ERROR ANALYSIS FOR HYBRID ESTIMATES THAT USE BIG DATA

Dennis Trewin* (Melbourne, Australia, dennistrewin@grapevine.net.au) and Siu-Ming Tam (University of Wollongong, Canberra, Australia, stattam@gmail.com)

Abstract

Big data, including administrative data, is seen as a source of data for official statistics especially given the increasing difficulty of getting acceptable response rates in sample surveys. It might be used directly, perhaps with the use of models to adjust for shortcomings in big data. Hybrid estimates are one such example. To make decisions on how it might be used we need to understand the nature of the errors in the big data source. The paper provides an Error Framework for the analysis of errors in big data, from both an input and output perspective. The paper also describes hybrid estimates and the circumstances under which they will provide more accurate estimates than big data in isolation.

Keywords: Big Data, Hybrid Estimates, Error Framework

Introduction

High quality traditional probability surveys are becoming more difficult and expensive to conduct. Non-response is a growing problem especially among some segments of the population and therefore the risk of bias in estimates derived from sample surveys. At the same time, there is a plethora of data from other sources. These data can have their own quality problems eg important parts of the population may not be covered, the data generation process is unknown or the concepts measured do not align with the target measures. However, they do have advantages over probability samples such as a greater number of observations, lower data collection costs and no additional reporting burden. These types of data are often referred to as big data. For the purposes of this paper, we are including administrative data as part of big data.

How can big data be used in official statistics? Option 4 is the subject of this paper.

1. Directly unmodified to produce statistics (more common for the use of administrative data).
2. Directly unmodified as a benchmark for a base period with subsequently collected data and models used to update estimates.
3. Directly but transformed -modelled but still likely to need other data (eg derived from a probability sample) or strong assumptions to support the models.
4. Combined data sources (integrated data sources or hybrid estimates).

We know there are both strengths and weaknesses with both probability surveys and big data. They are different so it does raise the question of whether they can be used in combination to build on their respective strengths and to compensate for their different weaknesses. Estimates combining both data sources are referred to as hybrid estimates.

Methodology

Total Error Framework for Big Data

A starting point for the consideration of hybrid estimates is understanding the sources of error in both the probability survey and the big data source. There is a lot of literature on Total Survey Error that covers the former (see Biemer and Lyberg, 2003). However, error frameworks for big data are still in the development stages. Total Error is a similar concept to Total Survey Error but recognises that the term 'survey' is misleading in the case of big data. Furthermore, the sources of error will be different, requiring a different framework for specifying the errors.

There have been some recent approaches to describe Total Error for Big Data (see Amaya et al, 2020). We have chosen to adapt the work of Reid et al (2017). Their work was targeted at administrative data but can be easily extended to Big Data. An important initiative in their work is to describe errors separately for each of 3 phases of the production process – (1) single input data, (2) integrated data sources, and (3) output. We will use the same structure except that the integration aspects of phase 2 are considered as part of output data as it is an integral part of the output process. That is, there will be separate Error Frameworks for input data and output data.

In this relatively short paper, only high level detail is shown in Table 1. Note that Model/Estimation Error only refers to that modelling that is undertaken at the input stage. Modelling used for output appears in Table 2.

Table 1: Summary of Total Error Framework for a Single Source of Big Data

ERRORS OF MEASURMENT – Type of Error	**Explanation**
Validity Error	Where the available measures differ from the target concept. Similar to specification error in TSE.
Measurement Error	Where the actual measures contain errors eg by poor form design when capturing the big data
Processing Error	Where errors are made in processing eg ETL activities, editing, coding
Modelling Error	Where models are used to change inputs such as conversion to the target variables.
ERRORS OF REPRESENTATION – Type of Error	**Explanation**
Coverage Error	Includes under-coverage, over-coverage and duplication of units
Selection Error	Where the data obtained from the ETL process does not represent the population
Missing Data Error	Includes both missing units and missing data items

An Error Framework for Hybrid Estimates

We are defining hybrid data as two or more data sets that compliment each other and are to be combined in some way (considered below) to provide more robust estimates (ie hybrid estimates). For example, a sample survey could be used to compliment a big data source by addressing its weaknesses eg under-coverage of certain populations or providing the data to enable the adjustments for any validity and measurement errors. This does not need to be done for every time period. It may be sufficient to run a survey every now and then to provide a benchmark which can be updated using the Big Data source.

