Catalogue no. 12-206-X
ISSN 1705-0820

Statistical Methodology Research and Development Program Achievements, 2023-2024

Release date: October 25, 2024

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, Statistics Canada has developed standards of service that its employees observe. 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.

Statistical Methodology Research and Development Program

Achievements, 2023-2024

This report summarizes the 2023-2024 achievements of the Methodology Research and Development Program (MRDP) sponsored by the Modern Statistical Methods and Data Science Branch at Statistics Canada. This program covers research and development activities in statistical and data science methods with potentially broad application in the agency's statistical programs; these activities would otherwise be less likely to be carried out during the provision of regular methodology services to those programs. The MRDP also includes activities that provide client support in the application of past successful developments in order to promote the use of the results of research and development work. Contact names are provided for obtaining more information on any of the projects described. For more information on the MRDP as a whole, please contact:

Jean-François Beaumont
(Email : jean-francois.beaumont@statcan.gc.ca)

Statistical Methodology Research and Development Program

Achievements, 2023-2024

Table of Contents

1 Data integration

1.1 Integration of probability and non-probability samples

PROJECT: Parametric estimation of participation probabilities for non-probability samples

Non-probability samples are currently being explored at Statistics Canada and other National Statistical Offices as an alternative to probability samples. However, it is well known that the use of a non-probability sample alone may produce estimates with significant bias due to the unknown nature of the underlying selection mechanism. To reduce this bias, data from a non-probability sample can be integrated with data from a probability sample provided that both samples contain auxiliary variables in common.

In this research, we focused on parametric estimation of the probability of participation in the non-probability sample, which is a key quantity to reduce participation bias. Several methods have recently been proposed to estimate the participation probability (e.g., Elliott, 2009; Chen, Li and Wu, 2020; and Wang, Valliant and Li, 2021). The pseudo likelihood method of Chen, Li and Wu (2020) has attracted the attention of many researchers, but it was recently found to be inefficient in some scenarios. The methods of Elliott (2009) and Wang, Valliant and Li (2021) may sometimes be more efficient, but they fail to reduce to the maximum likelihood method when the probability sample is a census.

The problem of estimating participation probabilities is similar to the problem of estimating probabilities of response to surveys, the main difference being the structure of the available information. The goal of this research was to find principled ways of leveraging auxiliary data from both the probability and non-probability samples.

Progress:

In the first part of the project, we developed an improvement of the method of Chen, Li and Wu (2020), based on best linear unbiased estimation theory, that more efficiently leverages the available probability and non-probability sample data. We also developed a sample likelihood approach, similar in spirit to the method of Elliott (2009), that properly accounts for the overlap between both samples when it can be identified in at least one of the samples. We used best linear unbiased prediction theory to handle the scenario where the overlap is unknown. Interestingly, our two proposed approaches coincide in the case of unknown overlap. Then, we showed that many existing methods can be obtained as a special case of a general unbiased estimating function. These theoretical results were documented and published in the June 2024 issue of *Survey Methodology* (Beaumont, Bosa, Brennan, Charlebois and Chu, 2024). This paper is a rejoinder to comments on a paper by the same authors entitled “Handling non-probability samples through inverse probability weighting with an application to Statistics Canada’s crowdsourcing data”, also published in the June 2024 issue of *Survey Methodology*.

In the second part of the project, we conducted simulation experiments that confirmed the efficiency gains of our proposed methods. We plan to document our findings in the fiscal year 2024-2025.

For more information, please contact:
Jean-François Beaumont (jean-francois.beaumont@statcan.gc.ca).

References

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Authors' response to comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data": Some new developments on likelihood approaches to estimation of participation probabilities for non-probability samples. *Survey Methodology*, 50, 1, 123-141. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00001-eng.pdf.

Chen, Y., Li, P. and Wu, C. (2020). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association*, 115, 2011-2021.

Elliott, M.R. (2009). Combining data from probability and non-probability samples using pseudo-weights. *Survey Practice*, 2, 813-845.

Wang, L., Valliant, R. and Li, Y. (2021). Adjusted logistic propensity weighting methods for population inference using nonprobability volunteer-based epidemiologic cohorts. *Statistics in Medicine*, 40, 5237-5250.

1.2 Record linkage

PROJECT: Capture-recapture with linkage errors

To prioritize the use of administrative sources, Statistics Canada is aiming to measure their coverage through field operations or clerical reviews, which may be expensive. The capture-recapture method may offer a cost-effective solution through a comparison to another source under standard assumptions, which include the independence of the sources and their perfect linkage. However, adaptations are required where the linkage is imperfect. So far, those refinements have required clerical-reviews (Ding and Fienberg, 1994; Di Consiglio and Tuoto, 2015; de Wolf, van der Laan and Zult, 2019) or the strong assumption that the linkage variables are conditionally independent (Racinskij, Smith and van der Heijden, 2019).

Progress:

A new method for capture-recapture estimation with linkage errors was proposed, which does not suffer from the above limitations. It operates by modeling the number of links from a record and builds on the statistical model previously described by Dasylva and Goussanou (2022). The methodology is described in a paper accepted for publication in Survey Methodology (Dasylva, Goussanou and Nambeu, 2024).

For more information, please contact:
Abel Dasylva (abel.dasylva@statcan.gc.ca).

References

Dasylva, A., and Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*. Retrieved from https://doi.org/10.1007/s42081-022-00153-3.

Dasylva, A., Goussanou, A. and Nambeu, C.-O. (2024). Models of linkage error for capture-recapture estimation without clerical-reviews. *Survey Methodology*, 50, 2. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024002/article/00007-eng.pdf.

de Wolf, P.-P., van der Laan, J. and Zult, D. (2019). Connection correction methods for linkage error in capture-recapture. *Journal of Official Statistics*, 35, 577-597.

Di Consiglio, L., and Tuoto, T. (2015). Coverage Evaluation on Probabilistically Linked Data. *Journal of Official Statistics*, 31, 415-429.

Ding, Y., and Fienberg, S.E. (1994). Dual system estimation of census undercount in the presence of matching error. *Survey Methodology*, 20, 2, 149-158. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1994002/article/14422-eng.pdf.

Racinskij, V., Smith, P. and van der Heijden, P. (2019). Linkage free dual system estimation. Available at https://arxiv.org/abs/1903.10894.

PROJECT: Measuring the linkage accuracy when many files are linked to a spine

To perform a statistical analysis, it may be necessary to link many files to a spine, which is essentially a list of unique units from a target population. However, the resulting links may be imperfect, where the identifying information is partial or erroneous. In such cases, the linkage accuracy must be evaluated based on record tuples with three or more records. Solutions based on clerical reviews are described by Binette, Baek, Engineer, Jones, Dasylva, and Reiter (2024). Yet, this approach is potentially costly. An alternative approach may be based on the extension of the Fellegi-Sunter model by Sadinle and Fienberg (2013) or Bayesian entity resolution models by Steorts, Hall, and Fienberg (2016), Marchant, Kaplan, Elazar, Rubinstein and Steorts (2021) or others. However, these models rely on the strong assumption of conditional independence (Fellegi and Sunter, 1969), or variations of this assumption.

Progress:

A new methodology is proposed, which dispenses with clerical reviews and the conditional independence assumption. In this methodology, the recall and precision are extended to record tuples with three or more records, and the resulting measures are evaluated by modeling the vector including the number of links to each file for each spine record. The actual model extends the one previously described by Dasylva and Goussanou (2022) into a finite multivariate mixture, where each component is the convolution of a multinomial distribution and an independent multivariate compound Poisson distribution. In addition to informing users of the linked data, the methodology has many other practical applications such as evaluating the reliability of mean that is based on the linked data, e.g., with a measure that is akin to a coefficient of variation, i.e., a CV-like measure. This is useful when determining if the linkage errors can be safely ignored or if further steps are required. The methodology also provides a basis for evaluating a spine, which is itself based on the imperfect linkage of many administrative files, such as the likelihood that many records actually represent the same person given that they are classified as such. See Dasylva and Goussanou (2024a) for details.

For more information, please contact:
Abel Dasylva (abel.dasylva@statcan.gc.ca).

References

Binette, O., Baek, Y., Engineer, S., Jones, C., Dasylva, A. and Reiter, J. (2024). How to evaluate entity resolution systems: An entity-centric framework with application to inventor name disambiguation. Available at: https://arxiv.org/pdf/2404.05622.pdf.

Dasylva, A., and Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*. Retrieved from https://doi.org/10.1007/s42081-022-00153-3.

Dasylva, A., and Goussanou, A. (2024a). Measuring the linkage accuracy when many files are linked to a spine. Internal report, Statistics Canada.

Fellegi, I., and Sunter, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Marchant, N., Kaplan, A., Elazar, D., Rubinstein, B. and Steorts, R. (2021). d-blink: Distributed end-to-end bayesian entity resolution. *Journal of Computational and Graphical Statistics*, 30, 406-421.

Sadinle, M., and Fienberg, S.E. (2013). A generalized Fellegi-Sunter framework for multiple record linkage with application to homicide record systems. *Journal of the American Statistical Association*, 108, 385-397.

Steorts, R., Hall, R. and Fienberg, S. (2016). A bayesian approach to graphical record linkage and de-duplication. *Journal of the American Statistical Association*, 111, 1660-1672.

PROJECT: Estimation of small area means with linked data

Small area estimation techniques are routinely used to produce estimates for areas where the sample is too small, by modeling the variable of interest based on auxiliary information (Rao and Molina, 2015). With a unit level model, this information must be obtained for each sampled unit, and this involves a record linkage if the auxiliary information is located on a separate file. In general, this linkage may be a source of errors that must be accounted for. Salvati, Fabrizi, Ranalli and Chambers (2021) have proposed a solution for applying the linear nested error model, where the auxiliary variables are obtained from a register, which is distinct from the sampling frame and contains an indicator of sample inclusion that is seldom available in practice. They also require clerical estimates of the rates of linkage error, which may be expensive to source.

Progress:

To overcome the above limitations, a solution is proposed that is based on extending the methodology from Salvati, Fabrizi, Ranalli and Chambers (2021), with the help of the model of linkage error described by Dasylva and Goussanou (2022). This methodology was presented at the Joint Statistical Meetings in 2023. See Dasylva (2024) for details.

For more information, please contact:
Abel Dasylva (abel.dasylva@statcan.gc.ca).

References

Dasylva, A. (2024). Estimation of small area means with linked data. Internal report, Statistics Canada.

Dasylva, A., and Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*, 5, 181-216. https://doi.org/10.1007/s42081-022-00153-3.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. New Jersey: Wiley.

Salvati, N., Fabrizi, E., Ranalli, M. and Chambers, R. (2021). Small area estimation with linked data. *Journal of the Royal Statistical Society B*, 83, 78-107.

