On data integrators for official statistics

By DR Siu-Ming Tam
Honorary Professorial Fellow, University of Wollongong
Ex Chief Methodologist, Australian Bureau of Statistics
Principal, Tam Data Advisory
Presented to
63rd Word Statistics Congress on 13 July, 2021
Outline

- Why and what of data integration
- Busting a Big Data myth
- 4 types of data integrators
- Data access issues
- 10 rules of engagement with big data
- Conclusion
Data integration - Why?

- Direct data collection is an expensive and increasingly unsustainable business model for NSOs.

- Declining budgets, increasing demands for more frequent and richer statistics, declining response rates etc. are strong drivers for NSOs to look at reusing/repurposing existing data sets e.g. adm data, Big Data etc.

- The public value of integrated data sets will be significantly higher than each of its component data sets, e.g. combining census data with migration data to assess how different migration cohorts settled in Australia.
What is Data integration?

- Two types data integration (DI) can be recognised

- Type I data integration (DI) - Micro level integration
  - defined by the UNECE as “the activity when at least two different sources of data are combined into a dataset. This dataset can be one that already exists in the statistical system or ones that are external sources”

- Type II DI - Macro level integration
  - defined in the statistical literature as a method to borrow strength from a non-random data set, say B, to improve the statistical value of estimates from a random sample, say A.
    - This definition includes the use of training data sets to “train” machine learning algorithms
    - We are particularly interested to integrate administrative data, or more generally Big Data, with survey data
What is Data integration? (cont’d)

- Type I DI method uses a methodology developed by Statistics Canada under in 1969 for probabilistic matching
  - This is still the methodology of choice today
  - The challenge is when the linkage variables are not statistically independent or have measurement errors
  - These are subjects of another talk and will not be covered today

- In this talk, I share some prevailing methodologies for data integration
  - Try to explain the underlying ideas using English rather than maths as much as possible, to suit a broad audience
Big does not necessarily mean it is good

- Suppose the population $U$ comprises 5 million males, and 5 million females; and the Big Data comprises 5 million males and 4 million females.

- The proportion of males in the population = 50%

- The Big Data (“sample” fraction of 90%) estimate = $5m/(5m+4m) = 55.56\%$
  - This has a bias of 5.56\% even though it has 90\% coverage.

- Why is there a bias in the Big Data estimate?
  - Because the propensity of males (100\%) included in the Big Data is NOT equal to the propensity of females (80\%) included in $B$.
    - This problem is often referred to as the under-coverage or self-selection bias of big data.
  - Another problem of big data is measurement errors in the variables of interest.
    - This includes also errors from the variable not measuring a concept of interest to the official statistician.
    - This causes significant problems for Type 1 DI but not so for Type II DI.
Inverse propensity weighted (IPW) estimator (Type 1 scenario)

- The previous example provides an insight into Big Data under-coverage adjustment
  - If we knew that the propensity for males is 100%, and female is 80%, then the estimated proportion of males would be $\frac{5m}{1/(5m + 1/0.8)} = 5/10 = 50$

- More generally, if the propensity, $\hat{\rho}_i$, could be estimated for each unit in the Big Data, B, then under a Quasi-Randomisation (QR) framework, the IPW estimator, $\hat{p}_{IPW} = \frac{\sum_{i \in B} y_i}{\sum_{i \in B} \frac{1}{\hat{\rho}_i}}$, is approximately unbiased and consistent.
For the estimator to work, we need to assume

- Every unit in the population has a positive propensity to be included in B (Positivity Assumption)

- The propensity can be estimated using auxiliary variables or covariates (MAR assumption)

- The propensity can be approximated by a (e.g. logistic regression) model, with the parameters, $\eta$, consistently estimated by an estimation equation
  - e.g. for the logistic regression model, the estimating equation is:
    - $\sum_{i \in B} x_i - \sum_{i \in U} \rho_i(\eta)x_i = 0$ to solve for $\eta$
Data integration - 4 data structure scenarios

Type 1 scenario has already been covered in previous slides
Main difference of others is the presence/absence of the response variable

