Practical Approaches to Design and Inference Through the Integration of Complex Survey Data and Non-Survey Information Sources

John L. Eltinge, Scott Fricker & Daniel Yang, BLS
Rachel M. Harter & Jamie Ridenhour, RTI International

Biometric Society – ENAR Meetings
Session #51 "Statistical Challenges of Survey and Surveillance Data in the U.S. Government"

March 16, 2015

BLS
BUREAU OF LABOR STATISTICS
U.S. DEPARTMENT OF LABOR

Acknowledgements and Disclaimer

The authors thank Jennifer Edgar, David Friedman, Steve Henderson, Brett McBride, Bruce Meyer, Bill Mockovak, Jay Ryan, Adam Safir and Lucilla Tan for many helpful discussions of topics considered in this paper. The numerical results in Section IV are part of a larger study in Fricker, Yang and Eltinge (2015).

The views expressed here are those of the authors and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics, the U.S. Food and Drug Administration nor RTI International.

Overview

I. Introduction:
   Data from Surveys and Alternative Sources

II. Framework for Integration

III. Design Example:
   Frame Enrichment from Alternative Data Sources

IV. Collection Example:
   Administrative Data for Sample Units

I. Data from Surveys and Alternative Sources

A. Prospective Data Sources for Government Agencies, Other Large-Scale Statistical Organizations

1. Traditional sample surveys (e.g., Fuller, 1999):

   a. High degree of design control, replicability

   b. Specification of
      - Target population(s), parameter(s)
      - Components of uncertainty considered in inference

I. Surveys and Alternative Sources (Continued)

2. Alternative Data Sources: “Big,” “Non-designed” or “Organic Data” (Groves, 2011, 2013; Couper, 2013):
   - Generated for non-statistical purposes
   - Limited (or no) “design control”
   - Often “tall and thin” = “variable poor”

   a. Specialized admin (taxes, regulation, benefits)

      Ex: Automobile titles (transactions & tax)

   b. Commercial transactions

      Ex: Subscription lists

I. Surveys and Alternative Sources (Continued)

c. Internal corporate files (with informed consent)

Ex: Employment, wage, benefit and price files

d. Web-scraped data on product features, prices

e. Social media

Ex: Unemployment, job openings (Shapiro, 2014)

f. Search engine results

Ex: Disease outbreaks (Google flu)

Ex: Demographics (Cressie et al., 2013)

I. Surveys and Alternative Sources (Continued)

C. Two General Approaches to Integration

1. Use alternative source to supplement the survey
   - Enhance sample frames
   - Target subpopulations in sampling
   - Improve unit contact, other fieldwork
   - Direct replacement of burdensome, expensive or error-prone survey items
   - Improved auxiliary info for edit, imputation or weighting

I. Surveys and Alternative Sources (Continued)

2. Focus primary attention on the “organic” data source (some good statistical properties, with limitations - coverage/representativeness, definitional issues, measurement biases, aggregation effects)

   a. Specialized sample survey - adjust for limitations

      Ex: U.S. Current Employment Survey

   b. Need to develop broader classes of supplementary survey designs

BLS

II. Framework for Integration

A. General Design Goal:

Balance Multiple Dimensions of Quality, Cost & Risk

B. Six Quality Dimensions (e.g., Brackstone, 1999):

Qualitative (timeliness, relevance, comparability, coherence and accessibility)

Quantitative (total survey error model components)

II. Framework for Integration (Continued)

C. Total Survey Error: An Estimator-Focused Approach:

(Estimator) – (True value)

= (frame error)
+ (sampling error)
+ (nonresponse effects)
+ (measurement error)
+ (processing effects)

Andersen et al. (1979), Groves (1989), Weisberg (2005), Biemer (2010), Lyberg (2012), Kenett and Shmueli (2014), many others