PROJECT: Model-based threshold selection for agricultural probabilistic linkages

With the increasing importance of administrative data usage in producing official statistics, the problem of matching records without a unique identifier has also become increasingly important. The probabilistic method proposed by Fellegi and Sunter (1969) is a common approach, which is implemented in the generalized system for probabilistic linkage at Statistics Canada, GLINK. However, the weight threshold, a parameter which the optimality of the procedure relies on, can prove challenging to set. Solutions for setting the threshold up to now have had major limitations, either relying on optimistic assumptions, or requiring training data or manual review, both of which can be costly to source. However, a model was developed which estimates the linkage error based on the number of links from a given record, accounting for all the interactions among the linkage variables while mitigating the need for any manual process (Dasylva and Goussanou, 2022).

Progress:

Building on Dasylva and Goussanou's error model, different algorithms were evaluated to select the linkage threshold. Between exhaustive search, binary search, or a more sophisticated recursive partitioning procedure proposed by Dasylva and Chen (2022), each method offers distinct advantages in terms of runtime and the quality metrics produced. The methodology was also compared to the method proposed by Belin and Rubin (1995) and the extreme value theory approach (Sariyar, Borg and Pommerening, 2011) using simulated data to show the results of using the model when the ground truth was known. In both cases, Dasylva and Goussanou's error model more accurately estimated the precision of the linkage. Lastly, following Dasylva and Goussanou's approach (2024b), chi-square goodness of fit tests were applied to farm data linkage pairs to ensure that the model applies to real data. For all the tested linkages, the data passed the goodness of fit tests, showing that the new error model and threshold selection method prove useful in a practical setting.

For more information, please contact:
Christian Arsenault (christian.arsenault@statcan.gc.ca) or
Abel Dasylva (abel.dasylva@statcan.gc.ca).

References

Belin, T., and Rubin, D. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 694-707.

Dasylva, A., and Chen, W. (2022). Probabilistic record linkage through recursive partitioning without training data. Presentation at the monthly meeting of the ONS-UNECE Machine Learning group, April 2022.

Dasylva, A., and Goussanou, A. (2022). On the consistent estimation of linkage errors without training data. *Japanese Journal of Statistics and Data Science*. Available at: https://doi.org/10.1007/s42081-022-00153-3.

Dasylva, A., and Goussanou, A. (2024b). Making statistical inferences about linkage errors. *Japanese Journal of Statistics and Data Science*. Available at: https://doi.org/10.1007/s42081-023-00228-9.

Fellegi, I., and Sunter, A. (1969). A theory of record linkage. *Journal of the American Statistical Association*, 64, 1183-1210.

Sariyar, M., Borg, A. and Pommerening, K. (2011). Controlling false match rates in record linkage using extreme value theory. *Journal of Biomedical Informatics*, 44, 648-654.

PROJECT: Secondary analysis of linked categorical data

It is increasingly common for analysts working with Statistics Canada data to utilize record linkage to add depth to survey files. For instance, social surveys are frequently linked to tax and immigration data through the Social Data Linkage Environment (SDLE). The use of linked data serves to reduce response burden and provides analysts more accurate values for certain variables (e.g. income) than what could be obtained directly from respondents.

However, many linkages within the SDLE are probabilistic, and linkage error is an inevitable by-product of probabilistic record linkage. Analysis of linked data that does not correct for the presence of linkage error may lead to biased results.

Most teams making use of Statistics Canada linked data are 'secondary' users of the data. This means that they have access to analytic variables on their files of interest and are provided with a set of linkage keys joining the files together, but they do not have access to the micro-data variables (e.g. names, birthdates, etc.) that were used to perform the linkages. In order to protect Canadians' sensitive information, employees or external researchers granted access to analytic variables are not generally granted access to linkage variables.

Recently, Li-Chun Zhang and Tiziana Tuoto (2021) proposed a promising new method of incorporating linkage error into the analysis of linked data. At the 2022 Statistics Canada's International Methodology Symposium, Li-Chun Zhang (2022) presented an approach they had developed to perform logistic regression from a secondary analyst perspective in adjusting for false matches. The primary goal of this project was to evaluate this approach and to develop programs to implement it on files linked through the SDLE.

Progress:

Zhang and Tuoto's algorithm was implemented in Python, and a simulation study was performed to compare its performance to that of naïve logistic regression. Artificial versions of SDLE linkages were created by sampling from the linked set of an actual SDLE linkage. The linkage weights available on the artificial linkage files were refined using the expectation maximization algorithm (a technique first suggested for record linkage by Winkler (2000) and used to help simulate the SDLE process for estimating the proportion of false matches for a given linkage. Finally, simulated categorical, analytic variables were added to each artificial linkage. These variables were inspired by the variables from an internal Statistics Canada analysis of linked data. The artificial sets of linked analytic data (and estimated error-rates) were fed into the Python program, and the performance of the algorithm was evaluated.

In some simulations, Zhang and Tuoto's approach led to reduced bias and slightly inflated variance, compared to what is obtained through naïve logistic regression. These results likely represent what would occur in a 'typical' case.

However, in certain other sets of simulations in which highly unbalanced covariates were present in the artificial sets of linked data, there were cases in which the standard naïve regression algorithm converged but Zhang and Tuoto's method failed to converge. A number of standard fixes were attempted, and internal partners were consulted to resolve this issue, but no solution was found. Zhang and Tuoto themselves wound up providing a helpful suggestion. They proposed using certain standard R packages (such as glm or glmet) to implement an iterative weighted least squares approach to estimate the regression coefficients. This approach is currently under investigation.

Finally, it should be noted that one of the (optional) inputs (the estimated variance of the false match rate estimator) to the Python function written to implement Zhang and Tuoto's method is not a standard output produced by the SDLE, but could be produced using a SAS program developed as part of a research project conducted in the previous fiscal year (Loewen and Millar, 2023).

The outputs from this project include a Python program for implementing Zhang and Tuoto's method with SDLE data that will be made available to users through the Record Linkage Resource Centre, an internship report by Toukal (2023), and a conference presentation by Millar (2024).

This project was done in collaboration with Julia Toukal and Abdelnasser Saidi, with the help of Li-Chun Zhang, Tiziana Tuoto, Abel Dasylva, Mark Stinner, and Kenza Sallier.

For more information, please contact:
Goldwyn Millar (goldwyn.millar@statcan.gc.ca).

References

Loewen, R., and Millar, G. (2023). Variance estimation for record linkage error-rates obtained via clerical review of stratified systematic samples of linked pairs. Presentation at the Methodology Seminar on May 10th, 2023, Statistics Canada, Ottawa.

Millar, G. (2024). Logistic regression on linked data from a secondary analyst perspective. Presentation at CANSSI-CRT Workshop on Modern Methods in Survey Sampling, University of Ottawa, July 8-10.

Toukal, J. (2023). Logistic regression in the context of record linkage. Internship report for l'École Nationale de la Statistique et de l'analyse de l'information.

Winkler, W. (2000). *Using the EM Algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage*. Bureau of the Census, Statistical Research Division, Statistical Research Report Series, No. RR2000/05.

Zhang, L.-C. (2022). Secondary analysis of Linked categorical data. Presentation at the 2022 Statistics Canada's International Methodology Symposium.

Zhang, L.-C., and Tuoto, T. (2021). Linkage-data linear regression. *Journal of the Royal Statistical Society, Series A*, 184, 522-554.

PROJECT: Machine learning blocking methods for record linkage: Application of the Sequential Covering Algorithm (SCA) and the High-Value Token-Blocking (HVTB) methods to the Visitors Record linkage project done by the Social Data Linkage Environment (SDLE) of Statistics Canada

A crucial pre-processing step in Record Linkage (RL) is to choose a blocking technique that ensures that the number of pairs of records to compare is computationally manageable while at the same time covers as many true matches as possible. An effective blocking technique aims at fulfilling both goals.

This project aims at evaluating two machine learning blocking methods using an example of record linkage done at Statistics Canada: 1) The Sequential Covering Algorithm (SCA) presented in Michelson and Knoblock (2006); 2) The High-Value Token-Blocking (HVTB) of O'Hare, Jurek-Loughrey, and De Campos (2021). Intuitively, SCA learns a conjunction of attributes, called a rule, that covers some set of positive examples, and learns another conjunction of attributes, repeating this step until it can no longer discover a rule with performance above a chosen threshold for the metric called Pairs Completeness (PC). PC measures the coverage of true positives, i.e. how many of the true matches are in the candidate set of pairs of records to compare according to the rules identified versus those in the entire set. HVTB is based on the well-known Term Frequency-Inverse Document Frequency (TF-IDF) and its ability to identify high-value tokens, i.e. a term or a word in a document. TF-IDF values indicate the significance of a token based on their commonality within a dataset, with lower values indicating highly frequent insignificant tokens, and higher values indicating rarer significant tokens.

Progress:

An evaluation of the performance of the algorithms was done on some small, medium, and large datasets before testing them on an example of record linkage that is done at Statistics Canada. Pairs Completeness (PC) and Reduction Ratio (RR) were used as metrics for the evaluation. RR measures how well a blocking method minimizes the number of candidates pairs to compare to catch matching records. The evaluation revealed that the algorithms performed well for small and medium size datasets, but required the use of techniques such as partitioning, parallel processing and sampling to deal with large datasets. Then, they were evaluated on a Statistics Canada's record linkage project that aims at matching records from a file containing information about visitors to Canada to another file called the Depot to identify matches. The goal of this evaluation was to determine if these algorithms provide better performance than the current blocking method used for this linkage. SCA performed well with a RR above 99% and a PC of 83%, which is similar to the current blocking method performance. However, the current method creates thousand times more candidate pairs to compare to catch matching records. Regarding the HVTB, the

implementation of the sampling approach is underway. A few more adjustments are required to make it efficient in terms of running time.

The next steps will be the following:

1) Finalize the implementation of the sampling approach for HVTB.
2) Determine a strategy for using SCA when a record linkage is being done for the first time, since it requires labelled data.

For more information, please contact:
Ronald Jean Paul (ronald.jeanpaul@statcan.gc.ca) or
Bassirou Diagne (bassirou.diagne@statcan.gc.ca).

References

Michelson, M., and Knoblock, C.A. (2006). Learning blocking schemes for record linkage. *Proceedings of the Association for Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 21, 440-445. Available at: https://www.aaai.org/Papers/AAAI/2006/AAAI06-070.pdf.

O'Hare, K., Jurek-Loughrey, A. and De Campos, C. (2021). High-value token-blocking: Efficient blocking method for record linkage. *ACM Transactions on Knowledge Discovery from Data*, 16, 1-17. Available at: https://doi.org/10.1145/3450527.

