Integration of Matrix Sampling and Multiple-Frame Methodology

John L. Eltinge
U.S. Bureau of Labor Statistics, Washington, DC USA Eltinge.John@bls.gov

Abstract

This paper considers the integration of matrix sampling and multiple-frame-multiple-mode methodology to reduce survey burden and to improve the balance of data quality, cost and risk. Primary emphasis is placed on applications that involve the combination of data from sample surveys and administrative records.

Keywords: Administrative record data, data quality, dual-frame survey, multiple-mode survey, partitioned design, respondent burden, survey cost structures, survey risk, total survey error, transaction data

1. Introduction: Modification of Survey Procedures to Balance Multiple Measures of Quality, Cost and Risk

Practical design of a survey procedure generally is intended to balance a relatively large number of performance criteria related to quality, cost and risk. The effort to strike this balance takes place under substantial constraints on the type and magnitude of available resources, as well as stakeholder expectations regarding quality and risk profiles.

In recent years, some surveys have encountered two important changes in the nature of these constraints. First, increasing concerns about cost and respondent burden have led statistical organizations to reconsider standard procedures of administering all relevant survey items to each selected sample unit. Second, in some cases statistical analysts have increased opportunities to access alternative data sources from, e.g., administrative records, transaction logs or social media.

This paper considers some ways in which to address these changed constraints through integration of previously developed methods for partitioned sample designs, and for multiple-frame, multiple-mode surveys. Section 2 introduces some relevant concepts and reviews selected previous literature. Section 3 presents a general framework to assess data quality under this integrated approach. Section 4 provides some concluding remarks. A longer version of this paper develops the applicable mathematical and methodological ideas in additional detail.

2. Two Approaches to Reduction of Burden and Improvement of Data Quality

2.1 Matrix Sampling and Other Forms of Partitioned Designs

The design of large-scale data collection processes often encounter a natural tension between a preference for in-depth coverage, and limitations related to cost and respondent burden. Under appropriate conditions, collection of more information from a sample unit can lead to estimation of population parameters for a larger number of substantive variables; production of more reliable estimates for a larger set of subpopulations of interest; and deeper multivariate analyses. However, collection of additional variables can increase burden on respondents, and thus lead to deterioration in data quality.

To address this issue, the methodological literature has developed a set of techniques for multiple matrix sampling and other forms of partitioned designs. The main idea is that a given sample unit is only asked to provide responses for a subset of the items covered in the full survey instrument, with the subset determined through a randomization mechanism. In some cases, the randomization mechanism may be based only on variables provided for all population units through a frame. In other cases, the assignment of sections may occur through a variant of two-phase sampling, in which all sample units may receive an initial set of core questions (e.g., for relatively simple demographic or other classification variables), and probabilities of assignment to more burdensome questions are based on responses to the core questions.

Early work in this area arose primarily in educational measurement; see, e.g., Sirotnik (1970), Shoemaker (1973), Munger and Lloyd (1988) and references cited therein. More recently, these ideas have been adapted and extended for use with complex sample surveys that collect data for a wide range of social and economic data. Examples include Hinkins (1984), Raghunathan and Grizzle (1995), Thomas et al. (2006), Gonzalez and Eltinge (2008), Chipperfield and Steel (2009) and Gonzalez (2012).

2.2 Multiple Data Sources: Multiple-Frame and Multiple-Mode Methodology

In some cases, one may further reduce burden by obtaining some data through, e.g., administrative records. (The current discussion with use the term "administrative record data" to cover a broad range of data produced through non-survey sources, including records related to taxation, administration of government or private-sector benefit programs, regulatory reporting, medical diagnoses and treatment, or commercial transactions. Somewhat similar forms of data have recently become available through social media streams, but these data may involve features that are beyond the scope of this paper.) Such records have some important potential strengths. For example, they may allow the coverage of some details (e.g., specifics of a given purchase or a given medical diagnosis and treatment) that may be difficult or impossible to capture through standard interview processes. This in turn may lead to a substantial reduction in respondent burden, provided the administrative record data are available for a large portion of the population at a reasonable cost.

