

Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Comments on “Trends and directions in sample survey theory and methods”

by Jean D. Opsomer, Daifeng Han and Medha Uppala

Release date: June 30, 2025



Statistics
Canada

Statistique
Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at infostats@statcan.gc.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

- Statistical Information Service 1-800-263-1136
- National telecommunications device for the hearing impaired 1-800-363-7629
- Fax line 1-514-283-9350

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, the Agency has developed standards of service which its employees observe in serving its clients. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under “Contact us” > [“Standards of service to the public.”](#)

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

© His Majesty the King in Right of Canada, as represented by the Minister of Industry, 2025

Use of this publication is governed by the Statistics Canada [Open Licence Agreement](#).

An [HTML version](#) is also available.

Cette publication est aussi disponible en français.

Comments on “Trends and directions in sample survey theory and methods”

Jean D. Opsomer, Daifeng Han and Medha Uppala¹

Abstract

In this discussion, we complement the excellent overview by Profs. Lohr and Rao with some additional topics. The first topic is a call for more recognition of the central role of modeling in survey estimation. The second is a brief discussion of the use of partial frame information in survey design. Finally, we draw the attention to recent increases of synthetic methods, in particular, multilevel regression and poststratification (MRP) in small area estimation applications.

Key Words: Design-based inference; Model-based inference; Multilevel regression and poststratification (MRP); Sampling design.

1. Introduction

We congratulate Profs. Lohr and Rao on an excellent and timely overview of the state of survey statistics, demonstrating that it continues to be a vibrant research area. In this discussion, we will highlight three topics that complement the authors’ overview: the increasing role of modeling in survey estimation, the application of partial frame information in survey design, and the recent increase of synthetic methods in small area estimation applications.

2. Generalized design-based estimation: We are all modelers now

Traditionally, design-based estimation has two key properties: (1) *representativeness*: the estimators are statistically representative of the population from which the sample originates, and (2) *model-independence*: inference is possible without appealing to model assumptions. These two properties are guaranteed within the pure randomization-based framework that constitutes the foundation of design-based inference. They also continue to hold within the model-assisted paradigm, in which the contributions of the model are fully captured by the randomization distribution of the estimators. In fact, properties (1) and (2) are often treated as interchangeable, with the representativeness of estimators justified by the fact that randomization was used to obtain the data. This is stated in the term “design-based”, which we still use despite the fact that in the presence of nonresponse (which can now reach 90% for some survey modes), both the model assumptions and the modeling approaches become an intrinsic part of the construction and distribution of the estimators, i.e. (2) no longer holds.

Violation of (2) does not automatically invalidate (1), although claiming that estimators are statistically representative without appealing to (2) requires a more complete and rigorous justification. As the authors

1. Jean D. Opsomer, Daifeng Han and Medha Uppala, Westat, Inc., 1600 Research Blvd., Rockville, Maryland, 14850, U.S.A. E-mail: jeanopsomer@westat.com.

demonstrate in their overview article, sophisticated statistical methods have been developed by the survey statistics research community, not only to account for unit and item nonresponse, but also to combine data from different surveys, even to integrate survey and non-survey data. As a result, the distinction between “design-based” and “model-based” is no longer meaningful in the sense of differentiating the methods and the inferential framework used to create survey estimates.

However, there are a number of key characteristics of survey inference that still set it apart from other areas of statistics. As “generalized design-based” statisticians, our inferential goal continues to be to provide data that make it possible to describe and/or estimate characteristics of $\mathcal{U} = \{\mathbf{y}_i, i \in U\}$ based on the sample data $\mathcal{S} = \{\mathbf{y}_i, i \in s\}$, with \mathbf{y}_i denoting a vector of variables for population unit i . This is in contrast to “model-based” statisticians, who are typically making inference on an underlying model for random variable \mathbf{Y} observed in the sample, denoted $\mathcal{F}_s(\mathbf{Y})$ (although they might make inferential claims about a population model $\mathcal{F}_U(\mathbf{Y})$ as well). Even though generalized design-based statisticians need to continue accounting for the original sample design, inference for \mathcal{U} requires modeling the complete selection process, including all forms of non-observation (including nonresponse, non-locating, matching errors, frame errors, etc). As long as the selection process is correctly modeled, the statistical representativeness of estimates with respect to the population continues to hold even though the inference is no longer model-independent.