1.3 Small area estimation

PROJECT: The use of random forests in small area estimation

When domain sample sizes are small, design-consistent direct estimators of population parameters are likely to be unstable. To improve the precision of direct estimators, the Fay-Herriot area level model is often used. It has two components: a sampling model and a linking model. The latter specifies the relationship between the population parameters of interest and auxiliary variables available at the domain level. In its original form, the Fay-Herriot model assumes a linear linking model with constant error variance. It also requires estimating the smooth design variance of direct estimators, i.e., the model expectation of the design variance of direct estimators. Design-based variance estimators can be considered as estimators of the smooth design variances, but they are typically unstable for small sample sizes. To solve this problem, design-based variance estimates are usually smoothed, often using a log-linear smoothing model.

The assumptions underlying the Fay-Herriot and smoothing models are not always satisfied in practice, and it may be difficult and time-consuming to adequately correct the models. In this context, it may be desirable to have access to non-parametric methods, especially when the number of domains is large, because they depend less heavily on the validity of model assumptions and may speed up the production of small area estimates. We are particularly interested in random forests for three reasons: i) they can be easily applied to the case of a mixture of categorical and continuous auxiliary variables, ii) they do not require specifying interactions, and iii) they produce predictions that always remain within the range of observed values. We consider a bootstrap procedure for the estimation of the mean square prediction error.

Progress:

In the previous year, we developed non-parametric versions of the Empirical Best (EB) predictor when random forests are used to replace fully parametric models. In 2023-2024, we refined our methodology and evaluated the properties of our proposed EB predictors using real data and through simulation studies. The retained predictor uses out-of-bag predictions as an auxiliary variable in a linear Fay-Herriot model. Our results show that random forests offer robustness to model misspecifications, increase the efficiency of small area estimates and simplify (but do not eliminate) the modelling effort. This project will be presented at the International Conference on Establishment Statistics in Glasgow in June 2024 and at the CANSSI conference in Ottawa in July 2024. In 2024-2025, we plan to write a paper that summarizes our results and develop an R program that implements our methodology.

For more information, please contact:
Jean-François Beaumont (jean-francois.beaumont@statcan.gc.ca).

PROJECT: HB inference for small area estimation using non-informative and informative priors

The Fay-Herriot (F-H) area level model is often used when direct estimators of population parameters are unstable because of small sample sizes. The idea behind the F-H model and other Small Area Estimation (SAE) models is to borrow strength from other domains. However, when the number of domains is small, the F-H and other SAE models are usually not performing well. The idea of this project is to borrow strength from other domains but also across time using a Hierarchical Bayes (HB) approach.

HB modeling is very popular in small area estimation, and prior specification is very important in the HB modeling approach. In this project, we study the performance of HB small area estimators using non-informative and informative priors for the regression parameters and variance components. We apply the Bayesian models of You and Chapman (2006) and You (2021) to the Canadian Labor Force Survey (LFS) data and evaluate the impact of the priors on the HB estimators. We study the impact of correct/incorrect informative priors and non-informative priors for HB small area estimation using both a LFS application and a simulation study.

Progress:

For variance components, we have studied prior specifications using inverse gamma and flat priors for HB small area models. We conducted a simulation study and applied the models to LFS data application. A research paper (You, 2023) has been published in Statistics in Transition new series.

For regression parameters, we investigated the use of noninformative and informative priors through LFS application and a simulation study. A research paper (You and Bosa, 2024) has been finished and submitted to a journal for possible publication. The paper is currently under revision.

For more information, please contact:
Yong You (yong.you@statcan.gc.ca).

References

You, Y. (2021). Small area estimation using Fay-Herriot area level model with sampling variance smoothing and modeling. *Survey Methodology*, 47, 2, 361-370. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00007-eng.pdf.

You, Y. (2023). An empirical study of hierarchical Bayes small area estimators using different priors for model variances. *Statistics in Transition New Series*, 24, 169-178.

You, Y., and Bosa, K. (2024). Performance of hierarchical Bayes small area estimators using non-informative and informative priors with LFS application. Submitted to *Survey Methodology* (under revision).

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 1, 97-103. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9263-eng.pdf.

PROJECT: Evaluation of small area confidence interval estimation

In this project, we study and evaluate confidence interval estimation for small area estimation using Empirical Best Linear Unbiased Predictors (EBLUP). In particular, we study the Wald confidence interval and the modified Wilson confidence interval for small area proportion estimation using an application to Labour Force Survey data and a simulation study.

Progress:

The results from the real data analysis and simulation study show that both the Wald and modified Wilson confidence intervals perform very well in small area estimation using EBLUPs. We have completed the project, and an internal research report has been written (You and Hidiroglou, 2024).

For more information, please contact:
Yong You (yong.you@statcan.gc.ca).

Reference

You, Y., and Hidiroglou, M. (2024). Empirical study of confidence intervals for small area proportion estimation with LFS application. Internal report, Statistics Canada.

2 Data science methods and applications

PROJECT: Multiparty privacy-preserving record linkage

Record linkage is a process combining multiple sources of tabular data that report on at least some of the same entities. Sometimes this process requires the cooperation of several parties, who all hold privacy concerns regarding the individuals featured within their data, and who may not hold a high degree of mutual trust. Therefore, methods of performing record linkages while preserving the privacy of those

within the data, by not leaking any information beyond the results of the required computations, are a valuable avenue of research. In this work, we investigated a case study where a National Statistics Organization wants to collaborate with two other organizations to compute aggregates while preserving privacy.

Progress:

Our protocol makes use of oblivious programmable pseudo-random functions, and secure circuit-based computing. This work is an extension of a two-party protocol outlined in (Dugdale, Santos and Zanussi, 2023) and is based on Private Set Intersection (PSI), where two parties, each in possession of a private dataset, aim to obtain the common elements (or identifiers) on both datasets, i.e., compute the intersection. Privacy-Preserving Record Linkage (PPRL) is related to PSI in the sense that PPRL aims to compute aggregations on the attributes belonging to the common identifiers found after the intersection. These aggregates can be quite simple, such as computing the cardinality of the linked set, or more complex, such as weighted sums based on the attributes.

We started researching multiparty extensions on PSI protocols that we could extend to manage aggregations. We found that Circuit PSI (Chandran, Gupta and Shah, 2022) enables computation over the intersection of two sets using Secure Multiparty Computation (SMPC). Moreover, this protocol was extended to treat multiple parties owning a dataset (Chandran, Dasgupta, Gupta, Obbattu, Sekar and Shah, 2021). We configured and compiled the existing implementation in a docker container to compute the intersection over three datasets. Then we outlined a protocol using a secondary SMPC circuit that enables the computation of aggregations on the attributes of the intersection (Santos, 2023). We found that the multi-party extension to PPRL is more complex and stiffer, meaning the solution must be tailored to the problem of study: nature of datasets and desired aggregations. Finally, these cryptographic protocols are built on codebases from several research groups, and this makes implementations hard to audit, modify, improve, deploy, and to document.

For more information, please contact:
Benjamin Santos (benjamin.santos@statcan.gc.ca).

References

Chandran, N., Dasgupta, N., Gupta, D., Obbattu, S.L.B., Sekar, S. and Shah, A. (2021). Efficient linear multiparty psi and extensions to circuit/quorum psi. *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, 1182-1204.

Chandran, N., Gupta, D. and Shah, A. (2022). Circuit-PSI with linear complexity via relaxed batch OPPRF. *Proceedings on Privacy Enhancing Technologies*.

Dugdale, C., Santos, B. and Zanussi, Z. (2023). Practical Privacy-Aware Data Linkage and Statistical Aggregation based on Privacy Enhancing Techniques. Conference on New Techniques and Technologies for Official Statistics, Brussels, Belgium.

Santos, B. (2023). Multi-Party Privacy Preserving Record Linkage based on Circuit Private Set Intersection. Canadian Mathematical Society Winter Meeting, Montréal, Canada.

PROJECT: Generation of synthetic tabular data from diffusion models

One of the primary goals of a National Statistical Organization (NSO) is the dissemination of data to the public and researchers while providing privacy, confidentiality, and security requirements (Statistics Canada Data Strategy, 2022). Traditionally, NSOs apply methods of statistical disclosure control (SDC) to the data to reduce the risks of disclosure of sensitive information when disseminating it. Dissemination of synthetic data (SD) has been proposed as an alternative to SDC, in this case a trained model transforms the original data into a synthetic one.

Diffusion models are suitable candidates for generating synthetic tabular data (Kotelnikov, Baranchuk, Rubachev, and Babenko, 2022). However, an assessment on the trade-off between data privacy and data utility should be made to validate these models. Our goals were i) to implement or modify existing diffusion models for tabular data, ii) to perform evaluation metrics, iii) to address the privacy and confidentiality aspects, and iv) to assess the trade-off between data privacy and data utility.

Progress:

First, we started investigating diffusion models for tabular data as well as other approaches, including existing packages for synthetic data generation. Next, we developed a preprocessing module to transform the tabular data into suitable form to be used for training the generators. We adapted an implementation of TabDDPM (Kotelnikov et al., 2022), which is a diffusion model for synthetic tabular data generation. We made substantial progress in implementing a dashboard for easy tabular data manipulation and preprocessing, for training the synthetic data generators, and to output quality and privacy metrics.

In the experimentation phase, we compared the performance of Generative Adversarial Networks (GAN, Figueira and Vaz, 2022) generators against TabDDPM. We found that the quality of the synthetic data generated by the diffusion model was superior in almost all metrics. Moreover, TabDDPM performed very well for privacy protection, although the mechanism of adding privacy is not completely understood. In that regard, Differentially Private GAN offers better control over the privacy-utility trade-off. However, recent advancements in Differentially Private Diffusion Models along with improved models (Truda, 2023; Zhang, Zhang, Srinivasan, Shen, Qin, Faloutsos and Karypis, 2023) open the door for more investigation on the usability of these generators on synthetic tabular data generation while preserving privacy and utility. Implementing and assessing Differentially Private Diffusion Models is our next step.

For more information, please contact:
Benjamin Santos (benjamin.santos@statcan.gc.ca).

References

Figueira, A., and Vaz, B. (2022). Survey on synthetic data generation, evaluation methods and GANs. *Mathematics*, 10(15), 2733. MDPI AG. Retrieved from https://www.mdpi.com/2227-7390/10/15/2733.

Kotelnikov, A., Baranchuk, D., Rubachev, I. and Babenko, A. (2022). TabDDPM: Modelling Tabular Data with Diffusion Models. arXiv preprint arXiv:2209.15421.

Statistics Canada Data Strategy (2022). Available at: https://www.statcan.gc.ca/en/about/datastrategy.

Truda, G. (2023). Generating tabular datasets under differential privacy. arXiv preprint arXiv:2308.14784.

Zhang, H., Zhang, J., Srinivasan, B., Shen, Z., Qin, X., Faloutsos, C. and Karypis, G. (2023). Mixed-type tabular data synthesis with score-based diffusion in latent space. arXiv preprint arXiv:2310.09656.