<table>
<thead>
<tr>
<th></th>
<th>Source</th>
<th>Response variable, Y</th>
<th>Auxiliary variables, X</th>
<th>Sampling weight, d</th>
<th>Representative?</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Big Data, B</td>
<td>A</td>
<td>A</td>
<td>NA</td>
<td>No</td>
</tr>
<tr>
<td>Type I</td>
<td>Prob. sample, A</td>
<td>NA</td>
<td>A</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>Type II</td>
<td>Prob. sample, A</td>
<td>NA</td>
<td>A</td>
<td>A</td>
<td>Yes</td>
</tr>
<tr>
<td>Type III</td>
<td>Big Data, B</td>
<td>A</td>
<td>A</td>
<td>NA</td>
<td>No</td>
</tr>
<tr>
<td>Type IV</td>
<td>Prob. sample, A</td>
<td>A</td>
<td>A</td>
<td>NA</td>
<td>No</td>
</tr>
<tr>
<td>Type IV</td>
<td>Prob. sample, A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>Yes</td>
</tr>
</tbody>
</table>

Legend: A = Available; NA = Not available

- Type 1 example - On-line panels to estimate LF participation
- Type 2 example - COVID-19 tests data + NSO LF survey to estimate COVID-19 infections
- Type 3 example - Admin data/on-line panels + NSO LF survey to estimate LF participation
- Type 4 example - Satellite imagery + NSO agricultural survey to estimate crop yields
Type II scenario - Imputation estimators

- Two classes of estimators have been advocated in the literature
  - They are both of the form, $\sum_{i \in A} d_i \hat{y}_i$, to estimate the population total
  - First method - $\hat{y}_i$ imputed using regression prediction, with model built using B
  - Second method - $\hat{y}_i$ imputed using the nearest neighbour (NN) to be found in B

- Assumptions
  - MAR - No difference between regression model between the big data and population
  - Positivity assumption as discussed before
Type II scenario - Doubly Robust (DR) estimator

- DR estimator is constructed by the sum of the regression-imputation estimator and an IPW estimator of the regression residual:
  \[ \sum_{i \in A} d_i x_i^T \hat{\beta} + \sum_{i \in B} \{ y_i - x_i^T \hat{\beta} \} / \hat{\rho}_i \]

- Doubly robust in the sense that the estimator is still unbiased if the assumed propensity model, or the regression model, is incorrect

  - If the propensity model is correctly specified, the first and the third terms cancel out under expectation, and the second term estimates QR unbiasedly the population total
  
  - If the superpopulation model is correctly specified, then the model expectation of the second and third terms cancel out, and leaving the first term which is the (design and model) expectation of the population total.
The idea is to use the big data as benchmarks to improve the usual survey estimator

- The benchmarks are the counts of unit in B, and the sums of the covariates in U (or B, if the covariates are not fully observed)
- This benchmarking process will result in a change of the weight from $d$ to $w$, using the calibration process
  - This process is the same to Generalised Regression (GREG) estimation commonly used in NSOs

Assumptions

- Correctness of the superpopulation model underpinning GREG
- However, some researchers consider calibration as a new estimation philosophy and will not consider the above assumption necessary
Type III scenario - Post-stratified (PS) estimators

- Total in U = total in B (fully observed) + estimated total in C, using the random sample in A∩C
  - In other word, use the random sample to adjust for under coverage bias of B
  - It is called a PS estimator because we consider the population comprises two stratum
  - Technically, the estimator is expressed as

  \[ \sum_{i \in B} y_i + \frac{N_C}{\sum_{i \in A \cap C} d_i} \sum_{i \in A \cap C} d_i y_i \]

  - Algebraically, the estimator is the same as a calibration estimator with \((N_C, N_B, \sum_{i \in B} y_i)\) as benchmarks. The estimator is also called Regression Data integrator (RDI)
Extensions of PS estimators

- The insight to express the PS estimator as a calibration estimator gives us ways to address 5 different types of challenges in the data:

  - **A - Availability of covariates information**
    - Add them in the benchmarks
    - \((N_C, N_B, \sum_{i \in B} y_i, \sum_{i \in U} x_i)\) or \((N_C, N_B, \sum_{i \in B} y_i, \sum_{i \in B} x_i)\) if \(x_i\) is fully/not fully observed

  - **B - Multiple appearances of the same unit in B**
    - Incorporate the multiple count in the calibration process
Extensions of PS estimators

- C - Measurement errors in B or A i.e. $y_i^*$ is observed instead of $y_i$

  - C1 - Measurement error in B case
    - Replace $y_i$ by $y_i^*$ in the $\sum_{i \in B} y_i$ benchmark

  - C2 - Measurement error in A case -
    - Use $\hat{y}_i$ instead of $y_i$ in $\sum_{i \in A} w_i y_i$, where
      - $\hat{y}_i$ = Regression-predicted value of $y_i$ using a measurement error model
Extensions of PS estimators

- **D - Non response in A**
  - Use a new calibration weight = old calibration weight * inverse of response propensity