II. Framework for Integration (Continued)

D. Integration of non-survey sources with survey data fits with extensions of TSE models to non-survey settings, e.g.,

Biemer (2014)
Davern (2007, 2009, 2010)
FCSM (1980, SPWP #6)
Herzog, Winkler and Scheuren (2007)
Iwig et al. (2013, Data Quality Assessment Tool)
IAOS (2008) Conference Proceedings
Jabine and Scheuren (1985)
Jeskanen-Sundstrom (2007)
Ord and Iglarsh (2007)
Penneck (2007)
Royce (2007)
Winkler (2009)
Zhang (2009, 2011, 2012)

II. Framework for Integration (Continued)

E. General approach: Integrate organic data source(s) to reduce magnitudes of TSE component(s), costs

F. Properties center on empirical results, but we often have limited information in initial exploration

1. Extend standard design ideas to obtain needed additional information quickly and at low cost.
2. Often must supplement with sensitivity analyses

III. Example: Frame Enrichment from Alternative Data Sources

A. Frames: List of Prospective Sampling Units

1. Example: Survey of youths aged 11 to 16 to evaluate an anti-tobacco media campaign

2. Design: Address-based sampling (ABS)

   a. Frames use updates from the U.S. Postal Service Computerized Delivery Sequence (USPS-CDS) file

   b. Overviews: Iannacchione (2011), Link (2010)

III. Frame Enrichment (Continued)

3. Supplementary information from some vendors

a. For field operations: Indicators for vacancy, educational housing (dormitories), seasonal housing, post office boxes, matching with telephone lists

b. Subpopulation membership indicators: Basic demographics, presence of children in the household

c. Geography: latitude and longitude (matching with specific blocks, other Census geographical groups, jurisdictions)

III. Frame Enrichment (Continued)

B. Some research on the auxiliary variables in address-based frames or on marketing database variables appended to the frames:

Amaya et al. (2014)
DiSogra et al. (2010)
Dekker and Murphy (2014)
Harter and McMichael (2013)
Hubbard et al. (2014)
McMichael et al. (2014)
Ridenhour et al. (2014)
Roth et al. (2013)
Valliant et al. (2014)

III. Frame Enrichment (Continued)

C. Tobacco-Use Study:

1. Initial strata and primary sample units based on large geographical aggregates

2. Within selected block groups: Further stratification of households based on refined frame information?

   Subpop membership: race-ethnic classification, presence of youth, specific ages of youth

   Smoking-related: education, income, other SES

III. Frame Enrichment (Continued)

D. Question:

Worthwhile to use additional variables in frame?

1. Evaluation: Cost-variance trade-offs

   a. Costs of acquiring more variables, linking to current frame

   b. Possible cost reductions

      - Facilitate unit contact
      - Screening for targeted subpopulations

III. Frame Enrichment (Continued)

2. Variances of key estimators

   a. Prevalence (and initiation) rates for usage of specific types of tobacco

   b. Coefficients of related logistic reg models

3. Important cautionary note:
   Under differential sampling rates facilitated by enriched frames, variance-cost trade-offs may not be well approximated by customary measures like raw sample counts, design effects

III. Frame Enrichment (Continued)

4. Further complications:

a. Quality of additional frame variables?

- Proportion missing? Informative missingness?

- Out of date (e.g., household move, dissolution)

- Misclassification

b. Impact on sampling properties of enriched frame?

III. Frame Enrichment (Continued)

E. Evaluate Efficiency of Prospective Alternative Design that Uses More Frame Variables

1. Use previous survey information for:

   a. Subpop "hit rates," means, variances

   b. Relative costs for screening, in-depth data collection

   c. Data quality

2. Sensitivity analyses often required

IV. Collection Example: Administrative Data for Sample Units

A. U.S. Consumer Expenditure Survey:
Consent to Link with Administrative Data

1. Large-scale household survey that collects many items on consumer expenditures, as well as income and assets

2. Voluntary responses, so respondent cooperation is crucial

3. Respondent concerns potentially include both burden and privacy

IV. Administrative Data for Sample Units (Continued)

4. From 2011 CE Research Section:

> "We'd like to produce additional statistical data, without taking up your time with more questions, by combining your survey answers with data from other government agencies. Do you have any objections?"