PROJECT: Functional encryption

Cryptographic schemes are the standard for protecting data in transit and at rest. In addition to schemes, which remove the utility of the ciphertext until it is decrypted, there are other schemes which permit the data to be utilized while encrypted. Functional Encryption is a cryptographic scheme which allows functions to be applied on encrypted datasets without decrypting the data. The result from the applied function is decrypted, allowing authorized users to obtain plaintext function outputs applied on encrypted data without compromising security. Individual keys can be generated for different functions such that only authorized parties can utilize and observe the plaintext function outputs. This project aims to explore different uses of Functional Encryption from the perspective of a National Statistical Office. For instance, rather than requiring access to be manually granted and approved to fully access sensitive encrypted datasets, access can be provided to subsets of datasets based on the user's privileges and attributes.

Progress:

Following a survey on Functional Encryption, tests have been conducted to explore how the techniques can be practically applied. First, the use of Functional Encryption to provide partial views of an encrypted dataset has been explored with an open dataset on employment income statistics by major field of study and highest level of education. With the dataset encrypted, seven Boolean expressions have been defined such that users with the appropriate combination of roles can view certain combinations of columns of the dataset in plaintext while the rest remain as ciphertext. These rules rely on the attributes assigned to a user, such as a privileged user, where a general user may only see encrypted data by default. This can provide real-time partial access to encrypted datasets without requiring individual requests to be approved or declined for full access.

In a second experiment, Functional Encryption has been utilized on a mock dataset of location data for metro connections. Here, a user's movement is tracked and encrypted on their mobile device. Rather than request a user to send their full data, the users can provide keys to allow only the sums of location visits to be retrieved and used to generate a heatmap. During a public health crisis for example, this could provide privacy-preserving heatmaps to identify hotspots without sacrificing an individual's privacy. Both experiments exhibit strong potential for its usage when working with encrypted data.

For more information please contact:
Julian Templeton (julian.templeton@statcan.gc.ca).

PROJECT: Machine learning for imputation

Imputation is a key step across all survey programs at Statistics Canada. Harnessing the power of modern supervised learning methods in the imputation process could improve the quality of data and resulting estimates. We explore the use of machine learning for imputation and compare its performance to traditional methods, specifically in the context of predictive mean matching (PMM). PMM typically employs linear regression as its underlying model and use predicted values as the single matching variable for donor imputation. Since PMM does not rely directly on the outputs of the regression model, it is

generally considered robust to model misspecification. Here, we replace the linear regression with a machine learning model and assess performance through a simulation study.

Progress:

Using a subset of variables from the 2016 Census Public Use Microdata File (Statistics Canada, 2023), we construct a complete population dataset, from which we draw samples for our simulation study. Within each sample, we generate a fixed proportion of missing values for a single variable under a Missing at Random (MAR) mechanism and use the remaining variables to fit imputation models. We fit various machine learning models, selecting the model with the most predictive power through cross-validation for use in our simulation study. We measure the performance of estimators of the population mean and median by examining bias, empirical standard error, and coverage.

With this study design, we find that the use of machine learning models with PMM yields marginal improvements in bias, coverage, and efficiency. While the improvements are modest, they highlight the potential of machine learning methods to enhance imputation processes. Additionally, the results demonstrate the effectiveness of PMM as an imputation method. To fully evaluate the effectiveness of machine learning models for imputation, further investigation is warranted. We recommend conducting additional simulation studies that consider various missingness mechanisms, levels of missingness, and different types of data.

For more information please contact:
Alden Chen (alden.chen@statcan.gc.ca).

Reference

Statistics Canada (2023). 2016 Census Public Use Microdata File (PUMF), hierarchical file. Available at: https://www150.statcan.gc.ca/n1/en/catalogue/98M0002X.

PROJECT: Masked AutoEncoders for generalized imputation of tabular data

The need to address and handle missing data is ubiquitous within Statistics Canada. Missing data can arise in many forms, such as survey non-response, and can hinder the accuracy and reliability of statistical inferences, often necessitating the use of imputation techniques and heuristics. Imputation approaches may suffer from requiring extensive domain knowledge or may lack sufficient fidelity.

A set of modern state-of-the-art machine learning paradigms (masked language/image modeling) fit models to reconstruct a randomly sampled masked (missing) portion of the input, resulting in models that are capable of performing arbitrary downstream (text/image) tasks. Masked AutoEncoders (MAE's) are one such approach. Recognizing that these paradigms are simply optimizing for imputation, this project applies masked autoencoders to tabular datasets, with the aim of producing a generalized imputation model that once fit using this paradigm, can perform high-fidelity imputation on arbitrary missing features from a given tabular dataset. The approach is simple, general in nature and requires no domain knowledge of the dataset.

Progress:

A transformers-based architecture was used to implement the MAE imputation models. Extensive testing against 15 established imputation baselines, ranging from classical techniques to machine learning and deep learning-based approaches, across 25 diverse tabular datasets.

The results consistently showed that MAE's were the strongest imputation technique as compared to the 15 baselines tested. When tested against datasets with 30% missing values (independently and randomly sampled), our approach resulted in the highest ranking for categorical feature accuracy, numerical feature mean absolute error, and categorical feature Wasserstein distance, and 2nd highest ranking for numerical feature Wasserstein distance. Amongst the imputation techniques tested, MAE's consistently were the most robust to the amount of missing data: though our approach fit models using 30% missing data, MAE's showed strong performance when evaluated on data with as much as 50% missing values and was consistently more robust than other baselines when evaluated on as much as 90% missing values. The results clearly demonstrate that MAE's for generalized tabular imputation is a viable imputation technique that can be applied to missing data for all variables within a particular tabular dataset and requires no domain expertise.

For more information, please contact:
Nicholas Denis (nicholas.denis2@statcan.gc.ca).

PROJECT: Time and effort optimization in data processing

The Time and Effort Optimization research project is an investigative study aimed at enhancing the efficiency of microdata preprocessing in survey data at Statistics Canada. The project's focus is on the robustness of data preprocessing for complex tasks and the performance of an imputation method under varying conditions. This research is designed to optimize the timeliness and reduce manual efforts in the Monthly Electricity Supply and Disposition Survey (MELE).

Progress:

In the initial phase of this research, we evaluated the imputation models for the Monthly Electricity Generation from previous project at different time points. In the second phase, detailed analysis of the MELE's IBSP (Integrated Business Statistics Program) processes was conducted, focusing on the crucial role that manual adjustments play in enhancing data quality. This is followed by developing a methodology that employs the Augmented Dickey-Fuller Test to select stable data series for simple imputation. The approach also assesses whether modeling those units with high predictive power improves overall estimation precision. The study's simple approach has demonstrated promising outcomes on the MELE survey, achieving a low Mean Absolute Percentage Error (MAPE) on the estimate, while the volume of applied manual correction has notably decreased. The next steps include applying the insights gained from this project to the new Methodological Acceleration Initiative, which aims to develop tools to monitor and reduce manual validation of data. This project has the potential to significantly reduce costs and timelines for validation, thereby improving the efficiency of economic statistics programs.

For more information, please contact:
Soufiane Fadel (soufiane.fadel@statcan.gc.ca).

3 Estimation issues in surveys

PROJECT: An estimator for concurrent use of full and reduced collection effort on random subsamples

This research is motivated by two important trends in recent years, at Statistics Canada as well as at other statistical organizations. Firstly, response rates to social surveys have continued to experience a decline. Secondly, there has been an increased use of online self-response as a collection mode due to its ease and relatively low cost. To adjust to these challenges, this project constructs an estimator (called the concurrent multi-mode estimator, CMME) for collection activities that use less expensive non-response follow-up (e.g., mail and email) for the whole sample concurrently with more expensive interviewer-facilitated non-response follow-up for a subset of non-respondents. The concurrent use of less expensive contact methods with interviewer-facilitated methods makes this different from previous work in which the non-response follow-up is carried out after a first self-response phase (e.g., Hansen and Hurwitz, 1946 and Neusy, Beaumont, Yung, Hidiroglou and Haziza, 2022) and from the experience of the National Household Survey in which the initial sampling fraction was large (Beaumont, Bocci and Hidiroglou, 2014).

Progress:

Work carried out in 2023-24 included theoretical development, a simulation study, and an experimental application of the method on the Canadian Social Survey. The theoretical work demonstrated that the CMME is unbiased with correct model specification for response propensities included in the estimator and that the CMME can leverage survey variables from respondents for some modeling, allowing for richer models than are available for many household survey non-response adjustments. Simulations further showed that, for certain variables that are similar to those found on Statistics Canada household surveys, the variance is lower and effective sample size higher under the CMME than under other estimators with comparably low bias. These results suggest that these methods could be used to give rise to substantial cost savings without unduly compromising the quality of estimates.

The results of this project were presented to the Advisory Committee on Statistical Methods in May 2024 (Mather, Boulet and Brennan, 2024) and at the Annual Meeting of the Statistical Society of Canada in June 2024.

For more information, please contact:
Anne Mather (anne.mather@statcan.gc.ca) or
Cilanne Boulet (cilanne.boulet@statcan.gc.ca).

References

Beaumont, J.-F., Bocci, C. and Hidiroglou, M. (2014). On weighting late respondents when a follow-up subsample of nonrespondents is taken. Paper presented to the Advisory Committee on Statistical Methods, 58, Statistics Canada.

Hansen, M.H., and Hurwitz, W.N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, 41, 517-529.

Mather, A., Boulet, C. and Brennan, A. (2024). An Estimator for Concurrent Use of Full and Reduced Collection Effort on Random Subsamples. Paper presented to the Advisory Committee on Statistical Methods, 78, Statistics Canada.

Neusy, E., Beaumont, J.-F., Yung, W., Hidiroglou, M. and Haziza, D. (2022). Non-response follow-up for business surveys. *Survey Methodology*, 48, 1, 95-117. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2022001/article/00006-eng.pdf.

PROJECT: Weight trimming methods for survey data

The typical weighting process for survey data involves three major stages: (1) assigning a base weight, (2) correcting the weights for unit nonresponse, and (3) modifying the weights to ensure consistency between survey estimates and known population totals, often called calibration. In some cases, the weights undergo a final modification through weight trimming to improve the efficiency of survey estimates by limiting the variability of the weights and thus ensuring greater stability of the estimates.

Trimming procedures require the specification of a tuning constant, where weights greater than the constant are trimmed to the value of the constant, and the excess weight is redistributed among remaining units. The procedures used in practice are often ad-hoc and a formal, theoretical treatment of weight trimming is currently lacking. The goals of this project are (1) for variable-specific procedures, develop some theoretical results along the lines of Ma and Wang (2020) on the choice of the optimal threshold, (2) develop some weight trimming procedures to handle multiple survey variables, (3) conduct a vast simulation study to compare the performance of several weight trimming procedures (see Potter (1988, 1990), Potter and Zheng (2015), Haziza and Beaumont (2017)) in terms of bias and efficiency and to assess the performance of mean square error estimators in terms of bias, and finally, (4) apply and demonstrate the proposed methodology on an existing dataset.