Survey statisticians, both researchers and practitioners, have already been operating in this generalized design-based framework for several decades. However, the balance between known and unknown sources of uncertainty (i.e. the sampling design and the other selection mechanisms) has shifted to the extent that the latter sources now frequently dominate the overall uncertainty. Therefore, given the increasing importance of selection models with their associated assumptions in estimation, it would behoove us to more explicitly recognize this model dependence in the information provided with survey datasets. In particular, claims of “statistical representativeness” that rely on model assumptions might be accompanied by a suitable disclaimer, for example: “the survey weights and replicate weights reflect the sampling design and modeled selection probabilities that are based on the following predictors: [...] Under the assumption that the selection model is correctly specified, the use of these weights in estimation and modeling ensure that the results are statistically representative of the population”.

This transparency will also make clearer to data analysts and researchers in other areas of statistics that today’s survey statisticians are no longer only applying design-based methods and have instead become sophisticated modelers as well. Our particular expertise is in modeling the complex selection processes that result in the collected data, blending these models with the sampling design randomization, and creating efficient and representative estimation methods. This expertise continues to represent a critically important contribution to the science of statistics. The broad scope of methodological advances discussed by Profs. Lohr and Rao demonstrate that we are well positioned to address the current and future challenges inherent in providing statistically valid inference for populations of interest.

3. Use of incomplete auxiliary information in sampling

In the section devoted to survey design, Profs. Lohr and Rao note that sample allocation continues to be an important research topic. We agree and wish to briefly highlight a particular type of allocation that is becoming increasingly useful in household sampling when there is no access to a high quality frame. In this situation, sampling can be very inefficient if the goal is to screen for subpopulations of interest (e.g., households with children) or to oversample particular subpopulations to increase the precision for domain estimates (e.g., estimates for rare racial/ethnic minority groups). In the US, address-based sampling (ABS) has become widely used in the past decade. The sampling frame consists of all postal addresses in the US, but lacks information on the characteristics of the households at those addresses. ABS frame vendors can append characteristics of addresses such as the number of adults living in the address, age of the head of the household home tenure, Hispanic origin, and the presence of children (Valliant, Hubbard, Lee and Chang (2014); Roth, Caporaso and DeMatteis (2022)), but that information is of varying quality and only available for a subset of the addresses.

A general approach for using such information in targeting a particular subpopulation is to form “high-density” and “low-density” sampling strata using the available frame variable(s). Sampling units with missing frame information are automatically classified in the low-density stratum. It is then possible to allocate the sample so that the “high-density” stratum is oversampled, with corresponding reduction in sampling rate in the “low-density” stratum. It is clear that the target subpopulation can be expected to be present in both strata, in different proportions. Two conditions are required for this approach to improve the efficiency of estimation (Waksberg, Judkins and Massey, 1997). First, in the “high-density” stratum, there has to be a high proportion of cases that belong to the subpopulation or domain of interest. Second, among all the cases that belong to the subpopulation or domain, those in the “high-density” stratum have to account for a sufficiently high proportion of their population total. Unless the vendor information is of sufficiently high quality, one or both conditions can be violated.

While both area-level (e.g., census block group or tract) and address-level information can be used for such targeting, an important advantage of using address-level information is that the subpopulation does not need to be geographically clustered (Chen and Kalton (2015); Dutwin, Coyle, Lerner, Bilgen and English (2024)). It is also feasible to use predictive modeling with area-level characteristics in combination with the ABS vendor-appended variables for stratification and oversampling. A promising approach in this direction is the Bayesian Surname and Geocoding (BSG) method for oversampling racial/ethnic minority groups. BSG and its variants use geographic information (e.g., census tract) along with surname to predict a set of probabilities of membership in five racial/ethnic groups (i.e., White, Black, Hispanic, Asian, and other) for each record (Elliott, Morrison, Fremont, McCaffrey, Pantoja and Lurie (2009); Khanna, Bertelsen, Rosenman, Olivella and Imai (2022)).

