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

Survey Methodology

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

by David Haziza

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

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2025
Vol. 51, No. 1, pp. 115-123
Statistics Canada, Catalogue No. 12-001-X

Comments on "Trends and directions in sample survey theory and methods"

David Haziza[1]

Abstract

This discussion of the paper by Rao and Lohr focuses on the use of machine learning procedures for estimating finite population parameters. While there is growing interest in these methods within national statistical offices, several areas remain largely unexplored and warrant significant attention in the coming years. In this discussion, I highlight potential topics for future research and development in this rapidly evolving field.

Key Words: Finite population parameter; Imputation; Model-assisted estimation; Prediction; Propensity score estimation; Small area estimation; Statistical inference.

First, I would like to congratulate the authors on an excellent paper commemorating the 50th anniversary of *Survey Methodology*. The paper provides a comprehensive review of recent developments in several areas of practical importance, including sampling methods, weighting methods, handling missing data, data integration, small area estimation, and variance estimation. Furthermore, the authors highlight important directions for future research, such as data equity and the ethical application of machine learning (ML). To paraphrase the authors, "survey sampling has come a long way since the first issue of *Survey Methodology*". In the 1980s and 1990s, estimators were predominantly derived from parametric methods (e.g., linear regression models). Gradually, survey statisticians began exploring and using nonparametric methods, and nowadays, machine learning procedures have become increasingly prevalent in the field.

One does not need to be a fortune-teller to predict that, in the coming years, ML procedures will play a significant role in producing point estimates of finite population parameters. For this reason, my discussion will focus on the use of ML procedures that have gained significant attention in recent years within national statistical offices. ML methods have recently been studied and applied in various settings, including model-assisted estimation, the treatment of missing data, small area estimation, and the data integration of probability and nonprobability samples. Unlike traditional parametric methods, ML models are inherently robust in capturing complex non-linear relationships between variables, even when specific polynomial terms (e.g., quadratic, cubic) or interaction terms are not explicitly included in the model, and this with minimal manual effort. Moreover, many ML algorithms are known for their ability to produce high-quality predictions, such as random forests (Breiman, 2001), Bayesian Adaptive Regression Trees (Chipman, George and McCulloch, 2010) XGBoost (Chen and Guestrin, 2016), and Cubist (Quinlan, 1992). However, because of their flexibility, ML procedures generally require larger sample sizes than parametric procedures (e.g., linear regression and logistic regression), especially when the number of predictors is large. Indeed, their rate of convergence

1. David Haziza, Department of mathematics and statistics, University of Ottawa, Ottawa, Ontario. E-Mail: dhaziza@uottawa.ca.

is typically (much) slower than the parametric rate of $1/\sqrt{n}$. Also, ML models require the specification of hyperparameters, often referred to as the model's architecture. The number of hyperparameters may vary significantly depending on the complexity of the model. Simpler methods, such as regression trees, may require tuning a few hyperparameters only, whereas more complex architectures like convolutional neural networks can involve tuning over fifteen hyperparameters. These hyperparameters are frequently set manually by survey statisticians but some of them can also be optimized using methods like cross-validation. Tuning is a crucial yet often overlooked aspect of ML, as the performance of ML procedures can be highly sensitive to the choice of hyperparameters.

In surveys, ML can be applied to both prediction and the estimation of finite population parameters, although these are distinct objectives. In some cases (e.g., Wieczorek, 2023), the goal is prediction, where the focus is on minimizing prediction error (e.g., mean squared error) for individual outcomes. In this context, Nalenz, Rodemann and Augustin (2024) and Toth and McConville (2024) recently developed design-based versions of random forest algorithms, while Wieczorek (2023) explored the use of conformal prediction in finite populations. In practice, however, the primary objective is often to estimate finite population parameters, such as a population mean, rather than to make predictions for individual values. While predictions are not the main target, they serve as basic "ingredients" in constructing estimators of population means. For example, model-assisted and imputed estimators rely on a set of predictions $\hat{m}(\mathbf{x})$, representing the predicted value of a survey variable y, where $\mathbf{x}$ denotes a vector of predictors. In the context of unit nonresponse, propensity score adjusted estimators utilize a set of estimated response probabilities $\hat{p}(\mathbf{x})$ for the responding units. Since neither $\hat{m}(\mathbf{x})$ nor $\hat{p}(\mathbf{x})$ is of primary interest, they can be viewed as nuisance functions, serving as intermediate steps toward the ultimate goal of estimating finite population parameters.