Progress:

A large simulation program was developed using a mixture of R and SAS, making use of Statistics Canada's generalized system for calibration, G-EST. The simulation incorporates more than ten trimming procedures and compares them in terms of bias and efficiency when applied to multiple survey variables with varying levels of correlation with the weights. Preliminary results show that trimming, when applied to weights with little or no correlation with the survey variables, improves efficiency without jeopardizing the bias, as expected. As correlation between the weights and the survey variables increases, the resulting estimators tend to exhibit appreciable bias which, in turn, may lead to a mean square error larger than that of the untrimmed estimator.

The immediate next steps of this project are to continue to grow the simulation study by adding additional trimming methods and incorporating categorical survey variables. Addressing the remaining goals will follow.

For more information, please contact:
Jody Krahn (jody.krahn@statcan.gc.ca) or
David Haziza (dhaziza@uottawa.ca).

References

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science*, 32, 206-226.

Ma, X., and Wang, J. (2020). Robust inference using inverse probability weighting. *Journal of the American Statistical Association*, 115, 532, 1851-1860.

Potter, F. (1988). Survey of procedures to control extreme sampling weights. *Proceedings of the Survey Research Methods Section,* American Statistical Association, 453-458.

Potter, F. (1990). A study of procedures to identify and trim extreme sampling weights. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association*, 225-230.

Potter, R., and Zheng, Y. (2015). Methods and issues in trimming extreme weights in sample surveys. *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

PROJECT: Degrees of freedom related to long-form census questionnaire estimation

Statistics Canada now publishes confidence intervals to express the quality of the estimates. The length of a confidence interval can testify to the quality of the estimate as long as the stated coverage is respected. A parameter that plays an important role in calculating confidence intervals is the number of degrees of freedom. In practice, this value is generally determined using a rule of thumb. In the case of small domains, this rule of thumb often overestimates the actual number of degrees of freedom, which leads to undercoverage of the confidence intervals.

In this project, the Satterthwaite approximation is used to derive a more precise estimate of the degrees of freedom in the context of the estimation of the long-form population census questionnaire. Variance estimation is based on an adaptation of the balanced half-sample method, as described by Devin and Verret (2016). A simulation study makes it possible to evaluate the gain in the coverage of confidence intervals for the estimation of a total of continuous and dichotomous variables. The results suggest that by using a more accurate number of degrees of freedom, the coverage is enhanced and often makes it possible to reach the nominal threshold in the problematic case of small domains.

The objective of this project for the year 2023-2024 was to finalize the writing and internal revision of an article by Toupin and Martin (2024) with the aim of submitting it to a scientific journal.

Progress:

The article by Toupin and Martin (2024) was submitted to *Survey Methodology* and is currently being revised.

For more information, please contact:
Marie-Hélène Toupin (marie-helene.toupin@statcan.gc.ca).

References

Devin, N., and Verret, F. (2016). The development of a variance estimation methodology for large-scale dissemination of quality indicators for the 2016 Canadian census long form sample. *Proceedings of the Survey Research Methods Section,* American Statistical Association, Alexandria VA.

Toupin, M.-H., and Martin, V. (2024). Améliorer la couverture des intervalles de confiance au niveau des degrés de liberté : application à l'estimation du questionnaire détaillé du recensement canadien. *Survey Methodology* (in revision).

PROJECT: Accuracy of machine learning predictions

An approach that is currently being investigated at Statistics Canada to cut costs and improve timeliness is to periodically replace the collection of survey data with predictions. These predictions would be obtained using current auxiliary data along with models determined using historical data. The estimation of finite population parameters, such as population totals, would then be achieved by aggregating these individual predictions. For instance, crop yields can be predicted based on remote sensing, agro-climatic variables and historical data, including past responses (Statistics Canada, 2020; National Academies of Sciences, Engineering, and Medicine, 2023, Chap. 8.3). Machine learning methods increasingly facilitate these predictions (Chu, 2022), which must include accuracy measures. However, traditional validation techniques like cross-validation are inadequate because the focus is on the uncertainty of the total prediction, rather than individual predictions (Hastie, Tibshirani and Friedman, 2001, Chap. 7), conditional on the covariates.

Progress:

To address the above need, a bootstrap methodology was developed and evaluated through preliminary simulations (Dasylva, Beaumont, Bosa, and Maranda, 2023), considering different models for the conditional mean response in the population and for the estimation of this conditional mean within the bootstrap procedure. Those simulations have shown that the mean square error may be estimated with a large relative bias when the bootstrap model greatly differs from the population model.

Further simulations were conducted, which revealed that the relative bias of the mean square error estimator is an increasing function of the coefficient of determination for a linear population model. Thus, it tends to be small when the coefficient of determination is moderate or small, as expected in practice, while it is large when the coefficient of determination is large. Additional simulations will be conducted in the next year to better understand the properties of our bootstrap mean square error estimator.

For more information, please contact:
Abel Dasylva (abel.dasylva@statcan.gc.ca).

References

Chu, K. (2022). Use of machine learning for crop yield prediction. Statistics Canada. Available at: https://www.statcan.gc.ca/en/data-science/network/yield-prediction.

Dasylva, A., Beaumont, J.-F., Bosa, K. and Maranda, G. (2023). Measuring the accuracy of a prediction for a finite population total. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, May 2023.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning*. New York: Springer.

National Academies of Sciences, Engineering, and Medicine (2023). *Toward a 2st Century National Data Infrastructure: Enhancing Survey Programs by Using Multiple Data Sources*. Washington, DC: The National Academies Press. Available at: https://doi.org/10.17226/26804.

Statistics Canada (2020). Integrated Crop Yield Modelling Using Remote Sensing, Agroclimatic Data and Survey Data. Available at: https://www.statcan.gc.ca/en/statistical-programs/document/5225_D1_T9_V1.

PROJECT: Bootstrap variance estimation for calibration estimators

Survey estimates are often calibrated to reflect known population totals. The general regression estimator (GREG) is a popular calibration estimator that assumes a linear relationship between the survey variable and auxiliary variables. The theory behind this estimator is well-established and a linearization variance estimator exists. Deville and Särndal (1992) showed that, for a family of calibration estimators, all members of that family are asymptotically equivalent to the GREG estimator. This suggests using the GREG variance estimator for other calibration estimators, which is the typical variance estimator implemented in practice.

Some calibration methods involve bound constraints on the weights, such as ridge calibration. Linearization in the presence of bound constraints is not always trivial, and the standard GREG variance estimator may not be appropriate. The objective of this research was to find an alternative variance estimator in that situation.

Progress:

We developed a bootstrap variance estimator that properly accounts for the bound constraints. We performed a simulation study where we compared the Horvitz-Thompson estimate, GREG and ridge calibration estimates along with their variance estimates. We showed in our simulations the importance of using pre-calibrated survey estimates (e.g., Horvitz-Thompson estimates) instead of known population totals as bootstrap control totals.

For more information, please contact:
Keven Bosa (keven.bosa@statcan.gc.ca).

Reference

Deville, J.-C., and Särndal, C.-E. (1992). Calibration estimators in survey sampling. *Journal of the American Statistical Association*, 87, 376-382.

PROJECT: Synthetic population for Labour Force Survey redesign simulations

The methodology of Canada's Labour Force Survey (LFS) undergoes a thorough review every ten years. For the current review, we developed a simulation system that reproduces the complex LFS survey process, from sampling to estimation, for investigating alternative methodologies or identifying places for improvement. The system draws from a synthetic population that is designed to be a reasonable approximation of the Canadian population in terms of demography, geography, household structure, and basic labour information, and to preserve realistic person-level labour transitions within any 6-month period (the duration a unit stays in the sample in the LFS's rotating-panel design). We accomplished this by constructing a "rotating panel population," modelled over a six-year period using a combination of cross-sectional and longitudinal modelling techniques. A population panel starts with a set of "clones" of a base population (the May 2016 long-form census respondents, passed through the cross-sectional model), and then each "clone" unit is longitudinally modelled for six months before being reset. The full synthetic population consists of six staggered panels in parallel. This structure produces the six-month series required for each simulated LFS respondent, while mitigating the drift that arises from projecting a population over an extended period.

Progress:

We finalized the longitudinal component: We had already decided on the class of model and set of potential predictors. In 2023, we produced a range of candidate models and evaluated them using cross-validation and other metrics to select our final model.

We finalized the cross-sectional component: We had already developed a plan to split the base population into classes (similar to post-stratification) and minimally adjust individuals' labour information to match LFS distributions for that month within each class. In 2023, we programmed an adapted classification tree algorithm so that splitting decisions could be constrained by conditions on an external dataset. Using this function, we established splitting conditions for the base population that ensured the monthly LFS sample for each class would be large enough to produce reasonable estimates.

We generated the data: We developed a strategy for first producing the initial month's data with the cross-sectional model and then projecting the six-month time series with the longitudinal model, maintaining consistency among employment, industry, and class of worker. We produced monthly synthetic population data files spanning 2013-2018.

We used the synthetic population to conduct simulations evaluating survey design and estimation methodologies. The results have played a major role in decision-making for most aspects of the redesign.

Greater detail can be found in Brennan and Summers (2023a, 2023b).

For more information, please contact:
Andrew Brennan (andrew.brennan@statcan.gc.ca) or
Pauline Summers (pauline.summers@statcan.gc.ca).

References

Brennan, A., and Summers, P. (2023a). Synthetic population modelling details. Internal report, Statistics Canada.

Brennan, A., and Summers, P. (2023b). Synthetic population overview. Internal report, Statistics Canada.

4 Confidentiality and Access

Confidentiality research at Statistics Canada continued to focus on developing new methods and ideas that offer alternative forms of access while continuing to ensure that personal individual and business information is not disclosed in any way. Progress was made on the projects described below. The team responsible for the Centre for Confidentiality and Access at Statistics Canada also continued to offer consultation services to internal and external partners as a way to help develop capacity in disclosure risk identification and treatment (see Section 5.5).

PROJECT: Confidentiality assessment for small area estimates

Currently, Statistics Canada has no official guidance on confidentiality rules for releasing small area estimates and no official study has yet been conducted on the subject. In recent years, there has been increasing demand from Research Data Centre (RDC) researchers for comprehensive confidentiality guidelines such that they can safely publish small area estimates. This confidentiality analysis was applied to area-level small area estimation and used simulations in R to investigate the impact of various sample sizes and various model error levels on disclosure risk.

