THE ABC OF BIG DATA

DATA INTEGRATION WORKSHOP – 26 NOV AND 3 DEC 2020

DR SIU-MING TAM

EX- CHIEF METHODOLOGIST AND GENERAL MANAGER, AUSTRALIAN BUREAU OF STATISTICS

HONORARY PROFESSORIAL FELLOW, UNIVERSITY OF WOLLONGONG

PRINCIPAL, TAM DATA ADVISORY

2 OUTLINE OF TALK

- Recap of main points from my talk on 16 Sept
- The ABC method of Big Data
- Extension of the ABC method to more complex situations
- Concluding remarks

TWO TYPES OF DATA INTEGRATION

- 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 (e.g. administrative dataset acquired from an owner of administrative registers or web-scraped information from a publicly available website)"
- Type II DI – Macro level integration
  - defined in the statistical literature as borrowing strength from a non-random data set, say B, to improve the statistical value of estimates from a random sample, say A.

This talk is about Type II integration

RECAP - INFERENTIAL VALUE OF BIG DATA SETS REDUCED BY THE EXTENT OF RESPONSE BIAS (MENG, 2018)

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

		Response bias, b		
Big Data fraction, f	Big Data size	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)

RECAP - 8 AND 13 FOLD IMPROVEMENT IN EFFICIENCY OF SURVEY ESTIMATES, USING ABS AG CENSUS AS BIG DATA (TAM ET AL, 2018B)

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.

6 TYPES OF DATA STRUCTURE – INGREDIENTS TO ADDRESS BIG DATA BIAS (TAM AND HOLMBERG, 2020)

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

We will cover scenario 3 in this talk – in essence, the method uses Big Data as benchmarks

7 THE ABC METHOD (AKA REGRESSION DATA INTEGRATION) IN A NUTSHELL - (KIM AND TAM, 2020)

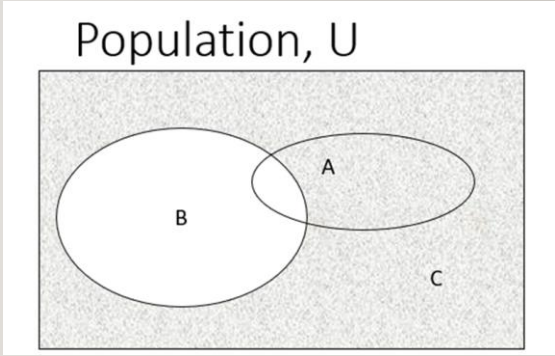
- Change the initial weight from d to w so that:

  Step A (Optimisation step)
  - The new weights minimise the Chi-square distance between the new weights and initial weights
    - This will ensure closeness between the weights; and
    - Asymptotic design unbiasedness

  Step B (Calibration step)
  - The new weights satisfy certain calibration conditions
    - To ensure that the weighted sums of known variables/counts match the benchmarks

