sources. This is done by illustrating four type of data structures that survey statisticians may come across. Naturally, the subject of new data sources is huge and growing by the day, and we will give our, undoubtedly, biased views about the subject. It is hoped that our paper will be a catalyst for other survey or official statisticians to give their perspectives on the subject, to enrich the debate on the future methodological directions of official statistics.

2 New data sources and their inferential value

In this paper we shall use the terms "big data" and new data sources interchangeably. Whilst big data often are characterised by a number of V's, e.g. Volume, Velocity, Variety etc., they are, from the official statistician's perspective, just data sources that, similar to administrative data, censuses or surveys, may be used in the production of official statistics. A diagrammatic illustration of the possible data sources for official statistics is provided in Figure 1.

Challenges in using big data for finite population inference are well documented in the literature – see for example, Couper (2013), Baker et al. (2013), Hand (2018), Tam and Clarke (2015), Japec et al. (2015), Macfeely (2019), Tam and Kim (2018). One often cited misconception of big data is that the size of the data set will compensate for any deficiencies in the data. Using the fundamental theorem of estimation error by Meng (2018), Tam and Kim (2018) showed that when the response variable is binary and if

$$
\begin{aligned}
p &= \Pr(Y_i = 1) \\
b &= \Pr(I_i = 1 \mid Y_i = 1) - \Pr(I_i = 1 \mid Y_i = 0) \\
r &= \frac{\Pr(I_i = 1 \mid Y_i = 1)}{\Pr(I_i = 1 \mid Y_i = 0)},
\end{aligned}
$$

where b and r are different measures of response bias, the effective sample of the big data set is given by $n_{eff} = \dfrac{f^2 N}{b^2 p(1-p) + f}$ where $f = n_B / N$, n_B and N denote the known size of the big data and population respectively. Furthermore, the bias of the sample mean compiled from big data as an estimator (also known as B-sample mean) of the population mean is $\dfrac{-p(1-p)(1-r)}{1-(1-r)p}$. When $f = 1$ and, using the Bayes' Theorem, it is easily shown that $b = 0$, and $n_{eff} = n_B$; also it follows that $r = 1$ and the bias of the sample mean is zero, as expected (Tam et al. (2019)).

Figure 1: *Possible data sources for the production of official statistics*

When the response bias is non-negligible, the inferential value of big data is substantially reduced, and the bias of the B-sample mean estimator is non-zero. For example, the inferential value of the big data to estimate the proportion of English speakers at home during the 2016 Australian Census is illustrated in Table 1, and the bias of the B-sample mean is given in Table 2 (Tam and Kim (2018)). Note that in Table 2, as illustrated in the formulae above, the bias, given r, is the same regardless of the size of the big data sample.

Other work to assess and analyse bias in this context has been given by Biemer (2019) (building on Meng (2018)) provided an expression of the estimation error in terms of data encoding error and sample recruitment error, and by Mercer et al. (2017) who developed a framework for a quality assessment of the level of selection bias.

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

Big Data fraction, f	Big Data size	Response bias, b		
		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 in estimating the proportion of English speakers at home, with different values of f and r*

Big Data fraction, f	Big Data size	Response bias, r		
		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: The proportion of English speakers at home in the 2016 Australian Census was 73%.

3 Validity of descriptive inference from new data sources – Type 1 data structure

By descriptive inference, we mean making inference of the parameters of a finite population, e.g. population means, proportions or totals, which are the “bread and butter” work for official statisticians. For analytic inference with big data, the reader may refer to Kohler et al. (2019). In this section, we assume that presence of an additional data source, A, to assist with inference using the new (big) data source B. We also assume the response variable of interest is only available from B, but the same auxiliary variables are available from both B and A. Data from on-line panels fall in this type of data structure, which is depicted as Type 1 in Table 3 below.

Table 3: *Four types of data structures*

	Source	Response variable, Y	Auxiliary variable, X	Representativeness?
TYPE 1	New data source, B	A	A	No
	Additional source, A	NA	A	No
TYPE 2	New data source, B	A	A	No
	Probability sample survey source, A	NA	A	Yes
TYPE 3	New data source, B	A	A	No
	Probability sample survey source, A	A	A	Yes
TYPE 4	New data source, B	NA	A	No
	Probability sample survey source, A	A	A	Yes

Note: A denotes available, NA denotes not available

Denote by U the population of known size N. Let each population unit be associated with an outcome of interest, denoted by $y_{i,}$ for $i \in U$ and let n_B. denote the sample size of B Here we assume the common assumption that $B \subset U$, and there are initially no duplicated units in B. Let $\delta_i = 1$ for $i \in B$, and 0 otherwise. Assume the response variable, y_i and the auxiliary variable vector, x_i are observed for $i \in B$, and $X = \sum_{i \in U} x_i$ is initially known. We are interested to estimate $Y = \sum_{i \in U} y_{i,}$ and $\bar{Y} = Y / N$. Let $\bar{Y}_B = \sum_{i \in B} y_i / n_B$.