Progress:

Simulated populations were created from which samples are selected. The simulated populations contained an auxiliary variable, a variable of interest, and domain information. The strength of the relationship between the auxiliary variable and the variable of interest was controlled through an “error” variable with a random component. Stratified random samples were drawn, and area-level small area estimates were calculated using the “sae” R package (Molina and Marhuenda, 2015). The risk of disclosure of small area estimates was compared against the Horvitz-Thompson direct estimates to demonstrate that small area estimates are inherently less risky than direct estimates, especially when sampling rates are extremely low. The results were then analyzed and finally, comprehensive confidentiality guidelines for the release of area-level small area estimates were proposed.

A paper outlining the simulation process and discussing the justifications for proposed confidentiality rules is near completion. An abstract for the 2024 International Methodology Symposium has been submitted and is awaiting approval.

For more information, please contact:
Cissy Tang (cissy.tang@statcan.gc.ca).

Reference

Molina, I., and Marhuenda, Y. (2015). sae: An R Package for small area estimation. *The R Journal*, 7, 81-98.

PROJECT: Synthetic data

Working towards more options for data users is essential. Creating synthetic data is a way to address confidentiality issues with personal data while retaining as much analytical value as possible. Synthetic data can be especially useful when looking for collaborative opportunities with external stakeholders that may not have access to the confidential microdata.

Progress:

Synthetic database for PASSAGES dynamic microsimulation model was completed. The data was released along with the model on April 23rd, 2024 (Statistics Canada, 2024). The synthetic starting population represents the Canadian population as of December 31, 2015, with family and earnings histories going back to 1966. The synthetic starting population was created primarily with machine learning methods that integrated census data, tax data and other administrative data sources. The creation of this synthetic database involved a combination of techniques to consider its hierarchical dimension (individual organized in family units) and its temporal dimension (variables relating to life trajectory with complex longitudinal correlations).

The Centre is working on updating internal guidelines for the creation of synthetic data files as well as completing a review on disclosure risk assessment for synthetic data.

For more information, please contact:
Héloïse Gauvin (heloise.gauvin@statcan.gc.ca) or
Steven Thomas (steven.thomas@statcan.gc.ca).

Reference

Statistics Canada (2024). *The Daily* – New retirement income microsimulation model now available. Available at: https://www150.statcan.gc.ca/n1/daily-quotidien/240423/dq240423c-eng.htm.

PROJECT: Optimization strategies for complementary cell suppression

Complementary Cell Suppression (CCS) is a standard method for suppressing confidentially sensitive cells when releasing tabular magnitude variables. This methodology is well developed and supported through Statistics Canada G-Confid solution, where optimal suppression solutions are obtained that ensure that suppression patterns are valid and minimize the amount of information being suppressed.

Statistics Canada is ensuring that its software solutions remain diverse and looking for solutions outside of the standard SAS-based solutions used up to this point. The challenge with moving G-Confid out of SAS is to find solutions that are compatible with the SAS solver OPTMODEL. Open-source solutions available through the PuLP Python package were studied as potential replacements. The results were analyzed, and

preliminary results were presented at the Joint Statistical Meetings (JSM) in 2023 (Chen and Thomas, 2023).

For more information, please contact:
Steven Thomas (steven.thomas@statcan.gc.ca).

Reference

Chen, H., and Thomas, S. (2023). Assessing the Performance of the Open-Source Linear Programming Solver in Cell Suppression Problems. In *Proceedings of the Survey Research Methods Section,* American Statistical Association. Available at: https://doi.org/10.5281/zenodo.10359791.

5 Support (Resource Centres)

5.1 Time Series Research and Analysis Centre

The objective of the Time Series Research and Analysis Centre is to maintain high-level expertise and offer consultation in time series throughout the agency. The centre provides consultation and advice on problems related to time series, explores problems that do not currently have known or satisfactory solutions, and develops and maintains tools to apply solutions to real-life time series problems.

The projects can be split into four sub-topics with emphasis on the following:

- Consultation and training in time series;
- Support and enhancement of the time series processing system and tools;
- Time series modelling and forecasting;
- Methodological support to consumer and producer price index programs.

Progress:

Consultation and training in time series

The Time Series Research and Analysis Centre is responsible for developing and delivering training on time series methods including seasonal adjustment, benchmarking, reconciliation, and time series modelling to participants from Statistics Canada as well as those from other agencies. In addition, the Centre provides guidance and consultation on time series projects in general for programs throughout Statistics Canada.

The Centre offered courses on time series components, seasonal adjustment, reconciliation, modelling and forecasting during the year to internal and external participants through the Statistics Canada training centre (Statistics Canada, 2024). Members of the centre also participated in outreach and training to other groups in Statistics Canada on time series topics as part of training for recent recruits (methodology branch seminar series for recruits and the data navigator course), and taught an introductory R course.

The Centre has also offered consultation to various internal programs (seasonal adjustment, time series modelling, backcasting, nowcasting, forecasting, trend estimation, calendarization, etc.). In particular, the centre provided time series support to the System of National Accounts in a number of areas, including

the monthly Gross Domestic Product, transportation, and housing investment. Representatives from the Centre also periodically attend a weekly analyst forum to maintain a presence in the analyst community. The Centre regularly consults on back-casting to preserve or restore comparability across time and has worked to produce guidelines on time series continuity for Statistics Canada's programs. This is a joint initiative with the System of National Accounts, and the recent work involved discussing the proposed guidelines with our Departmental Project Management Office to determine how the guidelines could be integrated into project management framework.

In addition, to support various internal programs, the Centre consulted and exchanged externally on time series topics (seasonal adjustment strategy during the pandemic, backcasting, deflation, software tools, etc.) with multiple federal and provincial public agencies, as well as national statistical organizations (Bank of Canada, *Institut de la statistique du Québec*, Australian Bureau of Statistics, US Census Bureau, BMO, BC Stats).

Support and enhancement of the time series processing system and tools

The Time Series Research and Analysis Centre develops and maintains a number of important tools used to process and analyse time series data for the Statistics Canada programs producing seasonally adjusted data, in particular the Generalized System G-Series, for benchmarking and raking/reconciliation/balancing (Statistics Canada, 2016), the Time Series Processing System (Ferland, 2022), and the Seasonal Adjustment Dashboard (Verret, 2021).

Work on a prototype version of G-Series in R has continued. The benchmarking and raking functionalities were implemented, packaged, and hosted in Statistics Canada's internal GitLab for internal use (package rgseriespt). The balancing functionality will be completed by fall 2024. Once the prototype is completed, it will be reviewed for official release (internally and externally).

The Time Series Processing System is a customizable SAS-based application to apply time series techniques including seasonal adjustment, benchmarking, and reconciliation, used extensively in the production of seasonally adjusted estimates for sub-annual programs within Statistics Canada (many of them being mission critical). The system is in a mature and stable state. However, it requires updating on an ongoing basis to broaden functionality and address new needs of programs in the agency. For the longer term, a new version of the system to allow flexibility to incorporate tools and new techniques available from open-source software are being investigated, in particular seasonal (R package providing an interface to the X-13ARIMA-SEATS software from the US Census Bureau), PyX13 (US Census Bureau, pre-release), JDemetra+/RJDemetra (Eurostat), and rgseriespt (R package for G-Series). Discussions related to time series processing systems were held with the *Institut national de la statistique et des études économiques* (France), the US Census Bureau, and the Australian Bureau of Statistics.

A number of improvements to the Seasonal Adjustment Dashboard were implemented this year. In particular, new functionalities to improve the efficiency were added, compatibility with different versions of R was added and minor bugs were resolved. The dashboard is in the process of being deployed for an additional labour program. A complete documentation has been produced.

Time Series modelling and forecasting

Increasing timeliness of statistical indicators is an important priority for Statistics Canada and one option for doing so is through time series modelling to nowcast economic indicators much earlier than the point

in time where the first traditional estimator is produced. Work on a nowcasting project to develop a more accurate method of estimating renovation activity expenditures (part of the monthly Investment in Building Construction Program, a key economic indicator that quantifies the state of building construction investment in the economy was presented at the Scientific Review Committee of the Modern Statistical Methods and Data Science Branch (Patak and Plunkett, 2023). Work on producing advanced indicators of energy statistics and a general framework for advance quality indicators was presented at 76th meeting of the Statistics Canada Advisory Committee on Statistical Methods (Le Moullec and Matthews, 2023; Matthews, 2022). An invited session was organized on the topic of real-time estimation at the International Statistical Institute's World Statistics Congress in July of 2023.

The Centre continued its investigation into ways to produce early indications of structural breaks using state-space models.

Methodological support to consumer and producer price index programs

The Time Series Research and Analysis Centre also has a unit dedicated to providing methodological support to consumer and producer price index (CPI and PPI) programs.

In recent years, the Consumer Price Index has undergone many changes to collection and methodology to both modernize practices and adapt to changing conditions since the COVID-19 pandemic. Traditional in-person collection has been replaced by a mix of online collection and alternate data sources including point-of-sale scanner data, administrative databases, web-scraping, and other surveys. Basket weights now come from the Household Financial Consumption Expenditure series instead of the Survey of Household Spending, to allow more frequent and timely basket updates. Record linkages have created a partial CPI frame linking businesses to commodity-level sales for some CPI aggregates. The Centre reviewed the CPI sampling methodology under the context of these changes. Several potential improvements were suggested including an updated sample allocation, sample rotation, probability sampling for some aggregates, and the use of quality indicators customized to unique characteristics of the CPI (Francis, 2023). Both design-based and model-based methods for variance estimation were explored. The centre also developed a strategy and an R tool that can be used to evaluate the sample of different components of the CPI. The method was applied to the clothing and footwear component this year and improvements were proposed.

Many of the Services Producer Price Indices (SPPI) use probability sampling of businesses at the first stage. Most use Sequential Poisson with probability-proportional-to-size sampling (Ohlsson, 1998), stratified by industry groups. Past designs allocated sample to industry groups using simpler methods such as proportional-to-revenue or proportional-to-size, in the absence of good variance estimates. Each SPPI sample was designed independently with budgets, sample sizes, and precision varying greatly. This project explored optimizing use of resources to ensure consistent quality across SPPIs and conserve budget where possible. The Beaumont-Patak generalized bootstrap (Beaumont and Patak, 2012) and Taylor linearization were used to estimate first-stage variances for Lowe price indices with Sequential Poisson PPS samples. Adjustments were made to account for characteristics unique to price indices: periodic nonresponse, parental imputation, price-updated weights, chaining, and chain drift (Francis, 2024). R functions were developed. Using these variance estimates, sample sizes were estimated to reach coefficient of variation targets for the lowest levels published. Results were tested for both a large survey (Wholesale Services

Price Index) and a small survey (Architectural, Engineering and Related Services Price Index) and presented to the Producer Price Division.