In recent years, ML methods have been extensively studied and applied in various settings for estimating finite population parameters: (i) Model-assisted estimation (e.g., Montanari and Ranalli, 2005; McConville and Toth, 2019; Dagdoug, Goga and Haziza, 2023a; Sande and Zhang, 2021). (ii) Propensity score estimation in the context of unit nonresponse (e.g., Phipps and Toth, 2012; Lohr, Hsu and Montaquila, 2015; Gelein, 2017; Kern, Klausch and Kreuter, 2019; Larbi, Tsang, Haziza and Dagdoug, 2025; Opsomer and Riddles, 2025). (iii) Single and multiple imputation for the treatment of item nonresponse (e.g., Tan, Flannagan and Elliott, 2019; Wang, Akande, Poulos and Fan, 2022; Deng and Lumley, 2024; Dagdoug, Goga and Haziza, 2023b; Dagdoug, Goga and Haziza, 2025). (iv) Small area estimation (e.g., Viljanen, Meijerink, Zwakhals and Van de Kassteele, 2022; Krennmair and Schmid, 2022; Michal, Wakefield, Schmidt, Cavanaugh, Robinson and Baumgartner, 2024). (v) Data integration (e.g., Ferri-García and Rueda, 2020; Kern, Li and Wang, 2021; Lee, Zhang and Chen, 2023; Liu, Gelman and Chen, 2023; Chen, Xu and Cutler, 2025; Beaumont, Bosa, Brennan, Charlebois and Chu, 2024).

A common practical question is how to assess the properties of a point estimator that incorporates predictions from a ML procedure. Since the primary interest lies in estimation rather than prediction, we continue to rely on our "old friends": bias, variance, and mean squared error of the point estimator. When estimating a finite population parameter, it is important to recognize that good predictions may produce poor estimators in terms of mean square error. A similar situation occurs in the context of causal inference, where the use of ML has been explored for estimating the average treatment effect (e.g., Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey and Robins, 2018). For example, in the context of propensity adjusted score estimation, the nonresponse-adjusted weights are given by $w_i = d_i / \hat{p}(\mathbf{x}_i)$, where d_i denotes the basic weight for unit i. However, models with high predictive power can sometimes produce very small estimated response probabilities $\hat{p}(\mathbf{x}_i)$, potentially leading to extreme adjusted weights w_i. These extreme weights can introduce significant instability into estimators of population means resulting in estimators with a large mean squared error; e.g., Little and Vartivarian (2005), Beaumont (2005), Park, Kim and Kim (2019), and Larbi et al. (2025). Similarly, imperfect predictions do not necessarily lead to a poor estimator. For instance, assuming no nonsampling errors, the model-assisted estimator of a population mean based on a given machine learning procedure remains design-consistent, even if the predictions are obtained through a suboptimal architecture. In fact, such an estimator can even outperform the Horvitz-Thompson estimator substantially in terms of mean square error; see Dagdoug et al. (2023a) for an illustration using random forests.

Because of the typically slower convergence of ML methods, point estimators (e.g., imputed estimators) based on a single ML procedure may be biased in finite samples, as these methods are vulnerable to the curse of dimensionality. Moreover, deriving valid variance estimators and establishing a central limit theorem can be very challenging. In the field of causal inference, several papers advocate the use of doubly robust (DR) estimators for estimating the average treatment effect (e.g., Chernozhukov et al., 2018; Newey and Robins, 2018; Kennedy, 2023) to address these challenges. DR estimators provide two opportunities to reduce potential confounding bias by fitting two models: one that relates the survey variable (prone to missing values) to a set of predictors, and another that relates the response (or participation) indicators to a set of predictors. DR estimators remain consistent if either model is correctly specified, offering protection against misspecification of one of the models. Unlike doubly robust estimators based on parametric models, which rely on strong modeling assumptions, DR estimators based on ML methods may achieve a more effective bias reduction due to their flexibility. An additional advantage of DR estimators based on ML procedures is their ability to achieve the parametric rate $1/\sqrt{n}$ even if both ML procedures converge slowly (e.g., Chernozhukov et al., 2018; Newey and Robins, 2018; Kennedy, 2023). The statistical inference tools recently developed in the field of causal inference may prove highly useful to survey statisticians in various settings, such as the treatment of missing data and data integration methods.

