Mining Big Data for Finite Population Inferences

**Keywords:** Big Data, Data Integration, Measurement Errors, Non-response adjustment, Under-coverage Bias.

# 1. Introduction

In 2006, an UK mathematician, Clive Humby, pronounced that “Data is the new oil. It’s valuable, but if unrefined, it cannot really be used. It has to be changed into gas, plastic, chemicals, etc. to create a valuable entity that drives profitable activity; so must data be broken down, analysed for it to have value.” We agree. Like oil, how the data is created in the first instance, how it is “refined” for statistics production, will determine its public value. Indeed, in official statistics, data that is not representative of the population on which public and private decisions are made, or data that are susceptible to measurement errors, are of at best limited value. However, rather than ignoring this type of data, which is increasingly becoming available from the Internet of things, the challenges for official statisticians are how to harness them effectively. Challenges in using Big Data for official statistics are outlined in Tam and Clark (2015) and Pfeffermann (2015).

# 2. Methods

Assume that the variable of interest, Y, is a continuous variable and we are interested to estimate the finite population total,  Assume further that we have a data source B whose data is fully observed without measurement error; and also a probability sample A with no measurement error, nor unit non-response in the data. If we consider the target population to be divided into two strata, one comprising the data from data source B, and the other comprising the complement of B in the population, i.e. U\B, the population total, , will be just the sum of the totals from both strata, i.e.  Using the sample units in AU\B, the total for the stratum U\B can be estimated by , and the population total by a post-stratified estimator, as follows:





If the size of the simple random sample, , it can easily be shown that





Letand . We have  Thus by using post-stratification to integrate the data source B, we can effectively increase the sample size of A, by a factor of  If = 0.5, this factor is 2; and is 5 if = 0.8.

Define for  and the Regression Data Integration (RegDI) estimator of , where the weights,  are determined by minimising defined by:



subject to



(2) requires the weights, , to be close to the Horvitz-Thompson weights, and (3) requires the weights to be calibrated to the known benchmarks,  It can be shown by tedious algebra that  Using this property, we can extend the integration of data source B with the probability sample A, to cover the following situations:

1. Availability of auxiliary variables (either in just B, or U) to improve the efficiency of 
2. The data in B is subject to measurement errors;
3. The data in A is subject to measurement errors
4. A is subject to unit non-response.

The methods developed above generalise those in Tam and Kim (2018).

# 3. Results from Empirical Studies

In the simulation study, a continuous *Y* variable is generated from the following model:

**

where  and  is independent of **. We generate a finite population of size *N* = 1,000,000 from this model.

In this simulation, we repeatedly obtain two samples, denoted by A and B, by simple random sampling of size *n* = 500 and by an unequal probability sampling of size = 500,000, respectively. In selecting sample B, we create two strata, where stratum 1 consists of elements with , and stratum 2 consists of those with *>* 2. Within each stratum, we select  elements by simple random sampling independently, where ** = 300,000 and  = 200,000.

Table: Results based on a Monte Carlo sample of size 1 000

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Scenario | Estimator | Bias | Standardised Variance | Standardised MSE |
| I | Sample Mean A  Sample Mean B  P  RegDI | 0.00  -0.11  0.00  0.00 | 100  0  47  47 | 100  1,116  47  47 |
| II | Sample Mean A  Sample Mean B  P  RegDI | 0.00  -1.10  -0.49  0.00 | 100  0  47  62 | 100  63,842  12,895  58 |
| III | Sample Mean A  Sample Mean B  P  RegDI | -1.00  -0.11  -0.50  0.00 | 108  0  51  83 | 52,831  668  13,521  79 |

# 4. Conclusions

In this paper, we have shown that Big Data sources subject to under-coverage bias can be used as benchmarks to calibrate the Horvitz-Thompson weights of a random sample to provide efficient estimation of the finite population total. Using the methods in Section 2, we showed the sample size,of a simple random sample, can be increased by a factor of  which provides an opportunity for official statisticians to reduce the cost of data collection by, if desired, reducing the sample size judiciously.

Noting that the post-stratified estimator is equivalent to a calibration estimator with the Horvitz-Thompson weights of the probability sample calibrating to the known benchmarks, , we can apply the calibration idea to address other biases in the data sources.

# References

[1] S. M. Tam and F. Clarke (2015). Big data, official statistics and some initiatives by the Australian Bureau of Statistics. *Internat. Statist. Rev. 83*, 436–448.

[2] D. Pfefffermann (2015). Methodological issues and challenges in the production of official statistics: 24th Annual Morris Hansen Lecture. *J. Surv. Stat. Methodol. 3*, 425–483.

[3] S. M. Tam, S. M. and J. K. Kim, (to appear). Big data, selection bias and ethics – an official statistician’s perspective. *Stat. J. IAOS*.