Quality of Geospatial Data Integrated from Multiple Sources

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

FCSM Workshop on Quality of Integrated Geospatial Data

October 26, 2018

Acknowledgements and Disclaimer

The author thanks many colleagues in the federal statistical system, academia and the private sector for very valuable discussions of the topics considered in this paper.

Overview: Dimensions of Data Quality

I. FCSM Workshops on Integration of Multiple Data Sources: Transparent Reporting & Practical Improvement

II. "Relevance" Dimension of Quality

III. "Accuracy" Dimension of Quality

Columns: Operational Goals

Rows: Quality Dimensions	**Transparent Reporting**	**Practical Improvement**
Relevance		
Accuracy		
Many others, plus risk & cost		

I. FCSM Workshops on Data Quality - 1

A. Prospective Integration of Multiple Data Sources: Wonderful Opportunity to

1. Improve quality/risk/cost profiles of current statistical information products and services

2. Expand statistical portfolios

II.B. Data Quality – Previous Workshops

- Input data quality (December 1, 2017)

- Processing quality (January 25, 2018)

- Output data quality (February 26, 2018)

- Metadata (September 14, 2018)

I.C. Data Quality – Multiple Dimensions - 1

Quantitative features: "accuracy"

Extend traditional "Total Survey Error" models (Biemer et al, 2017)

- Population coverage
- Linkage errors
- Definitional errors and inconsistencies
- Incomplete data
- Estimation errors (Lohr and Raghunathan, 2017; Elliott and Valliant, 2017)

I.C. Data Quality – Multiple Dimensions - 2

Qualitative features:

relevance, timeliness, comparability, coherence, accessibility

This talk: “relevance” and “accuracy” dimensions for geospatial data

II. "Relevance" Dimension - 1

A. General issue: Do our

- Formal conceptual and statistical framework
- Measurement and estimation methods
- Interpretation of results (and limitations)

align well with primary inferential questions of key stakeholders (and value conveyed)?

II.A. "Relevance" Dimension - 2

In other words: Spell out clearly

- What questions are we asking?
- Why (and when) are the questions (and answers) important for specified groups of data users?
- Consider both "use value" and "option value"

II.B "Relevance" - Definitions

For well-defined population (large literature)

Y = Outcome variables

$X = (X_G, X_I, X_A)$ = Predictor variables

X_G = Geospatial

X_I = Substantive interest: Intervention?

X_A = Other auxiliary vars considered important

II.C Relevance - Inferential Goals - 1

Understand conditional distributions

$F_Y(y|X_G)$ or $F_Y(y|X_G, X_I, X_A)$

and functionals thereof

Ex: conditional means, dispersion effects, quantiles, parameters of applicable models

II.C. Relevance - Inferential Goals - 2

1. Relevant level of geospatial granularity?

Focus on: $F_Y(y|X_G)$

a. Inherent interest in specified geography: "my county"

II.C. Relevance - Inferential Goals - 3

b. Substantial numerical differences in $F_Y(y|X_G)$ across specified areas
(implicit: relative to predictive uncertainty)

i. Empirical evidence (e.g., historical pattern)

ii. Substantial practical impact if present
(cf. "option value" in assessing utility)

II.C. Relevance - Inferential Goals - 4

2. Indications of prospective intervention effects?

Options: $X_I = X_{I1}$ or X_{I2}

Compare:

$$F_Y(y|X_G, X_I = X_{I1}, X_A) \text{ vs. } F_Y(y|X_G, X_I = X_{I2}, X_A)$$

and related quantities

II.D. Relevance – Applications - 1

Five Applications, with Prospective Interpretation (cf. CEP, 2017)

1. Purely descriptive reports (means or totals) and related ranks
   - Per "triple goal" estimation (Shen and Louis, 1998)

II.D. Relevance – Applications - 2

2. Tables: Describe association between Y and X_I, after accounting for X_G, X_A

3. Prediction:

   - Predictive distribution $F_Y(y|X_G)$?
   - Change $F_Y(y|X_G, X_I, X_A)$ w/different X_I ?

II.D. Relevance – Applications - 3

4. Perceived causality (per extensive literature, e.g., Imbens and Rubin, 2015):

   Change in X_I leads to change in Y ?
   Concrete mechanism? Level of granularity?

5. Perceived control: A decision to change from $X_I = X_{I1}$ to $X_I = X_{I2}$

   leads to a specified change in $F_Y(y|X_I, X_A)$
   accounting for "slippage" from nominal X_{I2}

II.D. Relevance – Applications - 4

For any of “description,” “association,” “prediction,” “causality” or “control”:

1. Level of aggregation (e.g., geography) for:
   - Practical distinctions among areas, groups
   - Inform realistic decisions on prospective intervention?

II. Relevance – Applications – 5

2. Quality of information at the specified level of aggregation (section III)?

3. Stakeholder risks incurred through poor quality or break in series?

4. Value conveyed, accounting for (1)-(3)?
   - Both “use value” and “option value”

III. “Accuracy” Dimension of Quality

A. Assessment of estimation (prediction) accuracy

   – Accounting for which components of variability?