Zhang (2019a) developed Missing-at-Random (MAR) conditions under a superpopulation (SP) approach, or quasi randomisation (QR) approach for commonly used estimators from B to be unbiased. These are summarised in Table 4.

Table 4: *Conditions for unbiasedness*[1]

Estimator name	Estimator	Unbiasedness condition(s)
$B-sample$ expansion estimator	SP: $\hat{Y} = N\bar{Y}_B$	SP: $E(y_i \mid \delta_i = 1, i \in U) = E(y_i \mid i \in U) = \mu,$ a constant
	QR: $\hat{Y} = \sum_{i \in B} (y_i / \hat{\pi}),$ where $\pi = \Pr(\delta_i = 1)$	QR: $E(\delta_i; y_i, i \in U) = \pi,$ a constant
$B-sample$ calibration estimator	SP: $\hat{Y} = \sum_{i \in B} w_i y_i,$ where $\sum_{i \in B} w_i x_i = X$	SP: $E(y_i \mid x_i, i \in B) = E(y_i \mid x_i, i \in U)$
	QR: If x_i is the post-stratum dummy index $\hat{Y} = \sum_{j,k} (y_{jk} / \hat{\pi}_j),$ where $\pi_j = \Pr(\delta_{jk} = 1)$, j denotes stratum index and k denotes sample unit index within stratum, such that $\sum_{j,k} y_{jk} = \sum_{i \in B} y_i$	QR: $E(\delta_{jk}; y_i, jk \in U) = \pi_j,$ where $\delta_{jk} = 1$ if the unit is included in B, and 0 otherwise.
$B-sample$ inverse propensity weighted (IPW) estimator[2]	QR: $\hat{Y}_{IPW} = N \frac{\sum_{i \in B} (y_i / \hat{\pi}_i)}{\sum_{i \in B} (1 / \hat{\pi}_i)}$	QR: $\Pr(\delta_i = 1 \mid x_i; i \in U) = \Pr(\delta_i = 0 \mid x_i; i \in U) = \pi_i$

Notes:

(1) Under the SP approach, the Expectation Operator is with respect to the superpopulation. Under the QR approach, it is with respect to the inclusion probability distribution. Unbiasedness also includes Asymptotic Unbiasedness as defined in Zhang (2019a).

(2) Assuming $\pi_i = \pi(x_i; \eta) > 0$, a parametric probability of inclusion function, is completely determined by x_i, then $\hat{\pi}_i = \pi(x_i; \hat{\eta})$ where $\hat{\eta}$ is determined by solving (a) $\sum_{i \in B} x_i - \sum_{i \in U} \pi_i x_i = 0$ if x_i is known for $i \in U$, or $\sum_{i \in B} x_i - \sum_{i \in A} w_i \pi_i x_i = 0$ otherwise (Chen et al. (2018)). This uses a generalised (pseudo) estimation equation approach and assumes π_i is modelled by a logistic regression model;

or (b) $\sum_{i\in B}\pi_i x_i - \sum_{i\in A} w_i x_i = 0$ *(Kott and Chang (2010)), which uses a calibration weighting approach. Note that the IPW estimator proposed by Zhang (2019a) is* $\hat{Y} = \sum_{i\in B}(y_i / \hat{\pi}_i)$.

Bethlehem (2016) used simulation to show that the B-sample calibration estimator may be able to reduce the bias due to under-coverage or self-selection from on-line web panels. However, he concluded that this only works if the proper auxiliary variables are available. His results are also reaffirmed by simulations in Dever et a. (2008) and Schonlau et al. (2009). Noted that the work of Zhang (2019a) showed that a MAR assumption is needed to underpin those simulations.

Lee (2006) examined the performance of B-sample IPW estimator for on-line panel surveys and concluded that it can reduce, but not eliminate bias, at the expense of increasing variance. They also found that the relationship between the covariates and the response variable was important in forming propensity models, as weak relationship not only did not decrease bias, but also increased variance. Similarly, a MAR assumption is required to justify the simulations.