Despite the widespread use of ML procedures in national statistical offices, several areas remain partially or largely unexplored. To conclude this discussion, I highlight below some potential topics for future research and development:

- How to draw statistical inferences (both point and variance estimation) remains a central question for survey statisticians. To date, the problem of statistical inference has primarily been studied for specific ML procedures (e.g., regression trees and random forests) within particular settings, such as model-assisted estimation or imputation. In the coming years, it would be highly desirable to develop "generic" statistical inference tools that are applicable to a wide range of ML procedures. For example, in the context of imputation for item nonresponse, some preliminary findings are available in Dagdoug and Haziza (2024), who considered DR point estimators as well as variance estimators based on cross-fitting. However, the problem of statistical inference in the presence of ML for the treatment of unit nonresponse through weight adjustment procedures remains an open question.

- Many surveys are multipurpose, and it is often unclear in advance which types of parameters will be of interest to data users. For instance, users may require estimates of finite population quantiles in addition to totals or means. While most of the existing literature on the use of ML procedures in surveys has focused on estimating population totals and means, inference for more complex parameters, such as quantiles, remains largely unexplored.
- Resampling methods, such as the bootstrap and jackknife, are widely used in national statistical offices to estimate the variance of point estimators. However, their properties and behavior remain to be explored when applied to variance estimation for point estimators derived from ML procedures.
- Common questions that often arise in practice include: In a specific setting (e.g., propensity score-adjusted estimation), which ML procedure should we choose? And for a given ML procedure, how do we select the best architecture? As mentioned above, in the context of propensity score-adjusted estimation, the method that generates the best predictions may not necessarily yield the best estimator of a population mean. To address this, it would be desirable to develop appropriate criteria, such as selecting the ML procedure that minimizes the estimated mean square error of the propensity score-adjusted estimator. An alternative to selecting the best ML procedure (or architecture) is to use an aggregation approach, which involves fitting multiple ML models to generate a set of predictions for each sample unit and combining these predictions to produce an aggregate score. The reason for using an aggregation procedure is that no single ML method consistently outperforms all others across all scenarios. A procedure that works well in one context may perform poorly in another, making it impossible to predict in advance which method will perform best. An empirical study by Larbi et al. (2025) on the performance of aggregation procedures shows promising results, but more work is still needed to fully understand both their theoretical properties and their behavior in real-life applications.
- With the growing popularity of ML methods and the relatively straightforward application of Rubin's rules for variance estimation (Rubin, 1987) with multiple imputation (MI), the method has come under increased scrutiny in national statistical offices. However, very little research has been conducted on the properties of Rubin's variance estimators in the presence of imputed survey data through ML procedures. The existing literature is mostly empirical (e.g., Wang et al., 2022; Deng and Lumley, 2024). In survey sampling, the properties of imputed estimators are most often evaluated with respect to the joint distribution induced by the imputation model, the sampling design,

and the nonresponse mechanism. Under this framework, the total variance of an estimator after imputation can be expressed as $\mathbb{V}_{\text{tot}} = \mathbb{V}_{\text{sam}} + \mathbb{V}_{\text{nr}} + \mathbb{V}_{\text{mix}}$, where $\mathbb{V}_{\text{sam}}$ and $\mathbb{V}_{\text{nr}}$ correspond to the sampling and nonresponse variance, respectively, and $\mathbb{V}_{\text{mix}}$ is a mixed term (Särndal, 1992). However, the MI variance estimator does not track the mixed term $\mathbb{V}_{\text{mix}}$; see Kim, Brick, Fuller and Kalton (2006). Depending on the sampling design, the finite population parameter of interest, and the imputation procedure, the term $\mathbb{V}_{\text{mix}}$ may be large (positive or negative), in which case Rubin's variance estimator may be significantly biased. More work is needed to better understand the properties of MI point and variance estimators based on ML procedures.