As noted, the address-level auxiliary information comes from many sources with varying quality and completeness. There has been some evaluation of the data quality of such auxiliary information, but few provide insight about whether the auxiliary information was accurate enough for effective targeting purposes

(Battaglia, Dillman, Frankel, Harter, Buskirk, McPhee, DeMatteis and Yancey (2016); Dutwin et al. (2024)). One exception is Roth et al. (2022), which discussed the tradeoff between nominal sample size and design effect when using imperfect auxiliary information to target subpopulations of interest. Since the goal of targeting is to improve the precision of the estimates under a fixed cost, it is critical to examine the *effective* sample size (calculated as the nominal sample size divided by the design effect) for the domain estimates and overall estimates.

4. Multilevel regression with poststratification

Small area estimation is a well-established and active research topic among survey statisticians, with sophisticated area-level and unit-level modeling approaches suitable for a wide range of applications. These small area estimators can often be expressed as combinations of synthetic and direct estimators and have the important property that, as the sample sizes in the small areas increase, the contribution of the direct estimators increase, i.e. the small area estimators converge to the direct estimators. In a somewhat new development outside of our research community, multilevel regression with poststratification (MRP) has gained increased acceptance as an alternative small area estimation method, especially in public health and related areas. See for instance Zhang, Onufrak, Holt and Croft (2013), Zhang, Holt, Lu, Wheaton, Ford, Greenlund and Croft (2014), Zhang, Holt, Yun, Lu, Greenlund and Croft (2015), Davila-Payan, DeGuzman, Johnson, Serban and Swann (2015), Zgodic, Eberth, Breneman, Wende, Kaczynski, Liese and McLain (2021) and Wang, Tevendale, Lu, Cox, Carlson, Li, Shulman, Morrow, Hastings and Barfield (2022).

Originally introduced by Gelman and Little (1997) as a hierarchical logistic regression with post-stratification, the goal was to improve on raking-type post-stratification adjustments by producing estimates for a large set of domains rather than a limited set of categories. Although similar to traditional unit-level small area estimators, MRP estimators are synthetic and not guaranteed to be close to the direct estimators even with large area sample sizes. Nevertheless, given their increased use in small area estimation applications, we feel that it is useful for survey statistics to be aware of this development.

We briefly describe the MRP approach in a simple unit-level small area estimation setting. Let y_{ij} represent a unit-level target variable, with unit $j = 1, \dots, N_i$ within small area $i = 1, \dots, m$, and a corresponding categorical covariate vector \mathbf{x}_{ij} observed in the sample. The number of sampling units in a cell defined by the intersection between all levels of the covariates in each small area is known from an external source.

The first step of MRP is to fit a regression model of y on \mathbf{x} using the sample data. Reflecting the “multi-level” portion of MRP, this model will often include random effects, which are intended to improve the overall fit and stability of the model and in particular, do not have to correspond to the small areas. Once the estimates for all model parameters are obtained, either via frequentist or Bayesian methods, the model is used to predict the expected value of y for each combination of the levels of the covariates in each small area. The second step of MRP, post-stratification, aggregates the predictions of y by averaging over the cells in proportion to their known totals in each small area.

A practical advantage of MRP is that it allows one to predict y for values of x that have not been sampled or with very small sample sizes. However, MRP's implicit assumption of an ignorable survey design is a significant barrier. Other barriers include incorporating cluster samples, availability of population counts and identifying strong group-level predictors for multilevel regressions. A growing literature on MRP addresses some of these barriers. Recent work by Gelman, Si and West (2024) incorporates sampling weights as regressors and estimates the population distribution of weights to eventually predict y in every cell; these survey weighted predictions of y are then poststratified by averaging over known population counts.