4 Validity of descriptive inference from new data sources – Type 2 data structure

There is a predominant view in recent literature that the best approach to harvest the information of big data is to combine them with probability sample survey data (Elliott and Valliant (2017), Hand (2018), Thompson (2018), Lohr and Raghunathan (2017)). We now consider the case when the additional source, A, comes from a probability sample, with the weight of the sampling units denoted by d_i. The Type 2 data structure is depicted in Table 3. Note that if the new data source, B, also comes from probability sample surveys, there is already a large body of literature covering this subject – see for example, Citro (2014), Kim (2011), Kim and Rao (2012) Kim et al. (2016), Merkouris (2004), Park et al. (2017), Wu (2004) – which will therefore not be covered in this paper.

In what follows, we outline the results known to the authors that describe methods to harness big data for official statistics.

Result 1 (Elliott and Valliant (2017)). Treating the additional source, A, as a reference sample, they proposed to the following to estimate π_i for the B-sample IPW estimator $\hat{Y}_{IPW}$ defined in Table 4:

$$\pi_i \propto \Pr(A_i = 1 \mid x_i, i \in U)\frac{\Pr(\delta_i = 1 \mid x_i, i \in B \cup A)}{\Pr(A_i = 1 \mid x_i, i \in B \cup A)},$$

where $A_i = 1$ if $i \in A$ and 0 if $i \in U \setminus A$. As for the IPW estimator, the conditions for their estimator to be asymptotically unbiased, in the QR sense, are

$$\Pr(\delta_i = 1 \mid x_i; i \in U) = \Pr(\delta_i = 1 \mid x_i; i \in A) = \Pr(\delta_i = 0 \mid x_i; i \in U) = \pi_i$$

(Zhang (2019a)).

Result 2 (Fuller (2009), Theorem 5.1.1). Let $\pi_i = \Pr(\delta_i = 1)$ and suppose that there exists a vector λ such that $\pi_i^{-1} = x_i'\lambda$. Assume that the weights for the units in the probability sample survey are d_i, then under some regularity conditions, the regression estimator, $\hat{Y}_{\mathrm{Reg}} = X\hat{\beta}$, or $\hat{Y}_{\mathrm{Reg}} = \sum_{i\in A} d_i x_i'\hat{\beta}$ if X is unknown, where $\hat{\beta} = (\sum_{i\in B} x_i x_i')^{-1}\sum_{i\in B} x_i y_i$, is asymptotically design unbiased.

Results 3 (Yang and Kim (2017), Kim (2018)). Suppose $y_i = m(x_i, \beta) + e_i$ for some β with known function $m(.)$, and $E(e_i \mid x_i) = 0$. Under some regularity conditions, the imputation estimator

$\hat{Y}_{IM} = \sum_{i \in A} d_i m(x_i, \hat{\beta})$, where $\hat{\beta}$ is a consistent estimator of β, is approximately asymptotically SP unbiased, provided that $E(y_i \mid x_i, i \in B) = E(y_i \mid x_i, i \in U) > 0$.

Results 4 (Rivers (2007), Yang and Kim (2018)). If the Horvitz-Thompson estimator from sample A is denoted by $\sum_{i \in A} d_i y_i$ and $\hat{y}_i$ is the "nearest neighbour" (NN) of unit i in A, defined by $\hat{y}_i = y_{k_i}$, where $k_i = \arg\min_{j \in B} \| x_i - x_j \|$, then the nearest neighbour estimator $\hat{Y}_{NN} = \sum_{i \in A} d_i \hat{y}_i$ is, under some regularity conditions, asymptotically design unbiased, provided that $E(y_i \mid x_i, i \in B) = E(y_i \mid x_i, i \in U)$. The condition holds if $f(y_i \mid x_i, i \in B) = f(y_i \mid x_i, i \in U)$.

Result 5 (Chen et al. (2018), Kim and Wang (2019). Assume a parametric propensity model $\pi_i = \pi(x_i; \eta) > 0$ and a SP model $E(y_i \mid x_i, i \in U) = x_i' \beta$. The estimator $\hat{Y}_{DR} = \sum_{i \in A} w_i \hat{y}_i + \sum_{i \in B} \hat{\pi}_i^{-1} (y_i - \hat{y}_i)$ is doubly robust (Robins et al. (1994)), where $\hat{\pi}_i$ is determined by one of the three methods in Notes 2 under Table 4 above, and $\hat{y}_i = x_i' \hat{\beta}$ and $\hat{\beta}$ is as defined in Result 2, provided that

$$E(y_i \mid x_i, i \in B) = E(y_i \mid x_i, i \in U).$$

Remark 1. Note that all the results in this section require a MAR assumption, with the exception of Result 2 which requires the inverse of the selection probabilities to be of a specific form.

