Practical Evaluation of Proposals for Integration of Multiple Data Sources

John L. Eltinge

FedCASIC Conference – April 16, 2019
Acknowledgements and Disclaimer

The author thanks colleagues in government statistical agencies, universities and private-sector organizations in many countries; and especially thanks the participants in the 2017-2018 series of WSS/FCSM workshops on data quality for many valuable insights.

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

Integrate multiple data sources to produce high-quality statistical information products and services on a sustainable and cost-effective basis

Data Integration – Four Examples (1)

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

Goals: Reduce cost (expenditures, burden), improve quality, esp for high-cognitive load items
Data Integration – Four Examples (2)

Example B (“backbone and bridge”):

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

Longstanding case: Current Employment Survey
Data Integration – Four Examples (3)

Example C: Multiple-source extensions of traditional multiple-frame/multi-mode methods (e.g., Lohr and Raghunathan, 2017)
Crucial issue: ests of domain sizes, features

Example D: Small domain estimation (Rao and Molina, 2015)
Crucial issues: predictor variables, quality of fit
Data Integration – Four Examples (4)

All Four: Spectrum of Statistical Products:

- Tabular publications, graphs, maps
- Microdata releases (caution re disclosure)
- In-depth modeling results (per Commission on Evidence-Based Policymaking, 2017)
One Remaining Question: How?

Practical Criteria for Evaluation of Proposals for Capture and Integration of Multiple Sources

1. Trajectory of development – systems, products

2. Required information flow & decision points

3. Forestall distractions from both:
   - “Hype cycle” phenomena (Gartner, 2016)
   - Excessive skepticism
Suggested Evaluation Criteria

I. Outcome Oriented: Quality, Risk, Cost

II. Cross-Cutting:
   Stakeholder Expectations
   Structure
   Processes
   Communication
I. Outcome-Oriented Criteria

A. Quality – Interface of Product & User

Accuracy (main technical focus)
Relevance, Timeliness, Comparability,
Coherence, Accessibility, Granularity

- Brackstone (1999), CNSTAT (2017)
I.B. Quality – “Accuracy” Dimension

1. Anchor in inferential goals:
   a. Estimands, sources of uncertainty
   b. Exploratory vs. standardized production: reproducibility & replicability
I.B. Quality – “Accuracy” Dimension

2. Extensions of “total survey error” terms, with extensive assessment of model fit

Ex: Population coverage, linkage errors & entity resolution, definitional errors, incomplete data; est errors (Lohr & Raghunathan, 2017; Elliott & Valliant, 2017, Meng, 2018)
I.D. Other Dimensions of Quality

Relevance, Timeliness, Comparability, Coherence, Granularity, Accessibility

Specific criteria often context-dependent:
- Users & uses
- Challenging with heterogeneous user base
- Use cases to connect specific criteria with concrete value delivered to key stakeholder?
I.E. Outcome Criteria: Risk (1)

Identifiable system-level events that degrade sustainability: disclosure, “break in series”:

Ex: Failure in development timeline, system quality

Ex: Loss or undetected major change in data source

I.E. Outcome Criteria: Risk (2)

Align with literature on:

- Complex supply chains
- Fault-tolerant designs
- “Normal accidents” (via complex and tightly coupled systems – Perrow, 1999)
- Related behavioral issues (e.g., risk homeostasis)
I.F  Outcome Criteria - Cost (1)

For proposed sources & integration methods, spell out:

- Cash expenditure – direct collection, systems
- Other scarce resources (burden, personnel)
- Contingencies for risk management
I.F. Outcome Criteria - Cost (2)

Cost models for integration of multiple sources

- Expected value (upper quantiles?) for fixed and variable cost components

- Fixed budgets, cost over-runs & related incentives

- Depreciation of (intangible) capital investments, accounting for multiple-source uncertainties on duration & magnitude of use & maintenance?
II. Cross-Cutting Issues: For Each of Quality, Risk and Cost

A. Stakeholder Expectations & Linkage w/Value

1. Context: One-off special study, prototype, pilot, or full-scale robust production?

2. Vision on quality/risk/cost criteria; related constraints; uncontrolled externalities?

3. Roles of inferential goals, data availability?
II.B. Structural Effects - Scale

1. Scale Issues: Examples

- Input data sources – number, complexity
- Processing: Actions, time, resources
- Output: Products and features thereof
II.B. Structural Effects: Scale

2. For each example

a. Relevant unit of scale?

b. Dominant scale issues: occasional “surge”, steady change?

c. Scale functions: predictors, curvature, asymptotes, quality of fit?
II.C. Structural Effects: Constraints

1. Resources: Cash, Equipment, Calendar Time, Intangible Capital (especially human capital)

2. Optionality structure:
   - Direct or indirect ability to adjust constraints?
   - Cost of adjustment? Who pays? Incentives?
II.D. Cross-Cutting: Processes

1. Technical Processes: Methodology, systems
   Directly applicable literature & practice?

2. Managerial Processes:
   - Transparent, Controllable, Accountable?
   - Internal: Financial, human resources
   - External: Contracting (multiple inputs)
II.E. Cross-Cutting: Communication

1. Language and standards to provide sufficient clarity on answers & crucial nuances
   - Concrete anchors, images for stakeholders?

2. Consistent with cultural expectations on clarity & uncertainty?
   - cf. Gartner “hype cycle” critiques; Perrow (1995) on adoption & diffusion of technology
III. Conclusions

Evaluation of Proposals for Data Integration

A. Outcome Oriented: Quality, Risk, Cost

B. Cross-Cutting: Stakeholder Expectations Structure, Processes, Communication

C. Capture and Use of Criteria at All Stages: Exploratory, Prototype, Pilot & Production
Thank You!

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

John.L.Eltinge@census.gov
Data Sources & Tools

Capture, Production, Dissemination

Information Needs
I.C. Quality – “Accessibility” Dimension

1. Dissemination Options (per CEP, 2017)
   a. Standard tables, graphs, maps – public
   b. Restricted-access research data centers

2. Impact of disclosure avoidance methods
   (changing technical and societal environment)