Mining Big Data for Finite Population Inference*

*Joint work with Professor Jae Kim of Iowa State University

Dr Siu-Ming Tam, Chief Methodologist

Australian Bureau of Statistics

Honorary Professorial Fellow

University of Wollongong

14 March, 2019

Australian Bureau of Statistics
Informing Australia's important decisions

Outline of talk

- Debunking a Big Data myth
- The set up for Big Data inference, and the theory
- An application in official statistics
- Concluding remarks

Big does not necessarily mean it is good

- Let B be the Big Data set
- Let $\delta_i = 1$ if $i \in B$, and 0 otherwise
- Let

$\bar{y}_B$ be the sample mean of y in B. The MSE of $\bar{y}_B$ as an estimator of the population mean $\bar{Y} = \sum_{i \in U} y_i / N$ is (Meng, 2018)

$$E\{(\bar{y}_B - \bar{Y})^2\} = E_\delta\{Corr(Y,\delta)^2\}(f^{-1}-1)\sigma_Y^2$$

where $f = E(\delta)$ = sampling rate for B.

- Meng calls it the Fundamental Identity of Estimation Error
- If Y is binary,

Let $p = \bar{Y} = \Pr(Y=1)$. Let $b = \Pr(\delta = 1 | Y = 1) - \Pr(\delta = 1 | Y = 0) > 0$. Then

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

for large N, where n_{eff} is the effective sample size of B.

Inferential value of Big Data sets

Australian Bureau of Statistics

Effective sample size for estimating the proportion of Australians speaking English at home in the 2016 Census

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

(Tam and Kim, 2018a)

- How best to use the Big Data set?
  - We rely on the use of a probability sample, A

Data structure

Scenario 1 – E.g. On line panels

Data	X	Y	Representivity
Probability sample, A	✓		Yes
Big Data, B	✓	✓	No

Scenario 2 – E.g. Satellite imagery data, social media, search engine terms

Data	X	Y	Representivity
Probability sample, A	✓	✓	Yes
B	✓		No

Scenario 3 – Special case of Scenario 1(to avoid making MAR assumptions)

Data	X	Y	Representivity
Probability sample, A	✓	✓	Yes
B	✓	✓	No

- For the rest of the talk, we shall consider only Scenario 3

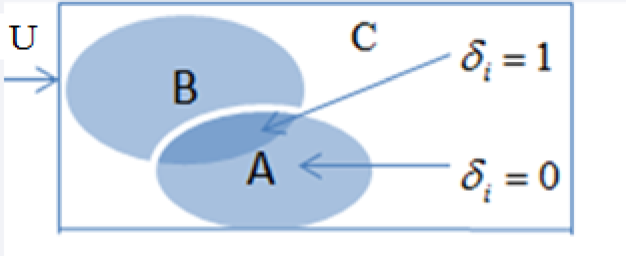
The set up – the ABC of Big Data (Tam and Kim, 2018b)

- Finite population: $U = \{1, \cdots, N\}$.
- Parameter of interest: $\bar{Y}_N = N^{-1} \sum_{i=1}^{N} y_i$ (Or equivalently: $\theta = \sum_{i \in U} y_i$)
- Big data sample: $B \subset U$.

$$\delta_i = \begin{cases} 1 & \text{if } i \in B \\ 0 & \text{otherwise.} \end{cases}$$

- Estimator: $\bar{y}_B = N_B^{-1} \sum_{i=1}^{N} \delta_i y_i$, where $N_B = \sum_{i=1}^{N} \delta_i$ is the big data sample size ($N_B < N$).

- Assume we have a random sample of U, denoted by A; N and N_B are also assumed known.