- How does the presence of influential units in the sample affect the performance of ML procedures? This question is particularly relevant in the context of business surveys, where ML procedures can be used, for instance, to impute missing values. Influential units are observations free of error that can disproportionately impact the sampling error of point estimators. That is, their presence in the set of observed data may have a considerable impact on the variance of point estimators. Understanding the behavior of ML procedures in the presence of influential units, as well as determining appropriate architectures and estimation procedures to improve robustness to the presence of influential units, remains an open question.

The use of ML methods is becoming increasingly important and will very likely be unavoidable in the coming years. However, as many questions remain unanswered, there will be numerous opportunities for research and development, both in academia and within national statistical offices.

References

Beaumont, J.-F. (2005). On the use of data collection process information for the treatment of unit nonresponse through weight adjustment. *Survey Methodology*, 31(2), 227-231. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9049-eng.pdf.

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology*, 50(1), 77-106. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00004-eng.pdf.

Breiman, L. (2001). Random forests. *Machine Learning*, 45, 5-32.

Chen, T., and Guestrin, C. (2016). XGBoost. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining – KDD 16*. ACM Press.

Chen, S., Xu, C. and Cutler, J. (2025). Integrating probability and non-probability samples through deep learning-based mass imputation. *Survey Methodology*, 51(2), (to appear).

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W. and Robins, J. (2018). Double/debiased ML for treatment and structural parameters. *The Econometrics Journal*, 21, 1-68.

Chipman, H.A., George, E.I. and McCulloch, R.E. (2010). BART: Bayesian Additive Regression Trees. *The Annals of Applied Statistics*, 4, 266-298.

Dagdoug, M., and Haziza, D. (2024). Statistical inference for a finite population mean with machine learning-based imputation for missing survey data. *Proceedings: Symposium 2024, The Future of Official Statistics.* Statistics Canada.

Dagdoug, M., Goga, C. and Haziza, D. (2023a). Model-assisted estimation through random forests in finite population sampling. *Journal of the American Statistical Association*, 542, 1234-1251.

Dagdoug, M., Goga, C. and Haziza, D. (2023b). Imputation procedures in surveys using nonparametric and machine learning methods: An empirical comparison. *Journal of Survey Statistics and Methodology*, 11, 141-188.

Dagdoug, M., Goga, C. and Haziza, D. (2025). Statistical inference in the presence of imputed survey data through random forests. *Scandinavian Journal of Statistics*, (to appear).

Deng, Y., and Lumley, T. (2024). Multiple imputation through xgboost. *Journal of Computational and Graphical Statistics*, 33, 352-363.

Ferri-García, R., and Rueda, M.D.M. (2020). Propensity score adjustment using ML classification algorithms to control selection bias in online surveys. *PloS One*, 15, e0231500.

Gelein, B. (2017). Handling missing data with superpopulation model, design-based approach and ML. Unpublished PhD thesis.

Kennedy, E.H. (2023). Semiparametric doubly robust targeted double ML: A review, 2023, arXiv:2203.06469.

Kern, C., Klausch, T. and Kreuter, F. (2019). Tree-based machine learning methods for survey research. *Survey Research Methods*, 13, 73-93.

Kern, C., Li, Y. and Wang, L. (2021). Boosted kernel weighting – Using statistical learning to improve inference from nonprobability samples. *Journal of Survey Statistics and Methodology*, 9, 1088-1113.

Kim, J.K., Brick, J.M., Fuller, W.A. and Kalton, G. (2006). On the bias of the multiple-imputation variance estimator in survey sampling. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(3), 509-521.

Krennmair, P., and Schmid, T. (2022). Flexible domain prediction using mixed effects random forests. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 71, 1865-1894.