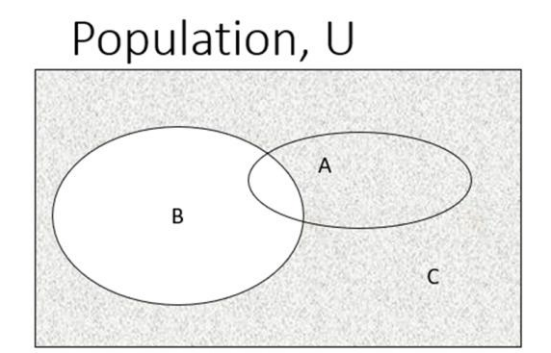
8 THE SET UP

- Finite population: $U = \{1, \ldots, N\}; F = \{y_1, \ldots, y_N\}$
- Parameter of interest: $Y = \sum_{i \in U} y_i$
- Big Data sample, $B \subset U$, of size n_B. Big Data sample mean: $\bar{Y}_B = \frac{1}{n_B} \sum_{i \in B} y_i$.
- Define Big Data inclusion indictor, δ_i, such that $\delta_i = 1$ if $i \in B$ or zero otherwise. Then $\bar{Y}_B = \frac{1}{n_B} \sum_{i \in U} \delta_i y_i$. Then $\bar{Y}_B$ is generally biased to estimate the population mean, $\frac{1}{N} \sum_{i \in U} y_i$.
- (Random) Survey sample, $A \subset U$, of size n_A. If the survey weights are d_i, then $\bar{Y}_A = \frac{1}{n_A} \sum_{i \in A} d_i y_i$. $\bar{Y}_A$ is unbiased.



9 THE KEY IDEA

- The total in U is the sum of know total in B and the unknown total in C.
- Thus we can estimate the total by $\hat{Y}_{PS} = \sum_{i \in B} y_i + N_C \hat{\bar{Y}}_C$, i.e. a post stratified (PS) estimator.
- $\bar{Y}_C$ can be estimated from the overlap between C and A. Formally, it is $\hat{\bar{Y}}_C = \frac{\sum_{i \in A} (1-\delta_i) d_i y_i}{\sum_{i \in A} (1-\delta_i) d_i}$
- Then PS estimator is the same as $\sum_{i \in A} w_i y_i$, where w_i minimises the following Chi-square distance $\sum_{i \in A} d_i (\frac{w_i}{d_i} - 1)^2$ subject to: $\sum_{i \in A} w_i \delta_i y_i = \sum_{i \in B} y_i; \sum_{i \in A} w_i \delta_i = N_B; \sum_{i \in A} w_i (1-\delta_i) = N_C$
- Solution for $w_i = d_i X^T (\sum_{i \in A} d_i x_i x_i^T)^{-1} x_i$, where $x_i^T = (1-\delta_i, \delta_i, \delta_i y_i), X^T = \sum_{i \in U} x_i^T$.
- We call $\hat{Y}_{RDI} = \sum_{i \in A} w_i y_i$ the Regression Data Integrator.



10 WHY THE TROUBLE?

- The equivalence of a regression data integrator to a post stratified estimator gives us the insight to extend the method to cover more complex situations, by imposing new calibration requirements, for example, when there is:
  - Auxiliary information for all units, or only the units in B;
  - Duplication in the units in B;
  - Measurement errors in A, or B; and
  - Non response in A.

- These have been discussed in the Kim and Tam (2020) paper, and may be the subject of future talks, if desired.

11 CONCLUDING COMMENTS

- The inferential value of Big Data is limited by the extent of under-coverage bias (and measurement errors).
- We show that we can use Big Data as benchmarks to improve substantially the efficiency of sample estimates
  - And by-pass the under-coverage bias issues if Big Data were used directly for inference
- The ABC method can be extended to other more complex cases as described in the previous slide

12 CONCLUDING COMMENT (CONT'D)

- We have not discussed variance estimation
  - but have derived the analytic results of variances and can use plug-in estimators
  - In the ABS example, we used bootstraps to estimate the uncertainty
- We have made substantial progress in:
  - Extending the above method to using Big Data for small area estimation; and
  - Adjust for linkage errors in matching A to B, used in estimating $\overline{Y}_C$
- Finally, random samples are here to stay in the Big Data world
  - Unless there are defensible ways to adjust for Big Data biases

References

- Kim, J.K. and Tam, S.M. (2020). Combing big data and survey sample data for finite population inference. International Statistical Review. To appear.
- 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, 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, 577-588.
- Tam, S.M., Kim, J.K., Ang, L. and Phan, H. (2018b). Mining the new oil for official statistics. Conference paper presented to BigSurv18, Barcelona. Also in "Big Data Meets Survey Science: A Collection of Innovative Methods", John Wiley and Sons.
- Tam, S.M. and Holmberg, A. (2020). New data sources for official statistics – a game changer for survey statisticians? The Survey Statistician, 81. 21-35.