Table 2 describes a high level framework for the output from hybrid estimates based on a combination of survey data and Big Data. They may or may not be based on the linking of common units. If some of the units can be linked, it is likely to enable more accurate estimates. The framework focusses on the residual errors after the hybrid estimates have combined the survey and Big Data sources. The residual errors will contain both systematic (bias) and random components. It is important to understand both.

For hybrid estimates, the validity of the modelling is arguably the most important consideration. There are at least three questions that should be addressed when considering modelling. What is the validity of the assumptions inherent in the model? What is the validity of the models used, including those derived from machine learning applications? What is the size of the residual error after the application of the models in hybrid estimates?

Table 2: Summary of Total Error Framework for Hybrid Data

ERRORS OF MEASURMENT – Type of Error	**Explanation**
Validity Error	Where there is systematic or random error after any adjustments when the available measures differ from the target concept source
Measurement Error	Where there are systematic or random errors (eg by poor form design) that have not been mitigated by modelling or other means
Processing Error	Where errors are made in processing eg ETL activities for Big Data, editing and coding on both data sources
ERRORS OF REPRESENTATION – Type of Error	**Explanation**
Frame/Coverage Error	Includes net under-coverage, over-coverage and duplication of units after the adjustments made through hybrid estimates
Sample Error	Due to the use of a sample to support hybrid estimates
Non-response/Missing Data Error	Includes both unit and item non-response that are not mitigated during the estimation process
Linkage Error	Residual impact of errors that are made in the process of linking common units
Modelling/Estimation	These are the systematic and random errors that remain after modelling/estimation
Analytic Processing	Seasonal Adjustment is an important example

Having defined the relevant frameworks, how do we use the framework to help with the design of the statistical collection? The first step is to understand the errors in the data sources irrespective of whether they are survey or Big Data based. Wherever possible, this should be based on quantitative information but we appreciate this is not always possible. Informed guesses may be necessary. Independent and expert third party advice can also be especially useful. The aim here is to identify the most important sources of error, not every source of error. These are the errors that need to be considered in determining whether the Big Data can be used or not, the design of the data collections and the hybrid estimation process itself. The frameworks described in Tables 1 and the standard TSE framework for survey data are relevant to this analysis.

The focus should then be on understanding these errors and how they might be measured, controlled or mitigated. Hybrid estimates, that adjust for the identified errors, are an important mitigation measure that should also be considered including the likely size of any residual errors in the hybrid estimates. The next step is to design the hybrid estimates themselves. This will include consideration of residual errors and how they might be further mitigated if they are important. It may be necessary to commission some special studies to support this analysis. The framework described in Table 2 is relevant to the error analysis of the hybrid estimates.

Hybrid estimates are considered in more detail below. The methodology can be very complex and the methodological capability may not be available in all statistical offices and external assistance may be needed at least in the initial stages until the required expertise can be developed.

Result

How to determine whether to use Hybrid Estimates?

If one has a probability sample, A, with full or partial response, and a big data set with under-coverage errors, which estimate out of the three, from the partially responding sample, big data set or a combination of the two, would be preferred?

Kim and Tam (2021) showed that, for simple random sampling with negligible finite population correction and full response, a hybrid estimator of the finite population total is preferred as it has smaller MSE than both the big data estimator or the survey estimator. In addition, they showed that , the hybrid estimator has a smaller sampling variance than the estimator from the probability sample, A, if $(1-\frac{N_B}{N})S_C^2 < S_U^2$ where N_B, N, S_U and S_C are the size of the Big Data set, size of the finite population, and the standard deviation of the population, and of the population segment missed by the Big Data set respectively. The inequality is almost always true when N_B is large in comparison with N.

