Discussion of "Identifying and Addressing a Break (Blip) in Series" by Lynn Langton

John L. Eltinge
Assistant Director for Research and Methodology
U.S. Census Bureau
Chair, Federal Committee on Statistical Methodology

FCSM/WSS Third Workshop on Transparent Reporting on the Quality of Integrated Data
February 26, 2018

Acknowledgements and Disclaimer

The author thanks Lynn Langton for the opportunity to review the material covered in this discussion.

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

Overview: Hearty thanks to Lynn Langton and BJS

A. NCVS survey example:

- Highlights important “break in series” issues for published estimates from surveys or from integration of multiple sources

- Important to share experiences, methods and results with stakeholders, other statistical organizations

Overview (continued)

B. Brief discussion

1. Review and summarize some general "break in series" issues and methods

Overview (continued)

2. Implications for transparent quality reporting

- Predominant factors that (may) affect quality/risk/cost profiles of our statistical products:
  Changes in populations, data sources, methodology (including adjustments and mitigation steps)

- Empirical results and impact on stakeholder value

3. Appendix: Additional technical material

I. “Break in Series” Phenomena

A. Essentially all data-capture systems are imperfect

Issue here: Comparability of results over time

1. NCVS example:

“new-interviewer” effects, respondent fatigue

Impact: “level shift” in crime rate estimates

I. Break in Series (continued)

2. Non-survey cases: Loss of, or major change in, data source

   a. Lose access to major third-party data source

   b. Quality of source changes – possibly undetected

   c. Production system incompatible with new system of third-party data provider

   d. Do not meet production schedule, quality standards

   e. Effects of some disclosure-limitation procedures

I. Break in Series (continued)

3. Impact on data quality, per Workshops #1 & 2:

- (Sub)Population coverage
- Incomplete data (group, unit, item)
- Definitional issues
- Imperfect web-scraping, record linkage, de-duplication, data fusion, imputation
- Model lack of fit

I. Break in Series (continued)

B. Much of methodology and practice:

Attempts to mitigate issues in (A.1)-(A.3)

C. BUT: Internal or external changes in sources, methodology or practice can produce a "break in series":

- Mean structure: Proportions, means, totals
- Dispersion structure: estimates "look less stable"
- Seasonal patterns (quarterly, monthly, weekly)
- Outliers (risk of gross errors)

II. Implications for Transparent Quality Reporting

A. NCVS Survey Example: Analysis and Communication

1. Diagnostics carefully calibrated with predominant features of underlying design (time-in-sample groups, new/experienced interviewers)

2. Practical impact of potential adjustments, costs

II. Implications for Transparent Quality Reporting (continued)

B. Principles for Integration of Multiple Data Sources:

1. Design data capture and integration methods to be robust against primary "break in series" risk factors

Issue: Many potential risk factors

- Some "ad hoc" adjustments, judgment calls

II. Implications for Transparent Quality Reporting (continued)

B.2. Resulting “robust” (“fault tolerant”) design inevitably requires complex trade-offs among (many?) quality/risk/cost profile components

II. Implications for Transparent Quality Reporting (continued)

B.3. Two-way stakeholder communication

a. What we know about potential "breaks in series" & prospective mitigation strategies

b. Stakeholder priorities and risk tolerance

- Concrete case studies?

III. Closing Remarks

A. Thanks to Lynn Langton and BJS:

- Important illustration of “break in series” issue from a prominent survey

- Impact on the “accuracy” and “comparability” dimensions of quality

III. Closing Remarks

B. Extend to Integration of Multiple Sources

- Empirical assessment of prospective impact on quality/risk/cost profile
- Robust (fault tolerant) design options
- Case-specific adjustments
- Two-way stakeholder communication

C. Examples from audience?

Thanks to all for your insights

Additional comments welcome:
John.L.Eltinge@census.gov

Appendix: Some Technical Features of "Break in Series" Phenomena

A. Formal description of "break in series" phenomena

1. Notation: estimand θ_{jt} for group j, period t
   (e.g., mean, proportion, total, regression coefficient)

   Estimator $\hat{\theta}_{jt} = \theta_{jt} + e_{jt}$

   Design and environmental variables: X

Appendix (continued)

2. Error distribution

$$e_{jt} \sim \left(\mu_{ejtX}, \sigma^2_{jtX}\right)$$

More formally: the random variable e_{jt} has a location-scale distribution function, within the family $F^*_{jt}(\cdot)$:

$$F_{ejtX}(y) = F^*_{jt}\left[\{y - \mu_{ejtX}\}/\sigma_{jtX}\right]$$

Issue: Realistic extent of empirical information on F_{ejtX}, and alignment with specific data sources and related risks?

Appendix (continued)

3. Under this framework, one may consider several types of "break in series" associated with changes in the distribution of e_{jt} . These include:

   Level shift: Change in μ_{ejtX}

   Dispersion effects: Change in σ_{jtX}

   Outlier effects: Change in $F_{jt}^{*}(\cdot)$

4. In addition, one may consider extensions of the abovementioned notation to characterize changes in patterns of autocorrelation or seasonality.

Appendix (continued)

B. Risk literature (Crockford, 1986; Perrow, 1999; Flyvbjerg and Budzier, 2011): Need systematic evaluation of:

1. Prospective causes of failure (system design flaws, single- or multi-point events)

2. Timelines, costs for identification and recovery from failure

3. Impact of failure and recovery on stakeholders

4. Robustness of process against failure
   - Esp. important for official statistics due to limited control over third-party providers of non-survey sources

Appendix (continued)

C. Of special interest: Perrow, C. (1999), *Normal Accidents:* Risks incurred in "complex and tightly coupled systems"

1. Deterioration in performance can occur more quickly than one can detect and mitigate the underlying problems

2. Potential application to integration of multiple data sources: Timely detection and mitigation of most likely problems

Appendix (continued)

D. “Fault tolerant” designs – allow quick recovery after failure

1. Literature from engineering, computer science:
   Denning (1976), Laprie (1985), Zhang, Gray and Gonzalez (2004, 2005), Monkman and Schagaev (2013)

2. Extend to integration of multiple data sources

   Ex: Parallel production during transitions

   Ex: Timely and cost-effective use of backup source if proposed data source fails?

Appendix (continued)

E. Time series literature on “change in regime” and “change-point estimation” – apply diagnostics to:

1. Estimators $\hat{\theta}_{jt}$

2. Source-specific components that contribute to $\hat{\theta}_{jt}$