The performance of some of these estimators (NN, IPW and DR) were compared by Yang and Kim (2018). DR work better than NN for the all three investigated scenarios. The IPW is sensitive to non-linearity misspecification but work better than DR in two of the three scenarios.

5 Validity of descriptive inference from new data sources – Type 3 data structure

Big data can be used to substantially increase the efficiency of estimators from a sample survey, if they are used as benchmarks in the estimation process (Kim and Tam (2018), Tam et al. 2019)). The idea is to treat the population as comprising two strata, namely, a big data stratum which is fully observed, and a missing stratum, the information from which will be obtained from a probability sample survey. The data structure Type 3 under this scenario is depicted in Table 3.

Result 6 (Kim and Tam (2018)). The post-stratified estimator, $\hat{Y}_{PS}$, given by

$$\hat{Y}_{PS} = \sum_{i \in U} \delta_i y_i + N_C \frac{\sum_{i \in A} d_i (1 - \delta_i) y_i}{\sum_{i \in A} d_i (1 - \delta_i)}$$

is approximately design unbiased, where $N_C = N - N_B$. In addition, $Var(\hat{Y}_{PS}) \approx (1 - W_B) \frac{N^2}{n} S_C^2$ where $S_C^2 = (N - N_B)^{-1} \sum_1^N (1 - \delta_i)(y_i - \bar{Y}_C)^2$, $\bar{Y}_C = \sum_1^N (1 - \delta_i) y_i / N_C$, n is the sample size of A, and $W_B = n_B / N$, assuming $n / N \approx 0$.

Remark 2. If $S^2 = N^{-1} \sum_1^N (y_i - \bar{Y})^2$ and $\hat{Y}_A = N \sum_{i \in A} y_i / n$, and assuming simple random sampling for A, then

$$\frac{Var(\hat{Y}_{PS})}{Var(\hat{Y}_A)} = (1-W_B)\frac{S_C^2}{S^2} \square \ 1,$$

$\text{if } S_C^2 \approx S^2$. The factor $(1-W_B)$ is the under-coverage rate of the big data. Therefore, we have the expected result that the higher is the coverage rate, the lower will be the sampling variance of the estimator.

Kim and Tam (2018) also show that the Data Integration Estimator under certain circumstances is equal to the post stratified and thereby asymptotically design unbiased.

Remark 3. If there are duplications in the units in B, the definition of δ_i can be modified from zero/one to zero/number-of-times that the unit appears in B. In addition, if auxiliary variables x_i are available for all the units in B and A, the information may be harvested by modifying (1) as follows: $\sum_{i\in A} w_i v_i = (N, N_B, \sum_{i\in U}\delta_i \mathrm{y}_i, \sum_{i\in U} x_i)$ where $v_i = (1, 1-\delta_i, \delta_i y_i, x_i)$ (or $\sum_{i\in A} w_i v_i = (N, N_B, \sum_{i\in U}\delta_i \mathrm{y}_i, \sum_{i\in U}\delta_i x_i)$, if $X = \sum_{i\in U} x_i$ is unknown, where $v_i = (1, 1-\delta_i, \delta_i y_i, \delta_i x_i)$).

Remark 4. If there are measurements errors in B such that y_i^* is measured instead of y_i, this can be accommodated by modifying (1) as follows: $\sum_{i\in A} w_i v_i = (N, N_B, \sum_{i\in U}\delta_i \mathrm{y}_i^*, \sum_{i\in U} x_i)$, where $v_i = (1, 1-\delta_i, \delta_i y_i^*, x_i)$. If the measurement errors occur in the units in A, this can be accommodated by using $\hat{\mathrm{Y}}_{\mathrm{Re}gDI} = \sum_{i\in A} w_i \hat{y}_i$, where $\hat{y}_i$ is estimated from a measurement error model based on the observations $\{(y_i, y_i^*), i \in A \cap B\}$.