For more information, please contact:
Etienne Rassart (etienne.rassart@statcan.gc.ca).

References

Beaumont, J.-F., and Patak, Z. (2012). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. *International Statistical Review*, 80, 127-148.

Ferland, M. (2022). *Time Series Processing System – v3.08*. Internal document, Statistics Canada.

Francis, J. (2023). *Canadian CPI Sample Design Methodology Review*. Internal document, Statistics Canada.

Francis, J. (2024). *Design-Based Estimates and Variance Estimation for SPPIs with Sequential Poisson PPS Samples*. Internal document, Statistics Canada.

Le Moullec, J., and Matthews, S. (2023). On the Path to Real-Time Economic Indicators: A use case in producing model-based flash estimates for monthly electricity generation: Simpler is better! Presented at the 76th meeting of the Advisory Committee on Statistical Methods, Statistics Canada.

Matthews, S. (2022). *A framework for Advance Indicators at Statistics Canada*. Internal document, Statistics Canada.

Ohlsson, E. (1998). Sequential Poisson sampling. *Journal of Official Statistics*, 14, 149-162.

Patak, Z., and Plunkett, K. (2023). Nowcasting monthly renovation activity expenditures. Presented at the April 28 meeting of the Scientific Review Committee of the Modern Statistical Methods and Data Science Branch. Internal document, Statistics Canada.

Statistics Canada (2016). *G-Series 2.00.001 User Guides*. Internal document, Statistics Canada.

Statistics Canada (2024). Workshops, training and references. Available at: https://www.statcan.gc.ca/eng/wtc/training.

Verret, F. (2021). Statistics Canada's Seasonal Adjustment Dashboard. Proceedings: *Symposium 2021, Adopting Data Science in Official Statistics to Meet Society's Emerging Needs*, Statistique Canada, Ottawa, Canada.

5.2 Economic Generalized Systems

The Economic Generalized Systems team is responsible for the support and development of three generalized systems, namely G-Sam – the generalized sampling system, Banff – the generalized system for edit and imputation, and G-Est – the generalized system for estimation.

Progress:

A typical volume of support cases for G-Sam, Banff and G-Est was processed by the project team. Most of these were resolved with suggestions on how to apply the systems in practical terms, however several required more involvement. In particular, the Banff team conducted an extensive review of the edit and imputation process flow for the Job Vacancy and Wage Survey. The Banff team has proposed updates to the process flow design and will continue to work with the survey team to implement and test those changes. The G-Sam team also worked with methodologists working on the Canadian Survey on Interprovincial Trade and the Canadian Survey on Business Conditions to optimize an allocation with respect to standard error precision targets on estimates of proportions. This support case included extensive consultations with the clients, a review of relevant theory, and the construction of several example programs.

The generalized systems team was heavily involved in various activities related to Statistics Canada's analytical diversification initiative, an initiative to transition systems to open-source alternatives. A member of the team was active in the initiative's task force and presented the topic for discussion at the 77th meeting of the Advisory Committee on Statistical Methods (Gray, 2023). Throughout the year, the unit updated the Generalized Systems Steering Committee on the diversification plans and progress for G-Sam, Banff and G-Est. Team members also met with representatives of foreign National Statistical Organizations on occasion to discuss analytical diversification, including the *Institut national de la statistique et des études économiques* (France) to discuss their own migration from SAS to R, and the Australian Bureau of Statistics to shed some light on Statistics Canada's experience to date.

Version 1.04 of G-Sam was released on November 9, 2023, and included several significant changes to the allocation module. This included the introduction of user-specified probabilistic constraints on the sample size and number of respondents for arbitrary domains, improved variance approximations appropriate for optimization solvers, and a new diagnostic output file. The underlying mathematics are described in the G-Sam user guides (Stinner, 2024). The G-Sam team intends to present the approach at a conference (or similar venue) when the R version of G-Sam is released next year. G-Sam version 1.04 is intended as the final SAS release.

Significant progress was made on the Banff modernization project, and a Python version is on track to be released in December 2024. A number of structural changes proposed by (Gray, 2022) are to be implemented in the release version, including new process controls, process blocks, and custom Python plugins. These improvements bring the system in line with the Generic Statistical Data Editing Model (UNECE, 2019), and the Banff team has been invited to give a keynote presentation discussing the new system at the next UNECE Expert Meeting on Statistical Data Editing (October 2024).

Prototype open-source versions of G-Sam and G-Est are currently in development. For each system, the team assessed options to adopt available open-source packages performing specific functions (e.g., stratification, calibration) but found these packages lacked production-ready performance standards. Diversification plans for both systems have been presented to the Generalized Systems Steering Committee.

For more information, please contact:
Etienne Rassart (etienne.rassart@statcan.gc.ca).

References

Gray, D. (2022). Banff's Next Step: An Open-Source Data Editing System for Advanced Tools and Collaboration. UNECE Expert Meeting on Statistical Data Editing.

Gray, D. (2023). Statistics Canada's Analytical Diversification Initiative – impact on Statistical Generalized Systems. Presented at the 77th meeting of the Advisory Committee on Statistical Methods, Statistics Canada.

Stinner, M. (2024). *G-Sam user guide (draft)*. Internal document, Statistics Canada.

UNECE (2019). Generic Statistical Data Editing Model. Available at: https://statswiki.unece.org/display/sde/GSDEM.

5.3 Record Linkage Resource Centre

The objectives of the Record Linkage Resource Center (RLRC) are to provide consultation services to internal and external users of record linkage methods, which includes making recommendations about the software and methods to be used, and collaborative work on record linkage applications. We also facilitate the dissemination of information on record linkage methods, software, and policy as well as the analysis of linked data to interested parties inside and outside Statistics Canada.

Progress:

We continued to support the development team of G-Link, the record linkage system developed at Statistics Canada, and to participate in the Record Linkage Working Group meetings of the Information Technology Lifecycle Management (ITSLM) and the Statistics Integration Methods Division (SIMD). The RLRC team met with ITSLM representatives every two weeks and followed up on minutes mentioning possible sources, past or present, of corrections, bugs or improvements for G-Link. The RLRC also offered support to internal and external G-Link users who requested assistance, provided comments or submitted suggestions through requests to G-Link_info.

During the year, most of the methodological work focused on the maintenance, development and support for users of version 3.5 of G-Link on SAS servers in cloud computing. A typical volume of support cases for G-Link was processed by the project team. Most of these were resolved with suggestions on how to apply the system in practical terms, however several required more involvement.

The development consisted of standardizing and integrating manual procedures for the estimation of record linkage errors and develop indicators of quality such as specificity and sensitivity, as well as ensuring the user interface made the outputs easy to interpret.

The RLRC has also worked on a variety of other probabilistic linkage in the Social Data Linkage Environment (SDLE). These linkages helped us to analyze the performance of the software and the solutions to be provided. Work on these projects has resulted in more systematic approaches to defining and adjusting record linkages on cloud-based SAS servers. Work was also undertaken in reweighting to compensate for bias introduced by missed links, including experimental methods for creating replicate weights in the case

of a linkage between two independently-drawn samples. Unit members also did further theoretical and prototype work on quality indicators for models fitted to linked data.

Members of the team offered formal courses with Statistics Canada's Training Centre, as well as seminars for recently recruited statisticians, a record linkage forum and other ad hoc presentations to analysts.

For more information, please contact:
Abdelnasser Saïdi (abdelnasser.saidi@statcan.gc.ca).

5.4 Data Analysis Resource Centre

The main goal of the Data Analysis Resource Centre (DARC) is to provide advice on the appropriate use of data analysis tools and methods, and to promote best practices in this area. DARC's services – which focus mainly on survey, census, or administrative data – are available to the employees of the Agency and other departments, as well as to analysts and researchers from academia and Research Data Centers (RDCs).

Progress:

Consultations

Consultation services were provided as requested by internal and external clients. Between April 1, 2023 and March 31, 2024, DARC responded to 37 requests. The questions varied in complexity and included topics such as analysis of Likert scales, interpretation of regression outputs, comparison of medians, logistic regression, quartiles and ratios estimations with survey data, specification of degrees of freedom and comparison of several cycles of a survey. DARC also helped clients with the implementation of statistical methods in SUDAAN, SAS, STATA, and R software.

Provision of Training

DARC redesigned and presented, in French, the internal course 0438A "Statistical Analysis of Survey Data – Module 1". R code was developed for exercises and examples, in addition to SUDAAN and SAS code. This six-day course is a mix of theory and practice.

DARC presented at Statistics Canada's Data Interpretation Workshop on data analysis with complex survey data, in English and French. DARC again presented the sessions on linear regression and on logistic regression, with complex survey data, of the Statistical Modelling Course at Statistics Canada, in French. DARC also gave the seminar for recruits on analysis of data from a complex survey.

Collaboration

DARC collaborated in developing measurement strategies for the Workplace Mental Health Performance Measurement Project with the Treasury Board Secretariat (TBS). This project used data from the 2022 cycle of the Public Service Employee Survey (PSES) to measure latent variables like psychological risk factors, behaviors, etc. and to calculate factor scores for different levels of aggregation. The factor scores developed for this project were used to create the Federal Public Service Workplace Mental Health Dashboard: Mental Health Dashboard - Canada.ca (tbs-sct.gc.ca). The measurement models were

developed using factor analysis and structural equation modelling as discussed by Blais, Mach, Michaud and Simard (2020) and Blais, Michaud, Simard, Mach and Houle (2021).

For further information, please contact:
Fritz Pierre (fritz.pierre@statcan.gc.ca) or
Isabelle Michaud (isabelle.michaud@statcan.gc.ca).

References

Blais, A.-R., Mach, L., Michaud, I. and Simard, J.-F. (2020). Analysis of the Public Service Employee Survey Items as Measures of the Psychosocial Risk Factors. Presentation to the Workplace Mental Health Performance Measurement Steering Committee, October 7, 2020.

Blais, A.-R., Michaud, I., Simard, J.-F., Mach, L. and Houle, S. (2021). Measuring workplace psychosocial factors in the federal government. *Health Reports*, 32, 12. Paper available at https://www150.statcan.gc.ca/n1/en/pub/82-003-x/2021012/article/00001-eng.pdf.

5.5 Centre for Confidentiality and Access

The methodology group responsible for confidentiality and access methods continued to offer consultation and support services to internal and external partners on the various access solutions and disclosure avoidance strategies.

De-identification

The confidentiality support group continued to offer its expertise in the understanding and development of ideas related to de-identification and anonymization. Statistics Canada is continuing to enhance its own internal strategies to ensure that internal information is de-identified whenever possible to minimize risks of disclosure.