We now use their ideas to extend to the case when A is partially responding, ie A_R , and compare the MSE of the estimators from the partially responding sample, big data set and a combination of both data sets. To illustrate ideas, we consider the simple case off estimating the finite population proportion, $P_U = \frac{\sum_{i \in U} Y_i}{N}$, where $Y_i = 1 \text{ or } 0$, without resorting to the use of auxiliary information in the estimation. We also assume that there are no measurement errors in the big data.

For this situation, how to choose between estimators from a Big Data sample, B, $P_B = \frac{\sum_{i \in B} Y_i}{N_B}$, a partially responding probability sample, A_R, $P_{A_R} = \frac{\sum_{i \in A_R} Y_i}{n_{A_R}}$, or hybrid estimates based on B and A_R ?

Let r_B denote the "representivity ratio" of B, i.e. the ratio of the probability of Y_i's $=1$ to <u>be included</u> in B to the probability of Y_i's $=0$ to be included in B. Likewise, let r_R denote the "representivity ratio" of A_R, i.e. the ratio of the probability of the Y_i's $=1$ included in B to <u>respond</u> to the survey to the probability of Y_i's $=0$ included in B to respond in the same survey. Then standard calculations show:

$$MSE(P_B) = \left\{ \frac{(r_B - 1)P_U(1-P_U)}{1+(r_B-1)P_U} \right\}^2$$ which is just the bias squared;

and $$MSE(\hat{p}_{A_R}) \square \frac{P_U(1-P_U)}{n_{A_R}} \frac{\{r_R + n_{A_R}(r_R-1)^2 P_U(1-P_U)\}}{\{1+(r_R-1)P_U\}^2}$$ comprising a sampling variance term and a bias square term.

To construct the hybrid estimator, let $A_R \cap C$ denote the responding sample in the population segment omitted by B, and $\hat{p}_{A_R \cap C}$ be the computed proportion of $Y_i = 1$ in this responding sample. Then the hybrid estimator of P_U is given by $\hat{p}_H = W_B P_B + (1 - W_B)\hat{p}_{A_R \cap C} = W_B P_B + W_C \hat{p}_{A_R \cap C}$, where $W_B = \frac{N_B}{N_U}$. $\hat{p}_H$ is unbiased because of correction for under-coverage from $\hat{p}_{A \cap C_R}$. Its MSE is given by:
$MSE(\hat{p}_H) = \frac{W_C^2}{n_{A_R \cap C}} \frac{r_R P_C (1 - P_C)}{\left\{1 + (r_R - 1) P_C\right\}^2} + \frac{W_C^2 (r_R - 1)^2 P_C^2 (1 - P_C)^2}{\left\{1 + (r_R - 1) P_C\right\}^2}$, where P_C denotes the proportion of $Y_i = 1$ in the omitted population segment. Note that $P_C = \frac{1}{W_C}(P_U - W_B P_B)$.

Which estimator is better under what conditions?

The choice is determined by comparing their MSEs. The following provide sufficient conditions for one estimator to be better than another. It is easily seen that $MSE(P_B) \le MSE(\hat{p}_{A_R})$ provided that $\frac{|(r_B - 1)|}{1 + (r_B - 1) P_U} \le \frac{|(r_R - 1)|}{1 + (r_R - 1) P_U}$, ie the absolute bias of P_B is smaller than that of $\hat{p}_{A_R}$ as the latter also has a sampling variance in its MSE. Otherwise, we have to assess the sign of:

$$MSE(P_B) - MSE(\hat{p}_{A_R}) = \left[\frac{(r_B - 1)^2 P_U^2 (1 - P_U)^2}{\left\{1 + (r_B - 1) P_U\right\}^2} - \frac{(r_R - 1)^2 P_U^2 (1 - P_U)^2}{\left\{1 + (r_R - 1) P_U\right\}^2} \right] - \frac{r_R P_U (1 - P_U)}{n_{A_R} \left\{1 + (r_R - 1) P_U\right\}^2}$$

which is data dependent and can only resolved numerically.