Remark 5. If there is non-response in the probability sample survey, A, we can use a QR approach for $\hat{\mathrm{Y}}_{\mathrm{Re}gDI}$ with a not missing-at-random (NMAR) propensity model, as follow. Let r_i be 1 if the i[th] unit in A is a respondent, and 0 if it is a non-respondent. Even if $r_i = 1$, y_i can be observed from the B sample for units with $\delta_i = 1$. One can therefore assume a more general parametric response model: $P(r_i = 1 \mid x_i, y_i) = P(x_i, y_i; \phi_x, \phi_y)$, where $\phi_x, \text{ and } \phi_y$ are unknown parameters. The parameters can be consistently determined by solving the following estimation equations, provided that $f(r_i, \delta_i) = f(r_i) f(\delta_i)$:

$$\sum_{i\in A} \frac{d_i r_i}{p(x_i, y_i; \phi_x, \phi_y)} \begin{pmatrix} x_i \\ \delta_i y_i \end{pmatrix} = \sum_{i\in A} \begin{pmatrix} x_i \\ \delta_i y_i \end{pmatrix}.$$

Once $\hat{\phi}_x \text{ and } \hat{\phi}_y$ have been determined, the final weights for $\hat{\mathrm{Y}}_{\mathrm{Re}gDI}$ are determined by minimising $\sum_{i\in A} d_i r_i^{-1} (\frac{w_i}{d_i r_i^{-1}} - 1)^2$ subject to a calibration constraint, e.g. $\sum_{i\in A} r_i w_i v_i = (N, N_B, \sum_{i\in U}\delta_i \mathrm{y}_i)$.

Kim and Tam (2018) showed through Monte Carlo simulations that the RegDI estimator outperforms other estimators in a set of scenarios that reflect measurement errors situations and model misspecifications. Result 6 can also be used for Type 2 data structure, by using the following vector, $v_i = (1, 1-\delta_i, x_i)$ in (1).

6 Validity of descriptive inference from new data sources – Type 4 data structure

Many of the big data sets do not have the variable of interest to the official statistician. For example, in agricultural statistics, official statisticians collect information on crops e.g. classifications and yields from agricultural censuses and surveys, but the Satellite imaginary data which may be used to predict crop classifications or yields, only contain information on wavelengths.

The general approach in using this type of big data – refer to Type 4 in Table 3 - is to use a training data set to develop a statistical model (or train an algorithm in data science) for prediction – see for example Handbook (2017). We see an explosion of machine learning and artificial intelligence methods applied to this type of data structure for prediction. It is beyond the scope of this paper to cover this big body of literature. Instead, we give a few relevant references on applications to official statistics for the interested reader (Carfagna and Gallego (2006), Daas and Puts (2014), Daas et al. (2015), Husek (2018) and Richman (2009), Saar-Tsechansky et al (2007), Tam (2015)).

Alternatively, this type of data structure can be handled by the method outlined in Remark 4 above, i.e. treat the satellite imagery data as y_i^*, and use the training data set to build a measurement error model to predict y_i.

7 Other innovative ways of using big data

There are many other innovative methods of using big data for finite population inference. For example, transactions data, also known as scanner data, are being used by a number of national statistical offices for the compilation of price relatives for the consumer price index (CPI). Accompanied with this, new methods, known as multilateral index methods, have been developed which are considered to be one of the most effective ways to exploit the full amount information available in the transactions data – see ABS (2017), Ivancic et al. (2011).

To address huge reporting load from household expenditure surveys, and address reporting errors, Zhang (2019b) proposed to use scanner data to compile the CPI weights, and use the household expenditure survey as an audit sample to assess the accuracy of the scanner data-based CPI weights. He also developed a test for assessing the accuracy, and also a measure for the uncertainty, of these weights.

In another application, Kim et al. (2018) used a two–level structural error model to combine the survey information for small areas, $\hat{Y}_{hi}$, which is subject to non-significant sampling error, with big data sources i.e. x_{hi}, which are subject to coverage and measurement errors. Their objective is to borrow strength from the different small areas to predict Y_i. The probabilistic structure of their model is summarised in Table 5.

Table 5: *Probabilistic structure of the Kim et al. (2018) model*

Model	Data	Parameter	Latent variable
Level-one	$\hat{\mathbf{Y}}_h = (\hat{Y}_{h1}, \ldots, \hat{Y}_{hn_h})$	$\boldsymbol{\theta}_h$	$\mathbf{Y}_h = (Y_{h1}, \ldots, Y_{hn_h})$
Level-two	$\hat{\boldsymbol{\theta}} = (\hat{\boldsymbol{\theta}}_1, \ldots, \hat{\boldsymbol{\theta}}_H)$	$\boldsymbol{\zeta}$	$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_H)$

Level 1 is essentially the two-equation Fay-Harriot model combined using Bayes formulae, i.e. $h(Y_{hi} \mid x_{i,}\hat{Y}_{hi};\theta_h) \propto g(\hat{Y}_{hi} \mid Y_{hi})h(Y_{hi} \mid x_{hi};\theta_h)$. The MLE of $\hat{\theta}_{hi}$ estimated by EM algorithm are then used as "observed" inputs to estimate the MLE of ζ, again using the EM algorithm. The best prediction of Y_{hi} is then given by $\hat{y}_{hi}^{**}$, where $\hat{y}_{hi}^{**} = E\{E(Y_{hi} \mid \hat{Y}_{hi}, x_{hi};\theta_h) \mid \hat{\theta}_h, \hat{\zeta}\}$.

