Use of Adaptive and Responsive Design Concepts and Methods in the Integration of Multiple Data Sources

John L. Eltinge John.L.Eltinge@census.gov

Sixth International Workshop for Advances in Adaptive and Responsive Survey Design

November 5, 2019

.

Acknowledgements and Disclaimer

The author thanks Stephanie Coffey, Jaya Damineni, Anup Mathur, Michael Thieme for very helpful discussions of adaptive, responsive and dynamic designs; and thanks participants in several FCSM-WSS data quality workshops for helpful discussions of numerous questions on data quality, sensitivity analysis and adaptive design.

The views expressed here are those of the speaker and do not represent the policies of the United States Census Bureau.

Overview: Use Adaptive Concepts and Methods for Integration of Multiple Data Sources?

I. Integration of Multiple Data Sources: Three Examples

II. Design Concepts – General and Adaptive

III. Questions on Goals, Methods & Impact

I. Integration of Multiple Data Sources

Expand data sources & tools (beyond surveys):

- “Non-designed data” (“organic” “big data”: Groves, 2012; Couper, 2013; Citro, 2014; CNSTAT, 2017b, many others)

- Modeling, data management

Changing expectations on privacy, granularity of information, “evidence-based policymaking”

I. Integration – Three Examples

Example A (“append microdata”): Link survey data with unit-level administrative/commercial records

e.g., CNSTAT report - Consumer Expenditure Survey

Goals: Reduce cost (expenditures, burden), improve quality, especially for high-cognitive load items

I. Integration – Three Examples

Example B (“backbone and bridge”):

- “Backbone”: administrative record sets
- “Bridge”: supplementary sample surveys to calibrate definitions; determine “domain sizes” in multiple-frame extensions

Longstanding cases: Current Employment Survey

Small domain estimation (Rao and Molina, 2015)

I. Integration – Three Examples

Example C (“cleaning data”):

- Preliminary exploration, de-duplication, analysis of incomplete-data and error patterns (especially important for previously unused or uncontrolled data sources)

- Formal edit and imputation procedures

II. Design Concepts – General - 1

A. Goal: Estimate parameters θ

(means, quantiles, regression coefficients, generalized linear models, hierarchical)

Multiple sources provide data Y

Integration based on models $f(Y|X, Z, \beta)$ for true outcomes, errors, missing-data patterns

II. Design Concepts – General - 2

B. Performance Profiles for Estimation of θ

Quality: Accuracy (MSE-TSE, interval properties), Relevance, Timeliness, Comparability, Coherence, Accessibility, Granularity (Brackstone, 1999; CNSTAT, 2017; others)

Also: Risk and cost (often dominate operations)

II. Design Concepts – General - 3

C. Operating Space Defined by

$Z =$ Environment (observed, uncontrolled)

and process outcomes

$$X = \left(X_{Source}, X_{Method}, X_{System}, X_{Admin}\right)$$

= Design vector (resource decisions)

II. Design Concepts – General - 4

Schematic model: “Performance profile” vector

$$P = \ (Quality, Risk, Cost) = g_{\theta}(X, Z;\ \gamma) + e$$

e = residual effects (uncontrolled, unobserved)

γ = parameters of performance profile, dispersion

Spell out dominant layers of conditioning

II. Design Concepts – Adaptive - 1

A. Adaptive/Responsive/Dynamic Survey Design:
Extensive literature

- Two-phase sampling (Cochran, 1977, others)

- Many recent developments, e.g., Groves and Heeringa (2006), Rosenblum et al (2019), Schouten et al. (2018), Tourangeau et al. (2017), this session

II. Design Concepts – Adaptive - 2

B. Broad Concept: Change (adapt) some of

$$X = \left(X_{Source}, X_{Method}, X_{System}, X_{Admin}\right)$$

based on refined information on Z, γ or β

to improve "performance profile" vector

$$P = (Quality, Risk, Cost) = g_{\theta}(X, Z;\, \gamma) + e$$

II. Design Concepts – Adaptive - 3

C. Common (not exclusive) focus:

- Survey nonresponse

- Refined information via paradata (may require extensive systems work)

- Related diagnostics (e.g., R-indicators)

III. Questions - Goals, Methods & Impact -1

Extend Adaptive Concepts and Methods to Integration of Multiple Data Sources?

Example A (“append microdata”):

Ex: Alignment of non-response follow-up with availability of imputation based on linked records, imperfect prediction models

III. Questions - Goals, Methods & Impact -2

Example B ("backbone and bridge"): Adaptive supplementary surveys to build the "bridges"?

Ex: Capture subpopulations not included
in the administrative data sources?

Ex: Estimate "domain sizes" in multiple-frame settings?

Ex: Estimate regression coefficients, other parameters
to calibrate administrative variables
with idealized concepts?