III.B. “Accuracy” – Performance - 1

Incremental improvements in accuracy of estimators (predictors) based on:

- Outcome data Y (sample survey, admin data)
- Additional geospatial data X_G (sample, population level)
- Further predictors X_A (sample, population)

III.B. "Accuracy" – Performance - 2

Of special interest: Incremental improvement in accuracy from including X_A, as well as Y and X_G

- i.e., extra effort (acquisition and management of X_A; additional modeling) worthwhile?

- Empirical question – diagnostics and commonly observed results?

III.C. "Accuracy" Measures - 1

1. Relevant conditioning:

   Extension of standard "total survey error" models to integration of multiple sources (Biemer et al, 2017; Japec et al., 2015)

   Esp: Population coverage, missing X variables, temporal effects, "unit problems" (filing units) and variable-specification issues

III.C. “Accuracy” Measures - 2

2. Record linkage effects

3. Adjust for exploratory-analysis and model-selection effects

   a. Contrast between formal inference and exploratory analyses (cf. Tukey, 1962, others)

   b. Nuances among multiple inferential goals (Shen and Louis, 1998, “triple goal” SAE)

III.C. “Accuracy” Measures - 3

3. Reporting summaries

   a. For specific estimands and point estimators

   b. Summaries across estimands
      - Of special interest for “unified” decision
        on spatial estimation methods

III.D. "Accuracy" – Implications

1. Solid inferences: What do (can) we know fairly well from current data, accounting for errors?

2. Response to abovementioned limitations:
   - Find better data sources (more admin records; calibration/bridge surveys)? Cost-effective?
   - Improve linkage, imputation, analysis methods?

III.E. Quality and Risk

1. Loss of, or major changes in, data sources

2. Production system changes (w/related costs)

3. Disclosure issues

Tools for identification and management of risks? Implications for management and integration of regional data sources?

III.F. Quality and Cost

For many resource dimensions

1. Incorporate both fixed and variable cost components

2. Additional costs incurred through integration of multiple data sources
   - cf. "complex supply chain management"

III.G. Quality, Risk and Cost

Empirical Information on Dominant Factors for Quality, Risk and Cost? Vary Across Sources?

1. Observational data (e.g., paradata)

2. Formal experiments – factorial designs or evolutionary operation?

3. Modeling diagnostics (Rao & Molina, 2015; Elliott and Valliant, 2017; Lohr & Raghunathan, 2017)

IV. Closing Remarks

A. Quality of Geospatial Data Based on Integration of Multiple Data Sources

1. Multiple dimensions of quality (plus cost and risk)

2. Today's talk:

   "relevance" and "accuracy" dimensions

IV.B. "Relevance" Dimension

Spell out via $F_Y(y|X_G)$ & $F_Y(y|X_G, X_I, X_A)$

- What questions are we asking?
- Why (and when) are the questions (and answers) important for specified groups of data users?
- Consider both "use value" and "option value"

IV.C. "Accuracy" Dimension

Extensions of "total survey error" models, to include:

- Overfitting effects?

- Incremental improvement in accuracy from including X_A, as well as Y and X_G:

$$F_Y(y|X_G) \;\&\; F_Y(y|X_G, X_A)$$

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.

Biemer, Paul P., Dennis Trewin, Heather Bergdahl and Lilli Japec (2014). A System for Managing the Quality of Official Statistics. *Journal of Official Statistics* **30**, 381–415. https://doi.org/10.2478/jos-2014-0022

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

Commission on Evidence-Based Policymaking (2017). *The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking.* Available through:

https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf

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

References (2)

Imbens, Guido W. and Donald B. Rubin (2015). *Causal Inference for Statistics, Social and Biomedical Sciences: An Introduction.* New York: Cambridge University Press.

Japec, Lilli, Frauke Kreuter, Marcus Berg, Paul Biemer, Paul Decker, Cliff Lampe, Julia Lane, Cathy O'Neil, Abe Usher (2015). AAPOR Task Force Report on Big Data. Available through:

https://www.aapor.org/Education-Resources/Reports/Big-Data.aspx

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

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.

References (3)

National Research Council (2013) . *Principles and Practices for a Federal Statistical Agency, Fifth Edition.* Committee on National Statistics. Constance F. Citro and Miron L. Straf, Editors. Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

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

Shen, W. and T.A. Louis (1998). “Triple-Goal Estimates in Two-Stage Hierarchical Models.” *Journal of the Royal Statistical Society, Series B* **60**, 455-471 **https://doi.org/10.1111/1467-9868.00135**

Simon, H. A. (1956). "Rational Choice and the Structure of the Environment." *Psychological Review* **63** (2), 129–138.

Tourangeau, Roger, Brad Edwards, Timothy P. Johnson, Kirk M. Wolter and Nancy Bates (eds.) (2014). *Hard-to-Survey Populations.* New York: Wiley.

Tukey, John W. (1962). The Future of Data Analysis. Ann. Math. Statist. 33, 1--67. doi:10.1214/aoms/1177704711

Wolter, K.M. (2007). *Introduction to Variance Estimation, Second Edition.* New York: Springer.