References

- Battaglia, M.P., Dillman, D.A., Frankel, M.R., Harter, R., Buskirk, T.D., McPhee, C.B., DeMatteis, J.M. and Yancey, T. (2016). Sampling, data collection, and weighting procedures for address-based sample surveys. *Journal of Survey Statistics and Methodology*, 4, 476-500.
- Chen, S., and Kalton, G. (2015). Geographic oversampling for race/ethnicity using data from the 2010 U.S. population census. *Journal of Survey Statistics and Methodology*, 3, 543-565.
- Davila-Payan, C., DeGuzman, M., Johnson, K., Serban, N. and Swann, J. (2015). Estimating prevalence of overweight or obese children and adolescents in small geographic areas using publicly available data. *Preventing Chronic Disease*, 12, 140229.
- Dutwin, D., Coyle, P., Lerner, J., Bilgen, I. and English, N. (2024). Leveraging predictive modelling from multiple sources of big data to improve sample efficiency and reduce survey nonresponse error. *Journal of Survey Statistics and Methodology*, 12, 435-457.
- Elliott, M.N., Morrison, P.A., Fremont, A.M., McCaffrey, D.F., Pantoja, P.M. and Lurie, N. (2009). Using the Census Bureau's surname list to improve estimates of race/ethnicity and associated disparities. *Health Services and Outcomes Research Methodology*, 9, 69-83.
- Gelman, A., and Little, T.C. (1997). [Poststratification into many categories using hierarchical logistic regression](https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997002/article/3616-eng.pdf). *Survey Methodology*, 23(2), 127-135. Paper available at <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997002/article/3616-eng.pdf>.
- Gelman, A., Si, Y. and West, B.T. (2024). MRPW: Regression, poststratification and small-area estimation with sampling weights. Technical report, Columbia University. Available at http://www.stat.columbia.edu/~gelman/research/unpublished/weight_regression.pdf.

- Khanna, K., Bertelsen, B., Rosenman, E., Olivella, S. and Imai, K. (2022). wru: Who are you? Bayesian prediction of racial category using surname and geolocation. R package version 1.0.0. Available at <https://cran.r-project.org/web/packages/wru/index.html>.
- Roth, S., Caporaso, A. and DeMatteis, J. (2022). Variables appended to abs frame: Has their data quality improved? *PLoS ONE* 17, e0269110.
- Valliant, R., Hubbard, F., Lee, S. and Chang, C. (2014). Efficient use of commercial lists in U.S. household sampling. *Journal of Survey Statistics and Methodology*, 2, 182-209.
- Waksberg, J., Judkins, D. and Massey, J.T. (1997). [Geographic-based oversampling in demographic surveys of the United States](#). *Survey Methodology*, 23(1), 61-71. Paper available at <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1997001/article/3107-eng.pdf>.
- Wang, Y., Tevendale, H., Lu, H., Cox, S., Carlson, S.A., Li, R., Shulman, H., Morrow, B., Hastings, P.A. and Barfield, W.D. (2022). U.S. county-level estimation for maternal and infant health-related behavior indicators using pregnancy risk assessment monitoring system data, 2016-2018. *Population Health Metrics*, 20, 14.
- Zgodic, A., Eberth, J.M., Breneman, C.B., Wende, M.E., Kaczynski, A-T., Liese, A.D. and McLain, A.C. (2021). Estimates of childhood overweight and obesity at the region, state, and county levels: A multilevel small-area estimation approach. *American Journal of Epidemiology*, 190, 2618-2629.
- Zhang, X., Holt, J.B., Lu, H., Wheaton, A.G., Ford, E.S., Greenlund, K.J. and Croft, J.B. (2014). Multilevel regression and poststratification for small-area estimation of population health outcomes: A case study of chronic obstructive pulmonary disease prevalence using the Behavioral Risk Factor Surveillance System. *American Journal of Epidemiology*, 179, 1025-1033.
- Zhang, X., Holt, J.B., Yun, S., Lu, H., Greenlund, K.J. and Croft, J.B. (2015). Validation of multilevel regression and poststratification methodology for small area estimation of health indicators from the Behavioral Risk Factor Surveillance System. *American Journal of Epidemiology*, 182, 127-137.
- Zhang, X., Onufrak, S., Holt, J.B. and Croft, J.B. (2013). A multilevel approach to estimating small area childhood obesity prevalence at the census block-group level. *Preventing Chronic Disease*, 10, 120252.