- **E - Linkage error between B and A**
  - Linking error results in not observing $\delta_i$ in A, i.e. whether unit i in A is included in B or not, correctly
  - In this case:
    - Use the EM algorithm to estimate $\delta_i$ and adjusted benchmarks for the calibration process
Performance of the RDI - an ABS example

- B = 2015-16 Ag Census, with response rate of 85%
- A = 2014-15 Ag Survey with response rate of 78%

Issues
- Under-coverage in B
- Non response error in A

In spite of the above issues, 2014-15 RDI estimates are still about 8 to 12 fold more efficient than the estimator from A

### Bias, Variance and Mean Squared Error of Selected Agricultural Commodities at the Australian level

<table>
<thead>
<tr>
<th>Variable</th>
<th>Estimator from</th>
<th>Bias ($\times 10^2$)</th>
<th>Var ($\times 10^6$)**</th>
<th>MSE ($\times 10^6$)</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 Census only (B)*</td>
<td>-362.45</td>
<td>0</td>
<td>131.37</td>
</tr>
<tr>
<td></td>
<td>RDI using (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 Census only (B)*</td>
<td>-2,389.53</td>
<td>0</td>
<td>5,709.86</td>
</tr>
<tr>
<td></td>
<td>RDI using (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 Census only (B)*</td>
<td>-2,043.52</td>
<td>0</td>
<td>4,176.00</td>
</tr>
<tr>
<td></td>
<td>RDI using (A) and (B)</td>
<td>0.00</td>
<td>20.83</td>
<td>20.83</td>
</tr>
</tbody>
</table>

Notes:
(1) * Estimated by the difference between the total from B and the published ABS estimate from the Agriculture Census adjusted for non-response.
(2) ** Bootstrap estimates from 100 bootstrap samples.
Accessing the big data

- Cannot be taken for granted and requires sustained investments in building rapport and trust with data custodians

  - In Australia, years invested in cultivating a positive, productive and trusting relationship allow ABS to access a number of unidentified and identified data sets from Govt agencies

  - The Australian Govt is considering to pass legislation for Govt agencies to opt in to share data between them, subject to the necessary confidentiality controls

- ABS is also able to gather scanner data from supermarkets for compiling its CPI
  - Success relied again on building trust
    - Found that if one company took the lead to provide the data, others followed
The ten rules of engagement with Big Data

- Previous slides dealt with points 2, 6, 7 and 10.
- The other points will also need to be considered, when deliberating on the merits of integrating Big Data with survey data.

<table>
<thead>
<tr>
<th>Non-negotiable</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Use big data as a solution to a well-defined statistical need;</td>
</tr>
<tr>
<td>2. The long term supply of the big data source should be certain;</td>
</tr>
<tr>
<td>3. Social license issues must be addressed;</td>
</tr>
<tr>
<td>4. The big data is impartial;</td>
</tr>
<tr>
<td>5. Security and confidentiality issues have been addressed;</td>
</tr>
<tr>
<td>6. The big data is a cost effective alternative or supplement to traditional</td>
</tr>
<tr>
<td>statistical data sources; and</td>
</tr>
<tr>
<td>7. Statistics are amenable to valid statistical inferences</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Essential</th>
</tr>
</thead>
<tbody>
<tr>
<td>8. The use of big data reduces provider load;</td>
</tr>
<tr>
<td>9. The use of big data produces better statistics; or</td>
</tr>
<tr>
<td>10. The use of big data is a fail safe</td>
</tr>
</tbody>
</table>
Concluding remarks

- Data integration is the future of official statistics because integrated data sets significantly increase the public value of the data.

- Need to be mindful that unadjusted big data estimators will provide misleading inferences, which could significantly affect policy formulation and evaluation.

- Different data structure scenarios require different bias adjusted estimators:
  - They are either IPW estimators, imputation (NN or regression) estimators, calibration estimators or regression data integrators (RDI).
  - Note that to date only RDI has provided a complete package to address the multiplicity of big data and survey data issues.
  - Need to check fulfilment of underlying assumptions and undertake sensitivity analysis.
Concluding remarks (cont’d)

- Getting access to data is a challenge
  - For official statistics, my view is that priority for access should be given to administrative data held by other govt agencies

- Although no mention has been made about variances and their estimates, exact formula for the different estimators can be found in the literature
  - They are quite complex to implement
  - Bootstrap variance estimates can be used as viable alternatives

- Most important of all, one should assess the risk-benefits of integrating big data with survey data
  - The ten rules of big data engagement provides a useful framework for assessment
References

Questions? Stattam@gmail.com