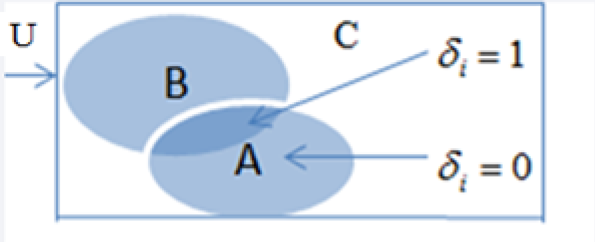
The key idea

- From $\theta = \sum_{i \in B} y_i + \sum_{i \in C} y_i$, $\hat{\theta} = \sum_{i \in U} \delta_i y_i + N_C \dfrac{\sum_{i \in A} w_i(1-\delta_i) y_i}{\sum_{i \in A} w_i(1-\delta_i)}$

  where the second component is provided by the random sample, A, of size n
- Significant improvement in efficiency in $\hat{\theta}$ due to B – e.g. for SRS, the effective sample size of n will be increased by a factor

  $= \dfrac{S^2}{S_C^2} \dfrac{1}{(1 - N_B / N)}$;
- Post-stratified estimator, $\hat{\theta}_A$, can be shown to be equivalent to a calibration estimator, $\hat{\theta}_A = \sum_{i \in A} w_i^* y_i$ where w_i^* minimises the Chi squared distance $D(w, w^*) = \sum_{i \in A} w_i (\frac{w_i^*}{w_i} - 1)^2$ subject to

  $\sum_{i \in A} w_i^* (1-\delta_i, \delta_i, \delta_i y_i, \delta_i x_i) = (N_C, N_B, \sum_{i \in B} y_i, \sum_{i \in B} x_i)$, where w_i is the HT weight



Addressing measurement errors in Big Data

Australian Bureau of Statistics

Extension 1 - Measurement error in sample B

Data Structure

Data	X	Y^*	Y	Represent?
A	✓		✓	Yes
B	✓	✓		No

Y^*: proxy variable for Y, $E(Y^*) \neq E(Y)$

Parameter of interest: $\theta = \sum_{i \in U} y_i$

Use $\hat{\theta}_A = \sum_{i \in A} w_i^* y_i$ where w_i^*'s minimises $D(w, w^*)$

subject to

$$\sum_{i \in A} w_i^* (1 - \delta_i, \delta_i, \delta_i x_i, \delta_i y_i^*) = \sum_{i \in U} (1 - \delta_i, \delta_i, \delta_i x_i, \delta_i y_i^*)$$

Addressing measurement errors in integrating sample

Australian Bureau of Statistics

Extension 2 - Measurement error in sample A

Data Structure

Data	X	Y^*	Y	Represent?
A	✓	✓		Yes
B	✓		✓	No

Use $\hat{\theta}_A = \sum_{i \in A} w_i^* \hat{y}_i$ where w_i^*'s minimises $D(w, w^*)$ subject to

$$\sum_{i \in A} w_i^* (1-\delta_i, \delta_i, \delta_i x_i, \delta_i y_i) = \sum_{i \in U} (1-\delta_i, \delta_i, \delta_i x_i, \delta_i y_i),$$ where $\hat{y}_i$ is imputed using a measurement error model for $i \in A$.

Extension 3 - Handling Unit Nonresponse in sample A

Australian Bureau of Statistics

Data Structure

Data	X	Y^*	Y
A_R	✓		✓
A_M	✓		
B	✓	✓	

$A = A_R \cup A_M$

Y is not observed in A_M

Use $\hat{\theta}_A = \sum_{i\in A} r_i w_i^* y_i / \hat{\pi}_i$ where $r_i = 0,1; \hat{\pi}_i =$ response propensity and

w_i^*'s minimises $D(w_i\hat{\pi}_i^{-1}, w^*) = \sum_{i\in A} r_i w_i \hat{\pi}_i^{-1}\left(\frac{w_i^*}{w_i\hat{\pi}_i^{-1}} - 1\right)^2$ subject to

$$\sum_{i\in A} r_i w_i^*(1-\delta_i, \delta_i, \delta_i x_i, \delta_i y_i^*) = \sum_{i\in U}(1-\delta_i, \delta_i, \delta_i x_i, \delta_i y_i^*)$$

An ABS example

Australian Bureau of Statistics

- Two data sources
  1. ABS (Australian Bureau of Statistics) 2015-16 Agricultural Census: 85% response rate
  2. REACS (Rural Environment and Agricultural Commodities Survey) data (2014-15), sample size $\cong$ 34K.
- Observation
  1. y_i: study variable for year 2015-16
  2. $\tilde{y}_i$: study variable for year 2014-15
- $\delta_i = 1$ if participated at Census and $\delta_i = 0$ otherwise.