III. Questions - Goals, Methods & Impact -3

Example C (“cleaning data”):

Ex: Capture and use paradata for modeling of (clustered) patterns of incomplete data, measurement error; impact on entity resolution performance

Ex: Adaptive capture of quality information to inform “fish or cut bait” decisions on data source

III. Questions - Goals, Methods & Impact -4

Extend adaptive-survey procedures (e.g., Tourangeau et al, 2017, others):

1. Performance profiles: error, cost, other

 - Collection of Y across multiple sources, phases

 - Truly focus on means (MSE, mean cost) or on controlling extremes (cf. “dashboards”)?

III. Questions - Goals, Methods & Impact -5

2. Dominant & modifiable design features X

   - Mechanisms for timely modification of X?

3. Dominant & observable environmental & process variables Z

   - Production-quality system for timely capture and operational use of Z?

III. Questions - Goals, Methods & Impact -6

4. Realistic approximations for

$$P = \ (Quality, Risk, Cost) = g_{\theta}(X, Z;\ \gamma) + e$$

5. Align information on Z, γ or β with feasible modification of X

6. Revisit (4) with data-driven design X:

   - (Conditional) bias, variance inflation?

III. Questions - Goals, Methods & Impact -6

7. Alignment with concepts and methods for:

   - Sensitivity analysis

   - Transparency, reproducibility and replicability (e.g., Stodden et al, 2014; NASEM, 2019)

IV. Summary: Use Adaptive Concepts and Methods for Integration of Multiple Data Sources?

I. Integration of Multiple Data Sources

II. Design Concepts -

General: Select X to Balance Multiple Criteria

Adaptive: Adjust X from Updated Z, γ or β

III. Questions on Goals, Methods & Impact

Thank You!

John L. Eltinge
Assistant Director for Research
and Methodology
U.S. Census Bureau
John.L.Eltinge@census.gov

References (1)

Biemer, Paul P., Edith de Leeuw, Stephanie Eckman, Brad Edwards, Frauke Kreuter, Lars E. Lyberg, N. Clyde Tucker, Brady T. West (Editors) (2017). *Total Survey Error in Practice.* New York: Wiley.

Brackstone, Gordon (1999). Managing Data Quality in a Statistical Agency. *Survey Methodology* **25**, 139-149.

Citro, Constance F. (2014). From Multiple Modes for Surveys to Multiple Sources for Estimates. *Survey Methodology Journal* **40,** 137-161.

Cochran, W.G. (1977). *Sampling Techniques, Third Edition.* New York: Wiley.

Elliott, Michael R. and Richard Valliant (2017). Inference for Nonprobability Samples. *Statistical Science* **32**, 249-264.

References (2)

Eltinge, John L. (2013). Integration of matrix sampling and multiple-frame methodology. Proceedings of the 59th World Statistical Congress. https://www.statistics.gov.hk/wsc/IPS033-P4-S.pdf

Groves, Robert M. and S.G. Heeringa (2006). Responsive Design for Household Surveys: Tools for Actively Controlling Survey Errors and Costs. *Journal of the Royal Statistical Society, Series A* **169,** 439-457.

Kaputa, Stephen J. and Katherine J. Thompson (2018). Adaptive Design Strategies for Nonresponse Follow-Up in Economic Surveys. *Journal of Official Statistics* **34**, 445-462.

Lohr, Sharon L. (2011). Alternative Survey Sample Designs: Sampling with Multiple Overlapping Frames. *Survey Methodology* **37**, 197-213.

References (3)

Lohr, Sharon L. (2018). Measuring Uncertainty with Multiple Sources of Data. *Proceedings of Statistics Canada Symposium 2018.*

Lohr, Sharon L and Trivellore E. Raghunathan (2017). Combining Survey Data with Other Data Sources. *Statistical Science* **32**, 293-312

Meng, Xiao-Li (2018). Statistical Paradises and Paradoxes in Big Data (I): Law of Large Populations, Big Data Paradox and the 2016 U.S. Presidential Election. *Annals of Applied Statistics* 1-42.

National Academies of Sciences, Engineering, and Medicine (2017). *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. Washington, DC: The National Academies Press. https://doi.org/10.17226/24893.

National Academies of Sciences, Engineering, and Medicine (2019). Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. https://doi.org/10.17226/25303

References (4)

Rao, J.N.K. and I. Molina (2015). *Small Area Estimation, Second Edition.* New York: Wiley.

Rosenblum, Michael, Peter Miller, Benjamin Reist, Elizabeth Stuart, Michael Thieme and Thomas Louis (2019). Adaptive Design in Surveys and Clinical Trials: Similarities, Diffferences, and Opportunities for Cross-Fertilization. *Journal of the Royal Statistical Society, Series A*

Schouten, Barry, Andy Peytchev and James Wagner (2018). *Adaptive Survey Design.* Boca Raton, Florida: CRC Press

Steorts, Rebecca (2015). Entity Resolution with Empirically Motivated Priors. *Bayesian Analysis* **10**, 849-875.

Stodden, V, F. Leisch and R.D. Peng (2014). *Implementing Reproducible Research*. London: CRC Press

Tourangeau, Roger, J. Michael Brick, Sharon Lohr and Jane Li (2017). Adaptive and Responsive Survey Designs: A Review and Assessment. *Journal of the Royal Statistical Society, Series A* **180** 203-223.