However, administrative and transaction records also have important potential limitations. For example, in many cases, they fail to cover the full target population; provide only a subset of the items of substantive interest; or are subject to nontrivial forms of error, including incomplete data, definitional effects, aggregation effects and reporting errors.

For these reasons, it is important to distinguish among several potential ways in which one might integrate survey data with administrative record data. Three cases are of special interest for the current work.

Case 1: Use of aggregated administrative record data with aggregated survey data. In some settings, administrative record data are available only in aggregated form. For example, due to confidentiality restrictions, the survey organization may only be allowed access to means or totals for specified subpopulations. In addition, these subpopulations may not provide full coverage of the entire population of interest, and the reported means or totals may be subject to some of the error sources described in the preceding paragraph. For this case, one would need to carry out supplementary data collection for the subpopulations not covered in the administrative records. In addition, the supplementary survey instrument generally would need to include questions that would determine which sample units were covered by the abovementioned administrative records, unless that record-coverage information was already provided by the frame for the supplementary survey.

The resulting design and estimation work is in large part an extension of previous literature on multiple-frame-multiple-mode surveys. For general background on this literature, see, e.g., Fuller and Burmeister (1972), Hartley (1974), Bankier (1986), Lepkowski and Groves (1986), Skinner and Rao (1996), Haines and Pollock (1998), Lohr and Rao (2000, 2006), Lu and Lohr (2010), Wolter et al. (2010), Lohr (2011) and references cited therein.

Case 2: Use of administrative record microdata for replacement or imputation of the full set of survey responses for some, but not all, units in the target population. For this case, one would ideally have administrative record microdata that align directly with concepts and measures covered by the survey instrument. Under those idealized conditions, one would still need to carry out a supplementary survey for the portion of the population not covered by the administrative records. The resulting design and estimation methods amount to an extension of previous literature on microdata-level work with multiple-frame-multiple-mode surveys, where now the errors in question involve sampling and nonsampling errors for units covered by the supplementary survey, and nonsampling errors for the units covered by the administrative records.

In many settings, however, the abovementioned ideal is not met, but the administrative record data provide a basis for imputation of the microdata of interest. The resulting design and estimation work must provide sufficient information (e.g., relevant regression coefficients) to produce efficient imputations; and must also account for imputation errors for the administrative record units, as well as all of the error sources mentioned in the preceding paragraph.

Case 3: Linkage of survey and administrative record microdata at the sample unit level for some units. For this case, one selects a set of sample units and collects some “core” variables for those units only through the survey instrument. For the same sample units, one attempts to collect other variables through linkage with administrative records at the microdata level. However, such linkage generally will not be feasible for all sample units, due to limitations of population coverage in the administrative record source, and due to limitations of record linkage methodology. For sample units for which this linkage is not feasible, one may attempt to collect the applicable data through a supplement to the “core” instrument; depending on costs and other considerations, the supplement may be administered to all applicable sample units, or to a subsample of units selected through a two-phase design. Under conditions, the resulting data structures are variants on those encountered in multiple-matrix sampling. For example, in the current case, the random-design-determined probabilities of assignment to a given instrument section are replaced by quasirandom-mechanism- determined probabilities of linkage of a sample unit with the administrative record source. This in turn leads to extensions of multiple-frame-multiple-mode methods for estimation. As in the previous two cases, it is important for the survey procedure to capture information on the proportions of the population that are covered by a specified administrative record source.