Larbi, K., Tsang, J., Haziza, D. and Dagdoug, M. (2025). On the use of machine learning methods for the treatment of unit nonresponse in surveys. *Survey Methodology*, 51(1), 275-303. Paper available at https://www150. statcan.gc.ca/n1/en/pub/12-001-x/2025001/article/00003-eng.pdf.

Lee, D., Zhang, L.C. and Chen, S. (2023). Robust quasi-randomization-based estimation with ensemble learning for missing data. *Scandinavian Journal of Statistics*, 50, 1263-1278.

Little, R.J., and Vartivarian, S. (2005). Does weighting for nonresponse increase the variance of survey means? *Survey Methodology*, 31(2), 161-168. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2005002/article/9046-eng.pdf.

Liu, Y., Gelman, A. and Chen, Q. (2023). Inference from non-random samples using Bayesian ML. *Journal of Survey Statistics and Methodology*, 11, 433-455.

Lohr, S., Hsu, V. and Montaquila, J. (2015). Using classification and regression trees to model survey nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, Alexandria, 2071-2085.

McConville, K.S., and Toth, D. (2019). Automated selection of post-strata using a model-assisted regression tree estimator. *Scandinavian Journal of Statistics*, 46(2), 389-413.

Michal, V., Wakefield, J., Schmidt, A.M., Cavanaugh, A., Robinson, B.E. and Baumgartner, J. (2024). Model-based prediction for small domains using covariates: A comparison of four methods. To appear in *Journal of Survey Statistics and Methodology*.

Montanari, G.E., and Ranalli, M.G. (2005). Nonparametric model calibration in survey sampling. *Journal of American Statistical Association*, 100, 1429-1442.

Nalenz, M., Rodemann, J. and Augustin, T. (2024). Learning debiased regression trees and forests from complex samples. *Machine Learning*, 113, 3379-3398.

Newey, W.K., and Robins, J.R. (2018). Cross-fitting and fast remainder rates for semiparametric estimation. arXiv preprint arXiv:1801.09138.

Opsomer, J.D., and Riddles, M.K. (2025). sCHAID: A tool for constructing nonresponse adjustment cells under a design-based framework. *Survey Methodology*, 51(1), 217-231. Paper available at http://www.statcan.gc.ca/pub/12-001-x/2025001/article/00006-eng.pdf.

Park, S., Kim, J.K. and Kim, K. (2019). A note on propensity score weighting method using paradata in survey sampling. *Survey Methodology*, 45(3), 451-463. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019003/article/00002-eng.pdf.

Phipps, P., and Toth, D. (2012). Analyzing establishment nonresponse using an interpretable regression tree model with linked administrative data. *Annals of Applied Statistics*, 6, 772-794.

Quinlan, J.R. (1992). Learning with Continuous Classes. *Proceedings of Australian Joint Conference on Artificial Intelligence*, Hobart, November 16-18, 1992, 343-348.

Rubin, D.B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, Inc.

Sande, L.S., and Zhang, L.C. (2021). Design-unbiased statistical learning in survey sampling. *Sankhyā A*, 83, 714-744.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18(2), 241-252. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14483-eng.pdf.

Tan, Y.V., Flannagan, C.A.C. and Elliott, M.R. (2019). Robust-squared imputation models using BART. *Journal of Survey Statistics and Methodology*, 7, 465-497.

Toth, D., and McConville, K.S. (2024). Design consistent random forest models for data collected from a complex sample. *Survey Methodology*, 50(2), 185-207. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024002/article/00015-eng.pdf.

Viljanen, M., Meijerink, L., Zwakhals, L. and Van de Kassteele, J. (2022). A ML approach to small area estimation: Predicting the health, housing and well-being of the population of Netherlands. *International Journal of Health Geographics*, 21. https://doi.org/10.1186/s12942-022-00304-5.

Wang, Z., Akande, O., Poulos, J. and Li, F. (2022). Are deep learning models superior for missing data imputation in surveys? Evidence from an empirical comparison. *Survey Methodology*, 48(2), 375-399. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022002/article/00009-eng.pdf.

Wieczorek, J. (2023). Design-based conformal prediction. *Survey Methodology*, 49(2), 443-473. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2023002/article/00007-eng.pdf.