8 Understanding the whole statistical production process

So far, we have only examined one, albeit vital, phase of the statistical process, namely estimation. Lothian et al. (2019) argued that we need to understand the whole statistical production process when dealing with non-traditional data sources. To achieve this understanding, they proposed a strategy for structuring, analyzing problems and answering questions, based on a system of statistical base registers, plus consistent monitoring and maintenance strategies. These statistical

registers are to serve as lighthouses for illuminating 'trusted' estimation procedures and provide a benchmark for comparing and investigating representativeness concerns. Their schema includes a framework for:

- structuring the non-probabilistic data;
- making it useful for cause and effect statistical inference;
- incrementally developing, designing and maintaining the database system; and
- inserting total survey error concepts into the schema.

Recognising that non-representativeness is a key issue for new data sources, they recommended the B-sample calibration estimator. We believe that the methods outlined in the earlier sections of this paper will provide a larger tool set for bias reduction to be used in the schema.

9 Concluding remarks

When should big data be used? Tam and Van Halderen (2019) outlined ten rules of big data engagement for the production of official statistics. These are summarised in Table 6 below.

Table 6: *Ten rules of big data engagement in official statistics?*

Non-Negotiable	Essential
1 Use big data as a solution to a well-defined statistical need	8 The use of big data reduces provider load
2 The long-term supply of the big data should be certain	9 The use of big data produces better statistics
3 Social license issues must be addressed	10 The use of big data is a fail safe
4 The big data is impartial	
5 Security and confidentiality issues have been addressed	
6 The big data is a cost effective alternative or supplement to traditional, statistical data sources	
7 Statistics are amenable to valid statistical inferences	

Of these, seven rules are considered as "non-negotiables", and the remaining three rules are considered essential. Even though in this paper, we have only discussed one of the seven "non-negotiables", i.e. statistics produced from new data sources are amenable to valid statistical inferences, it should be remembered that there are other important considerations to be made before using them.

From the results presented in this paper, it can be concluded that:

- where the response variable is available from a probability sample, A, and where it is possible to match the units in A with B, the RegDI estimator is the preferred estimator. Where there is no measurement errorin A, the estimator is approximately design-unbiased. If there is partial or unit non-response in A, the non-response can be modelled using NMAR assumption;

- where the response variable is not available in the probability sample, A, but auxiliary variables are available from both A and B, such that MAR can be assumed, the DR estimator is a failsafe estimator and preferred. Alternatively, the RegDI may also be used where matching of the units in A and B is possible;
- where the response variable is not available, and where the new data source does not come from a probability sample, but where MAR can be assumed, the B-sample IPW estimator or the B-sample calibration estimator may be used;
- Regardless, the availability of good auxiliary variables which are correlated with the response variable is vital for bias reduction for these types of estimators (Bethlehem (2016)). In passing, we note the simulation results in Buelens et al. (2018) which, by comparing the B-sample expansion estimator, and calibration estimator, with a number of commonly used machine learning techniques e.g. regression trees, artificial neural networks and support vector machines, showed that the latter techniques perform better in bias reduction; and
- Importantly, it is vital to decide when to engage with new data sources. Where new data sources are used, it is also important to get on top of the whole statistical production process involved in their use and apply the total survey error framework to assessing the quality of the resultant official statistics.

Finally, given the scope of this paper, we have not included any measures of uncertainties with the above estimators. The interested reader should refer to the relevant papers included in the References for the details.

Disclaimer: The views expressed in this paper are those of the authors and do not necessarily represent the views of the Australian Bureau of Statistics nor University o Wollongong.

References

Australian Bureau of Statistics (2016). Information paper: Making Greater Use of Transactions Data to compile the Consumer Price Index, Australia. Catalogue Number 6401.0.60.003. ABS, Canberra.