3. Integration of Partitioned-Design and Multiple-Source Methodology

Efficient estimation and inference based on data collected under any of cases 1, 2 or 3 requires balanced consideration of a wide range of sources of variability. This generally will include the superpopulation considered to have generated the finite population of interest; the sampling mechanisms used for the formal survey design; record-linkage effects; and nonsampling error processes associated with incomplete data, measurement errors and aggregation effects encountered in the survey and administrative data. Detailed exploration of these effects involve extensions of standard “total survey errors” approaches to design and analysis of sample surveys, and related work with errors in administrative record systems. For general background on "total survey error" approaches, see, e.g., Andersen et al. (1979), Weisberg (2005) and references cited therein. For extensions of some "total survey error" ideas to cases involving data from registers and other forms of administrative records, see, e.g., Davern (2009), and Zhang (2011, 2012).

For extensions of "total survey error" ideas to partitioned-design and multiple-source methodology, three issues are of special interest. First, it is important to have a clear consensus regarding the level of precision required in estimation of a given population or subpopulation parameter. As with standard matrix-sampling work, one may use a relatively small effective (sub)sample size in collection of data that will contribute only to estimation of certain parameters that do not require high levels of precision. Conversely, one may need much higher subsampling rates for variables that contribute to parameters that do require precise estimation. After adjusting for subpopulation membership issues, similar comments apply to combined collection of survey and administrative record under a matrix design.

Second, in development of supplementary survey plans, it is important to distinguish carefully between supplements intended for direct collection of data that are comparable to data provided by administrative record sources for other units; and supplements intended primarily to collect data that will be used to support methodological adjustments. Examples of important methodological adjustments include estimation of overlap rates among the various frames and administrative record sources; estimation of regression coefficients and related parameters for use in imputation and other microdata adjustments; and estimation of variance terms used in weighting and inference.

Third, data provided through administrative records are often viewed as high-quality and cost-effective alternatives to data collected through sample surveys. However, as indicated in the discussion of cases 1 through 3 above, administrative data can themselves have nontrivial error properties which in turn can have a substantial impact on the quality of the resulting combined-data estimators. In addition, there can be substantial costs associated with each additional source of administrative record data to be integrated into the overall estimation system. Important cost factors may include data acquisition; record linkage; additional field collection work required by use of multiple data collection instruments; data analysis required for evaluation of error components and development of the resulting combined-data estimator; and development and maintenance of production systems to implement all features of the combined-data procedure. In evaluation of these costs, both fixed and marginal cost components are important for the decision of whether to include a given data source, and for decisions on specific ways in which that data source will be used. Also, some administrative record sources are subject to substantial continuity risks. For example, some data elements from records based on tax, regulatory or benefit-administration processes may cease to exist or undergo fundamental changes if there are changes in underlying legislation or regulations. Similar comments apply to some data provided on a voluntary or contractual basis from private-sector sources. Optimization of the resulting design requires a balance of each of these cost, quality and risk factors.

4. Discussion

This paper has provided a brief overview of ideas related to the integration of matrix sampling and multiple-frame-multiple-mode methodology to reduce survey burden and to improve the balance of data quality, cost and risk. In principle, one could apply these ideas to cases in which all of the data were collected through sample survey methods. However, the current work has placed principal emphasis on applications in which some of the data sources arise from administrative records. For those applications, the paper identified three distinct cases defined by the features of the administrative-record data, and by the ways in which those data are subsequently combined with sample survey data. Such cases require in-depth extensions of "total survey error" models to administrative-record settings, and also involve consideration of multiple components of cost and risk.

Acknowledgement and Disclaimer

The author thanks Mick Couper, Don Dillman, Jennifer Edgar, Scott Fricker, Jeffrey Gonzalez, Michael Horrigan, Carol House, William Iwig, Frauke Kreuter, Sharon Lohr, William Mockovak, Trivellore Raghunathan, Jay Ryan, Adam Safir, Nathaniel Schenker and Richard Valliant for many helpful discussions of some of the ideas considered in this paper. The views expressed in this paper are those of the author and do not necessarily reflect the policies of the U.S. Bureau of Labor Statistics.

References

Andersen, R., J. Kasper and M.R. Frankel (1979). *Total Survey Error.* Jossey_Bass, San Francisco.

Bankier, M.D. (1986). Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys. *Journal of the American Statistical Association* **81**, 1074-1079.