Provided that $W_C = (1 - W_B)$ is sufficiently small such that $W_C^2 \approx 0,$ we have $MSE(\hat{p}_H) \le$ $MSE(P_B)$ or $MSE(\hat{p}_{A_R})$ This is because $\hat{p}_H$ is bias free, and has a sampling variance component of order W_C^2 , whereas the estimator from B or A_R suffers from both bias, and for the latter sampling variation as well.

How to determine r_B and r_R ?

Let $\theta_A = \frac{\sum_{i \in A} (1 - Y_i)}{\sum_{i \in A} Y_i}, \theta_B = \frac{\sum_{i \in B} (1 - Y_i)}{\sum_{i \in B} Y_i}$ and $\theta_{A_R} = \frac{\sum_{i \in A_R} (1 - Y_i)}{\sum_{i \in A_R} Y_i}$, then it can be shown using the Bayes Theorem (Tam and Kim, 2018) that: $\hat{r}_B = \frac{\theta_A}{\theta_B}$ and $\hat{r}_R = \frac{\theta_A}{\theta_{A_R}}$. As both θ_B and θ_{A_R} are observed, we only need to estimate θ_A which is given by $\hat{\theta}_A = \frac{\sum_{i \in A_R} (1 - Y_i) / \hat{\rho}_i}{\sum_{i \in A_R} Y_i / / \hat{\rho}_i}$, where ρ_i is the propensity of the ith sampling unit to respond to the survey, with ρ_i to be estimated using a logistic regression model. However, if one can assume $W_C^2 \approx 0,$ the hybrid estimator is always preferred, and there is no need to estimate $\hat{r}_B$ and $\hat{r}_R$.

Discussion and Conclusions

Hybrid estimates play an important role in potentially making greater use of Big Data. The first step should be to assess the error structure of the Big Data. Are the errors sufficiently unimportant for it to be used in isolation to produce official statistics perhaps with the type of adjustments described in the first three options outlined in Section 1.

Alternatively, does the Big Data need to be supplemented by a survey data source to adjust for the most important error sources? This is where hybrid estimates come into play. They might be used, for example, to adjust for coverage error. This would be a common reason. Another common reason would be to adjust for validity error where the data concept available in the big data source is different to the target concept. There is always the option of just using the survey data and not worrying about using the Big Data at all.

The paper provides error frameworks that can be used to make decisions between these alternatives. They help to identify the most important potential error sources so consideration can be given to how they might be best mitigated. In this analysis, it is important to understand both the systematic and random components of these errors. It may be necessary to commission some special studies to support this analysis.

This paper also provides a method for helping make these decisions under restrictive (but realistic) assumptions on representational error and negligible finite population corrections. It assumes the most important sources of error are non-response for surveys and lack of coverage in Big Data. Representativity ratios are the key statistic and paper defines them and explains how they can be estimated. Where it can be assumed that the big data "sampling" fraction is close to 1, the hybrid estimate always outperforms the big data estimator or the estimator from the partially responding sample, regardless of the representivity ratios.

References

Amaya.A., Biemer.P. and Kinyon. D (2020). *Total Error in a Big data World. Adapting the TSE Framework to Big Data.* Journal of Survey Statistics and Methodology.

Biemer.P. and Lyberg.L. (2003). *Introduction to Survey Quality,* John Wiley and sons.

Kim.J.K. and Tam.S.M. (2021). *Data integration by combining big data and survey sample data for finite population inference. International Statistical Review*. To appear

Reid.G., Zabala.F. and Holmberg. A. (2017). *"Extending TSE to Administrative Data: A Quality Framework and Case Studies from Stats NZ".* Journal of Official Statistics

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.