Baker, R., Brick, J., Bate, N., Battaglia, M., Couper, M., Dever, J., Gile, K., and Tourangeau, R. (2013). Report of the AAPOR Task Force on non-probability surveys. https://www.aapor.org/Education-Resources/Reports/Non-Probability-Sampling.aspx. Accessed 3 November 2019.

Bethlehem, J. (2010). Selection bias in web surveys. *International Statistical Review* **78**, 161 -188.

Bethlehem, J. (2016). Solving the nonresponse problem with sample matching?, Social Science Computer Review, 34, 59-77.

Biemer, P. (2019) Can a Survey Sample of 6000 Records Produce More Accurate Estimates than an Administrative Data Base of 100 Million? *The Survey Statistician* **80**, 11-15.

Buelens, B., Burger, J. and van den Brakel, J. (2018). Comparing inference methods for non-probability samples. *International Statistical Review* **86**, 322-343.

Burger, J., Buelens, B., de Jong, T. and Gootzen, Y. (2019). Replacing a survey question by predictive modelling using register data. Paper presented to the 62nd World Statistics Congress, Kuala Lumpur.

Carfagna, E. and Gallego, F. (2006). Using remote sensing for agricultural statistics. *International Statistical Review* **73**, 389-404.

Chen, Y., Peng, L. and Wu, C. (2018). Doubly robust inference with non-probability survey samples. https://arxiv.org/abs/1805.06432. Accessed 7 October 2019.

Citro, C. (2014). From multiple modes for surveys to multiple data sources for estimates. Survey Methodology **40**, 137-161.

Couper, M. (2013). Is the sky falling? Survey Research Methods 7, p.145-156.

Daas, P. and Puts, M. (2014). Social media sentiment and consumer confidence. *European Central Bank Statistical Paper Series* **5**, 1-29.

Daas, P., Puts, M., Buelens, B. and van den Hurk, P. (2015). Big data as a source for official statistics. *Journal of official statistics* **31**, 249-262.

Dever, J., Rafferty, A. and Valliant, R. (2008). Internet surveys: can statistical adjustments eliminate coverage bias. *Survey Research Methods*, **2**, 47-62.

Dzhumasheva, S. (2019). Improving census occupancy determination – the potential of administrative for the Census. Paper presented to the 2019 ASEARC Workshop, Sydney.

Elliott, M. & Valliant, R. (2017). Inference for non-probability samples, Statistical Science, **32,** 249-264.

Gweon, H., Schonlau, M., Kaczmirek, L., Blohm, M. and Steiner, S. (2018). Three methods for occupation coding based on statistical learning. *Journal of Official Statistics* **33**, 101-122.

Fuller, W. (2009). *Sampling techniques.* John Wiley and Sons. Hoboken.

Hand, D. J. (2018). Statistical challenges of administrative and transaction data. *Journal of the Royal Statistical Society A 181*, 1–24.

Handbook on earth observations for official statistics (2017). Report prepared by the Satellite Imagery and Geo-spatial Statistics Task Team of the United Nations Global Working Group on Big Data. New York. https://unstats.un.org/bigdata/taskteams/satellite/. Accessed 22 October 2019.

Husek, N. (2018). Telematics data for official statistics: An experience with big data. *Statistical Journal of the International Association for Official Statistics*, 34, 499-504.

Ivancic, I., Diewert, D. and Fox, K. (2011). Scanner data, time aggregation and the construction of price indexes. *Journal of Econometrics,* 161, 24-35.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. and Usher, A. (2015). Report of the AAPOR Task Force on big data.

Jentoft, S., and Zhang L-C., (2018) Two phase and double machine learning for data editing and imputation. UNECE Workshop on statistical data editing. Neuchatel 18-20 September. https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T4_Norway_ZHANG_Paper.pdf.

Kim, J.K. (2011). Parametric fractional imputation for missing data analysis, *Biometrika* **98**, 119-132.

Kim, J.K. (2018). Unpublished survey data integration lectures delivered to the Australian Bureau of Statistics.

Kim, J.K., Berg, E. & Park, T. (2016). Statistical matching using fractional imputation, *Survey Methodology* **42**, 19-40.

Kim, J.K. and Rao, J.N.K (2012). Combining data from two independent surveys: a model-assisted approach. Biometrika, **99**, 85-100.

Kim, J.K. and Tam, S-M. (2018). Data integration by combining big data and survey sample data for finite population inference. Submitted.

Kim, J.K. and Wang, Z. (2018). Sampling techniques for big data analysis. *International Statistical Review* **87**, 177-191.