Chipperfield, J.O. and D.G. Steel (2009). Design and Estimation for Split Questionnaire Surveys. *Journal of Official Statistics* **25**, 227-244.

Davern, M. (2009). What Is Needed to Build the Next Generation of Linked Survey and Administrative Data Files for Policy Research. Paper Presented to the Washington Statistical Society, Jyly 30, 2009.

Fuller, W.A. and L.F. Burmeister (1972). Estimators for Samples Selected from Two Overlapping Frames. *Proceedings of the Social Statistics Section, American Statistical Association*, 245-249.

Gonzalez, J.M. (2012). *The Use of Responsive Split Questionnaires in a Panel Survey.* Unpublished Ph.D. Dissertation, Joint Program in Survey Methodology, University of Maryland/University of Michigan.

Gonzalez, J.M. and J.L. Eltinge (2008). Adaptive Matrix Sampling for the Consumer Expenditure Quarterly Interview Survey. *Proceedings of the Section on Survey Research Methods, American Statistical Association.*

Haines, D.E. and K.H. Pollock (1998). Combining Multiple Frames to Estimate Population Size and Totals. *Survey Methodology* **24**, 79-88.

Hartley, H.O. (1974). Multiple Frame Methodology and Selected Applications. *Sankhya, Series C* **36**, 99-118.

Hinkins, S.M. (1984). Matrix Sampling and the Effects of Using Hot Deck Imputation. *Proceedings of the Section on Survey Research Methods, American Statistical Association*, 415-420.

Lepkowski, J.M. and R.M. Groves (1986). A Mean Squared Error Model for Dual Frame, Mixed Mode Survey Design. *Journal of the American Statistical Association* **81**, 930-937.

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

Lohr, S.L. and J.N.K. Rao (2000). Inference from Dual Frame Surveys. *Journal of the American Statistical Association* **95**, 271-280.

Lohr, S.L. and J.N.K. Rao (2006). Estimation in Multiple-Frame Surveys. *Journal of the American Statistical Association* **101**, 1019-1030.

Lu, Y. and S. Lohr (2010). Gross Flow Estimation in Dual Frame Surveys. *Survey Methodology* **36**, 13-22.

Munger, G.F. and B.H. Lloyd (1988). The Use of Multiple Matrix Sampling for Survey Research. *Journal of Experimental Education* **56**, 187-191.

Raghunathan T.E. and J.E. Grizzle (1995). A Split Questionnaire Survey Design. *Journal of the American Statistical Association* **90**, 54-63.

Shoemaker, D.M. (1973). *Principles and Procedures of Multiple Matrix Sampling.* Ballinger Publishing Company, Cambridge, Massachusetts.

Sirotnik, K.A. (1970). An Investigation of the Context Effect in Matrix Sampling. *Journal of Educational Measurement* **7**, 199-207.

Skinner, C.J. and J.N.K. Rao (1996). Estimation in Dual Frame Surveys with Complex Designs. *Journal of the American Statistical Association* **91**, 349-356.

Thomas, N., T.E. Raghunathan, N. Schenker, M. Katzoff and C. Johnson (2006). An Evaluation of Matrix Sampling Methods Using Data from the National Health and Nutrition Examination Survey. *Survey Methodology* **32**, 217-231.

Weisberg, H.F. (2005). *The Total Survey Error Approach: A Guide to the New Science of Survey Research.* University of Chicago Press, Chicago, Illinois.

Wolter, K.M., P. Smith and S.J. Blumberg (2010). Statistical Foundations of Cell-Phone Surveys. *Survey Methodology* **36**, 203-215.

Zhang, L.-C. (2011). Topics of Statistical Theory for Register-Based Statistics and Data Integration. *Statistica Neerlandica* **66**, 41-63.

Zhang, L.-C. (2012). A Unit-Error Theory for Register-Based Household Statistics. *Journal of Official Statistics* **27**, 415-432.