5. Previous studies:

   Davis, Elkin, McBride and To (2013)

IV. Administrative Data for Sample Units (Continued)

B. Unweighted Summary of Responses (Davis et al., Table 2A.1):

RESPOBJ	Count	Percent
Yes, object	942	18.89
No, do not object	3951	79.24
Do not Know/Refusal	93	1.87

IV. Administrative Data for Sample Units (Continued)

C. Sensitivity analysis - Impact on estimator quality if:

1. Ask “do you object” question:

a. If unit objects: Treat as nonrespondent for linkage to certain govt-related variables

b. If unit does not object: Link unit to obtain those variables from government source

2. Emphasize: Approach (1) not currently carried out in the field – requires exploratory analysis and legal/regulatory clearance

IV. Administrative Data for Sample Units (Continued)

D. Steps in Sensitivity Analysis: Extending previous "consent to link" literature (privacy vs. burden)

1. Propensity models for P(Do not object)

Predictor variables from:

- Demographics, socioeconomic status

- Behavioral variables (respondent's effort in responding, stated concerns about privacy, being too busy, etc.)

- Interactions among some predictors

IV. Administrative Data for Sample Units (Continued)

2. Explore Estimators of Population Means for:

FINCBTAX: Total amount of family income before taxes = "Income"

ZPROPTAX: Property taxes

EVEHPUR: Vehicle purchase cost (outlays for vehicle purchases including downpayment, principal and interest paid on loans, or if not financed, purchase amount) = "Vehicle cost"

IV. Administrative Data for Sample Units (Continued)

3. Three estimation approaches:

   a. Use all data (as in current production)

   b. Use only data from the "not object" units, but with no weighting adjustment

   c. Use only data from the "not object" units, and adjust customary weights with additional factor:

      1/P(Do not object to linkage with government data)

4. Standard errors account for the sample design through balanced repeated replication

Full Sample Analysis

Variable			
	N	Mean	SE
Income	4893	50939.00	1227.51
Property tax	4893	454.15	10.41
Vehicle cost	4893	599.59	33.22

For “Not Object” Units: No Adjustment

Variable	N	Mean	SE
Income	3951	52869.00	1535.04
Property tax	3951	429.12	10.76
Vehicle cost	3951	619.14	37.05

For “Not Object” Units: Propensity Adjustment

Variable			
	N	Mean	SE
Income	3915	52117.00	1523.85
Property tax	3915	434.74	11.39
Vehicle cost	3915	607.80	36.63

Note:

P.S.: Propensity Scores = 1 - P(Object).

Full Sample vs. Unadjusted "Not Object"

Variable	Point Est. Diff.		
	Agree - Full	$SE_{BRR}(\hat{\theta}_{Agree} - \hat{\theta}_{Full})$	$t(\hat{\theta}_{Agree} - \hat{\theta}_{Full})$
Income	1930.00	580.98	**3.32**
Property tax	-25.02	6.08	**-4.12**
Vehicle cost	19.55	10.83	1.81

Full Sample vs. “Not Object” with Adjustment

	Point Est. Diff.		
Variable	Agree P.S. - Full	$SE_{BRR}(\hat{\theta}_{Agree\ P.S.} - \hat{\theta}_{Full})$	$t(\hat{\theta}_{Agree\ P.S.} - \hat{\theta}_{Full})$
Income	1178.00	619.09	1.90
Property tax	-19.41	6.48	**-2.99**
Vehicle cost	8.21	15.34	0.54

V. Closing Remarks

A. Summary: Integration of Multiple Data Sources

1. Distinct cases:

   a. Survey dominant, but supplemented by "organic" source – tobacco & CE examples

   b. "Organic" source dominant, with survey to "fill in the gaps"

2. Practical issue: Design and analytic approaches based on limited information, sensitivity analyses

V. Closing Remarks (Continued)

B. Prospective Extensions

1. Empirical assessment of cost and risk components

    a. Initial exploratory analyses & pilot studies

    b. During production processes

2. Other components of total survey error

Contact Information

John L. Eltinge

Associate Commissioner

Office of Survey Methods Research

www.bls.gov/ore

202-691-7404

Eltinge.John@bls.gov