Kim, J.K, Wang, Z., Zhu, Z. and Cruze, N. (2018). Combining surveys and non-survey big data for improved sub-area prediction using a multi-level model. Journal of Agricultural, Biological and Environmental Statistics **23**, 175-189.

Kohler, U, Kreuter, F. and Stuart, E. (2019). Nonprobability sampling and causal analysis. Annual Review of Statistics and Its Applications, **6**, 149-172.

Kott, P., and Chang, T. (2010). Using Calibration Weighting to Adjust for Nonignorable Unit Nonresponse, *Journal of the American Statistical Association* **97**, 1265-1275.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for voluntary panel web surveys. *Journal of Official Statistics* **22**, 329-349.

Lothian, J., Holmberg, A. and Seyb, A. (2019). An evolutionary schema for using "it-is-what-it-is" data in official statistics. Journal of Official Statistics **35**, 137-165.

Lohr, S. and Raghunathan, T. (2017). *Combining survey data with other data sources. Statistical Science* **32**, 293-312.

Macfeely, S. (2019). Big data and official statistics. In *Big Data Governance and Perspectives in Knowledge Management*. IGI Global

Meng, X. (2018). Statistical paradises and paradoxes in Big Data (I): Law of large populations, big data paradox, and the 2016 US Presidential Election. The Annuals of Applied Statistics, 12, p 685 -726.

Mercer, A., Kreuter, F, Keeter, S. and Stuart, E. (2017). Theory and practice in non probability surveys – parallels between casual inference and survey inference. Public Opinion Quarterly **81**, 250-279.

Merkouris, T. (2004). Combining independent regression estimators from multiple surveys. *Journal of the American Statistical Association*, **99,** 1131-1139.

Park, S., Kim, J.K. & Stukel, D. (2017). A measurement error model for survey data integration: combining information from two surveys, Metron **75**, 345-357.

Richman M. B., Trafalis T. B. & Adrianto I. (2009). Missing data imputation through machine learning algorithms. In *Artificial Intelligence Methods in the Environmental Sciences*. Springer.

Rivers, D. (2007). Sampling for web surveys. In *Proceeding of Section on Survey Research Methods*. American Statistical Association.

Ruiz, C., (2018) Improving Data Validation using Machine Learning. UNECE Workshop on statistical data editing. Neuchatel 18-20 September.
https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.44/2018/T4_Switzerland_RUIZ_Paper.pdf

Saar-Tsechansky M. & Provost F. (2007). Handling missing values when applying classification models. *Journal of Machine Learning Research* **8**, 1623-1657.

Schonlau, M, van Soest, and Kapteyn, A (2009). Selection bias in web surveys and the use of propensity scores. Sociological Methods and Research, **37**, 291-318.

Tam, S-M. (2015). A statistical framework for analyzing Big Data. The Survey Statistician 72, 36-51.

Tam, S-M. and Clarke, F. (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. International Statistical Review, **3**, 436-448.

Tam, S-M. and Van Halderen, G. (2019). The five V's, seven virtues and ten rules of big data engagement for official statistics. Submitted.

Tam, S-M and Kim, J.K. (2018). Big data ethnics and selection bias: an official statistician's perspective. Statistical Journal of the International Association of official statistics, **34**, 577-588.

Tam, S-M., Kim, J.K., Ang, L. and Pham, H. (2019) (in press). Mining the new oil for official statistics in Big Data Meets Survey Practice: A Collection of Innovative Methods, John Wiley and Sons, Hoboken.

Thompson, M. (2018). Combining Data from New and Traditional Sources in Population Surveys, International Statistical Review, **87**, 79-S89.

Wang, S. (2019). Using predictive response propensity scores with the random forests method to direct a responsive intensive follow up strategy. Paper presented to the 62nd World Statistics Congress, Kuala Lumpur.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. Canadian Journal of Statistics **32**, 15-26.

Yang, S. and Kim, J.K. (2017). Predictive mean matching imputation in survey sampling. https://arxiv.org/abs/1703.10256.

Yang, S. and Kim, J.K. (2018). Integration of survey data and big observational data for finite population inference using mass imputation. https://arxiv.org/abs/1807.02817. Accessed 5 November 2019.

Zhang, L. (2019a). On valid descriptive inference from non-probability sample, Statistical Theory and Related Fields, 3:2, 103-113, https://doi.org/10.1080/24754269.2019.1666241. Accessed 3 December 2019.

Zhang, L. (2019b). Proxy expenditure weights for Consumer Price Index: Audit sampling inference for big data statistics. https://arxiv.org/abs/1906.11208. Accessed 8 November 2019.