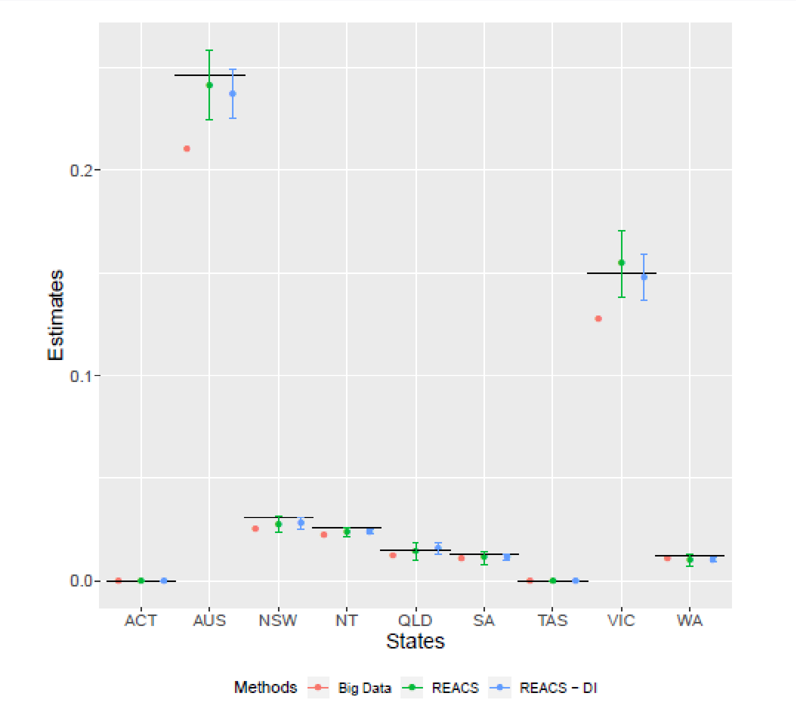
Assume no measurement error in REACS case- 8 and 13 fold improvement in efficiency

Table: Bias, Variance and Mean Squared Error of Selected Agricultural Commodities

Variable	Estimator from	Bias (x10^3)	Var (x10^9)**	MSE (x10^9)
DAIRY	REACS only (A)	0.00	6.19	6.19
	Agricultural Census only (B)*	-362.45	0	131.37
	(A) and (B)	0.00	0.43	0.43
BEEF	REACS only (A)	0.00	85.00	85.00
	Agricultural Census only (B)*	-2,389.53	0	5,709.86
	(A) and (B)	0.00	6.79	6.79
WHEAT	REACS only (A)	0.00	171.29	171.29
	Agricultural Census only (B)*	-2,043.52	0	4,176.00
	(A) and (B)	0.00	20.83	20.83

* Estimated by the difference between the total from B and the published ABS estimate from the Agriculture Census adjusted for non-response.

Only 1.5 fold increase in efficiency with measurement errors in REACS – DAIRY cattle results



Concluding comments

- Random samples are here to stay in the Big Data world
  - Unless there are defensible ways to adjust for Big Data biases
- We have not discussed variance estimation
  - but the methods will be published elsewhere
  - In the ABS example, we use bootstrap samples to estimate uncertainty

References

- Meng, X.L. (2018). Statistical paradises and paradoxes in big data (I): Law of large population, big data paradox, and he 2016 US presidential election. The Annals of Applied Statistics 12(2), 685-726
- Tam, S.M and Kim, J.K. (2018a). Big Data ethics and selection bias: An official statistician's perspective. Statistical Journal of the International Association for Official Statistics 34(4), 577-588.
- Tam, S.M and Kim, J.K. (2018b). Mining the new oil for official statistics. Conference paper presented to BigSurv18, Barcelona.

Questions?

Siu-Ming.Tam@abs.gov.au