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
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

$$E(\bar{y}_x - \bar{y}) = \text{Var}(\bar{y}_x) = \frac{1}{N} \text{Var}(y_i)$$

where $\text{Var}(y_i)$ is the variance of $y_i$ in $B$.

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

$$n_{eff} = \frac{f^2}{\hat{\beta}^2 \hat{\pi}(1 - \hat{\pi})}$$

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

(\text{Meng, 2018})
How best to use the Big Data set?

- We rely on the use of a probability sample, A
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: \( U = \{1, \ldots, N\} \).
Parameter of interest: \( \hat{Y}_N = N^{-1} \sum_{i=1}^{N} y_i \) (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: \( \hat{y}_B = N_B^{-1} \sum_{i=1}^{N} \delta_i y_i \), 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.

Diagram:
- \( U \)
- \( B \)
- \( C \)
- \( A \)
- \( \delta_i = 1 \)
- \( \delta_i = 0 \)
The key idea

- From \( \theta = \sum_{x \in \mathcal{A}} y_i + \sum_{x \in \mathcal{B}} y_i \), \( \hat{\theta} = \sum_{x \in \mathcal{A}} \delta y_i + N_x \sum_{x \in \mathcal{B}} \frac{w_i(1 - \delta_i)y_i}{\sum_{x \in \mathcal{A}} w_i(1 - \delta_i)} \)

  where the second component is provided by the random sample, \( \mathcal{A} \), of size \( n \).

- Significant improvement in efficiency in \( \hat{\theta} \) due to \( \mathcal{B} \) - e.g. for SRS, the effective sample size of \( n \) will be increased by a factor

  \[
  \frac{\frac{S^2}{\hat{\sigma}^2}}{\left(1 - \frac{N_x}{N}\right)^2} ;
  \]

- Post-stratified estimator, \( \hat{\theta}_A \), can be shown to be equivalent to a calibration estimator, \( \hat{\theta}_d = \sum_{x \in \mathcal{A}} w_i^* y_i \) where \( w_i^* \) minimises the Chi squared distance

  \[
  D(w, w^*) = \sum_{x \in \mathcal{A}} w_i (\frac{w_i^*}{w_i} - 1)^2 \text{ subject to } \sum_{i \in A} w_i(1 - \delta_i, \delta_i, \delta y_i, \delta x_i) = (N_x, N_y, \sum x_i, \sum y_i) , \]

  where \( w_i \) is the HT weight.
Addressing measurement errors in Big Data

Extension 1 - Measurement error in sample B

Data Structure

<table>
<thead>
<tr>
<th>Data</th>
<th>X</th>
<th>Y*</th>
<th>Y</th>
<th>Represent?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>B</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>No</td>
</tr>
</tbody>
</table>

$Y^*$: 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

\[
\sum_{i \in A} w_i^*(1 - \delta_i, \delta_i, \delta_i x_i, \delta_i y_i^*) = \sum_{i \in A} (1 - \delta_i, \delta_i, \delta_i x_i, \delta_i y_i^*)
\]
Addressing measurement errors in integrating sample

Extension 2 - Measurement error in sample A

Data Structure

<table>
<thead>
<tr>
<th>Data</th>
<th>X</th>
<th>Y*</th>
<th>Y</th>
<th>Represent?</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>Yes</td>
</tr>
<tr>
<td>B</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>No</td>
</tr>
</tbody>
</table>

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 \( i \in A \).
Addressing non-response in random sample A

Data Structure

<table>
<thead>
<tr>
<th>Data</th>
<th>X</th>
<th>Y*</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>$A_R$</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>$A_M$</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>B</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

$A = A_R \cup A_M$

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

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

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 $\approx 34K$.

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>
<thead>
<tr>
<th>Variable</th>
<th>Estimator from</th>
<th>Bias ($\times 10^3$)</th>
<th>Var ($\times 10^5$)**</th>
<th>MSE ($\times 10^5$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAIRY</td>
<td>REACS only (A)</td>
<td>0.00</td>
<td>6.19</td>
<td>6.19</td>
</tr>
<tr>
<td></td>
<td>Agricultural</td>
<td>-362.45</td>
<td>0</td>
<td>131.37</td>
</tr>
<tr>
<td></td>
<td>Census only (B)*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(A) and (B)</td>
<td>0.00</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>BEEF</td>
<td>REACS only (A)</td>
<td>0.00</td>
<td>85.00</td>
<td>85.00</td>
</tr>
<tr>
<td></td>
<td>Agricultural</td>
<td>-2,389.53</td>
<td>0</td>
<td>5,709.86</td>
</tr>
<tr>
<td></td>
<td>Census only (B)*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(A) and (B)</td>
<td>0.00</td>
<td>6.79</td>
<td>6.79</td>
</tr>
<tr>
<td>WHEAT</td>
<td>REACS only (A)</td>
<td>0.00</td>
<td>171.29</td>
<td>171.29</td>
</tr>
<tr>
<td></td>
<td>Agricultural</td>
<td>-2,043.52</td>
<td>0</td>
<td>4,176.00</td>
</tr>
<tr>
<td></td>
<td>Census only (B)*</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>(A) and (B)</td>
<td>0.00</td>
<td>20.83</td>
<td>20.83</td>
</tr>
</tbody>
</table>

* 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
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


Questions?
Siu-Ming.Tam@abs.gov.au