External consultation

Statistics Canada has offered its expertise to several groups both domestically and internationally. Internationally, Statistics Canada has shared its Random Tabular Adjustment tool with Statistics Sweden as well as National Agricultural Statistics Service in the United States. The team has also been consulting with outside groups including IBM and Gurobi to find appropriate alternatives to using SAS OPTMODEL in its complementary cell suppression strategies. The confidentiality group has also been working closely with their health colleagues in developing new ways of sharing cancer data with international counterparts.

Domestically, Statistics Canada has had several meetings with the Bank of Canada to advise on its disclosure control strategies, the Canadian Revenue Agency to discuss re-identification risks, and the Public Health Agency of Canada to discuss synthetic data.

For more information, please contact:
Steven Thomas (steven.thomas@statcan.gc.ca).

5.6 Support and Research Activities at the Data Science Innovation Division

The Methodology Research and Development Program (MRDP) at Statistics Canada has supported many activities for the Data Science and Innovation division. The support from the MRDP has enabled many services, community of practice, further research, centre of expertise and guidelines that can profit the agency.

Activities, Mandates and Products:

The funding enabled the Machine Learning Community of Practice (COP) to fulfill its mandate to increase knowledge and build capacity across Statistics Canada with respect to machine learning. The activities related to the COP covers a wide range, such as hosting weekly Machine Learning Technical Series seminars, weekly newsletters and podcasts, coding challenges and weekly drop-in "office hours". Another centre that benefited from the funding is the Natural Language Processing (NLP) Centre of Expertise (COE) that is mandated to centralize resource for knowledge sharing and capacity building in text analytics using machine learning and create, maintain, and promote best practices and guidelines in text analytics. Their activities comprise providing reviews, consultation, and guidance to NLP practitioners within Statistics Canada, as well as creating a list of realized and on-going NLP projects across the agency.

Aside from the communities and centres, the Machine Learning guidelines were elaborated, to develop an ever-green Machine Learning Guidelines document for practitioners. These guidelines comprise responsible artificial Intelligence (AI) and machine learning, explainable AI, and fairness in machine learning.

Lastly, two prominent research projects were supported for further experimentation, as their results were promising for the agency. The first one tackles synthetic data generation for health by evaluating the safety and effectiveness of synthetic healthcare data generation methods, focusing on a balance between utility and privacy. This project is aligned with the Canadian Institute for Advanced Research (CIFAR) initiatives and aims to modernize data sharing and decision-making within the healthcare sector. A workshop on the project will be offered in the near future.

The second promising research project funded by the committee is an investigative study aimed at enhancing the efficiency of microdata preprocessing in survey data at Statistics Canada. The project's focus is on the robustness of data preprocessing for complex tasks and the performance of an imputation method under varying conditions. This research is designed to optimize the timeliness and reduce manual efforts in the Monthly Electricity Supply and Disposition Survey (see also the project "Time and effort optimization in data processing" in Section 2).

For more information, please contact:
Marie-Eve Bedard (marie-eve.bedard@statcan.gc.ca).

5.7 Questionnaire Design Resource Centre

The Questionnaire Design Resource Centre (QDRC) is a focal point of expertise at Statistics Canada for questionnaire design and evaluation. The QDRC provides consultation and support services, and carries out projects and research related to the development, testing and evaluation of survey questionnaires. The QDRC plays a very important role in quality management and responds to program requirements

throughout Statistics Canada by consulting with clients, respondents and data users and by pre-testing survey questionnaires.

While much of the QDRC's work is carried out on a cost-recovery basis, the section is frequently approached on an ad hoc basis for expert reviews and consultation services on a wide variety of surveys. The group also offers courses on questionnaire design.

Progress:

The QDRC conducted many reviews of survey questionnaires. While most of these involved Statistics Canada questionnaires, several were conducted for surveys being done by other government organizations such as Transport Canada, Public Services and Procurement Canada and others.

The group also contributed to various corporate consultation initiatives.

For further information, please contact:
Paul Kelly (paul.kelly@statcan.gc.ca).

5.8 Quality Assurance Resource Centre

The Quality Assurance Resource Centre (QARC) is dedicated to advancing research and development in statistical methods aimed at enhancing the quality assurance and control processes. Our primary objective is to elevate the standards of survey data collection and processing operations within the bureau. Achieving this goal involves exploring various methodologies, with a particular emphasis on refining the outgoing quality of data.

At the heart of our efforts lies the provision of methodological services for G-Code, a generalized system developed at Statistics Canada for creating coded databases and implementing machine learning algorithms in data processing. Our research encompasses a broad spectrum of quality assurance and control practices, addressing issues of efficiency and automation. These findings are not only pertinent to our specific operations but also have wide-reaching applicability across various stages of survey operations.

Progress:

The methodological support team assisted the G-Code development team and monitored user inputs to identify potential improvements for G-Code. Additionally, QARC extended support to both internal and external G-Code users whenever help, comments, or suggestions concerning G-Code were required.

Throughout the year, QARC has focused on developing a novel methodology termed "Quality Control by Score" to enhance the Quality Control (QC) of Machine Learning (ML) text coding processes. With ML technology playing an increasingly pivotal role, ensuring the quality of generated codes becomes paramount. In response to this demand, Statistics Canada has actively pursued the development of a strategy for determining optimal QC sampling rates, utilizing scores derived from the ML process. This methodology will enable a responsible approach to classifying data with a broader implementation of machine learning. Our objective is to utilize this approach for QC purposes across various classifications within essential surveys such as the Labour Force Survey (LFS), Job Vacancy and Wage Survey (JVWS),

Canadian Community Health Survey (CCHS), and the Statistical Business Register (SBR). Notably, a paper detailing this methodology was presented at the Advisory Committee on Statistical Methods (Oyarzun, Wile and Evans, 2023).

The QARC team has undertaken an investigation into a calibration methodology aimed at establishing a more coherent relationship between machine learning scores and the accuracy of the coded data. Calibration in this context involves fine-tuning the ML model to align its scoring system with the actual accuracy of the coded data. By refining this relationship, we strive to enhance the reliability and precision of the classification process, ultimately improving the overall quality assurance measures within our data processing operations. This calibration methodology represents a crucial step towards optimizing the performance of machine learning algorithms in coding tasks, ensuring that the output aligns closely with the intended classifications, and bolstering confidence in the final data.

For more information, please contact:
Javier Oyarzun (javier.oyarzun@statcan.gc.ca).

Reference

Oyarzun, J., Wile, L. and Evans, J. (2023). Quality Control by Score. Paper presented at the Advisory Committee on Statistical Methods, October 2023, Statistics Canada.

5.9 Data Ethics Secretariat

The role of the Data Ethics Secretariat is to implement the Necessity and Proportionality Framework. Concretely, the Data Ethics Secretariat conducts ethical reviews on new data acquisitions via survey or other sources, and new data uses such as microdata linkages. The purpose of these ethical reviews is to ensure responsible use of data throughout the data lifecycle. The Data Ethics Secretariat raises ethical considerations, holds discussions with program managers and makes recommendations to the Principal Data Ethics and Scientific Integrity Officer. The Data Ethics Secretariat also supports the internal Data Ethics Committee and has a capacity building role.

Progress:

On top of conducting roughly 180 ethical reviews, members of the Data Ethics Secretariat have given numerous presentations to inform internal partners, colleagues from other federal departments as well as from international organizations on Statistics Canada's approach on data ethics. The team gathers information to remain up to date on topics perceived as sensitive by the public. This is done by conducting literature reviews on some targeted topics, informal discussions with internal partners such as Communications and the Questionnaire Design Resource Centre or counterparts from other federal departments or foreign National Statistical Offices.

In addition to its internal activities, the team is also very active internationally, playing a leadership role in the UNECE Task Team on Ethical Leadership. The main objective of this task team is to write a reference book on ethics for National Statistical Organizations. Work on this reference book took a big step forward in March 2024 at the Workshop on Ethics in Modern Statistical Organisations, where discussion sessions

on each section of the reference book were organized. The reference book is expected to be completed in 2025.

For more information, please contact:
Martin Beaulieu (martin-j.beaulieu@statcan.gc.ca).

5.10 Quality Secretariat

The Quality Secretariat's mandate includes designing and managing quality management studies and responding to requests for quality management information or assistance from Statistics Canada's various programs or other organizations.

PROJECT: Capacity building with internal, national and international partners

The Quality Secretariat's objective is to provide advice and undertake capacity-building measures internally, with national partners (other departments or other organizations) and international partners, primarily by giving a general overview of Statistics Canada's quality management practices and official quality-related documents (the Quality Assurance Framework and the Quality Guidelines) and by providing quality management support services.

Progress:

The Quality Secretariat undertook capacity building for many partners during the reporting period. Internally, training was offered through various courses for staff. At the national partner level, formal presentations on quality management practices were made to two organizations, in addition to holding workshops and seminars. The Quality Secretariat collaborated with Statistics Canada's Data Literacy Training Initiative in the development of on an online training module entitled Data Quality as Fitness for Use. This course introduces a clear and easy-to-use framework to help the learner to define their data need, assess the fitness of potential data sources for a specific purpose and make a decision about whether a potential data source is in fact, fit for a specific purpose.

Discussions occurred within the Government of Canada Enterprise Data Community of Practice Data Quality Working Group. This working group, co-chaired by Statistics Canada, aims to define a data quality framework applicable to all Government of Canada organizations as part of the implementation of the Data Strategy. A draft Data Quality Framework is available for partners from other federal departments and a shorter version of the framework, called Guidance on Data Quality, was approved and made publicly available in January 2024.

At the international level, involvement with the United Nations Expert Group on National Quality Assurance Frameworks increased, taking on the role of co-chair of the Subgroup on administrative and other data sources. The purpose of the Subgroup is to prepare a Module for Quality Assurance when using Administrative and Other Data Sources to produce Official Statistics. This module aims to provide practical and concise guidance and best practices for statistical agencies in assuring the quality of official statistics when administrative data sources, "other" data sources or multiple data sources are used for the production of official statistics, to be used as a complement to the United Nations National Quality Assurance Framework Manual for Official Statistics (United Nations, 2019). The module will be the subject

of worldwide consultation in the Spring of 2024 and submitted for approval to the United Nations Statistical Commission towards the end of 2024.

For more information, please contact:
Martin Beaulieu (martin-j.beaulieu@statcan.gc.ca).

Reference

United Nations (2019). United Nations National Quality Assurance Frameworks Manual for Official Statistics. Available at: https://unstats.un.org/unsd/methodology/dataquality/un-nqaf-manual/.

PROJECT: Quality indicators for statistics from integrated data

In order to provide users with quality indicators for programs that combine administrative data sources, the Quality Secretariat has worked on the development of a composite indicator that combines quality indicators related to different stages of data processing (record linkage, imputation, geocoding, etc.) into a single indicator. The objective is to give a global view of the quality of an estimate by considering several factors that can introduce errors (Gagnon, Qian, Yeung, Lebrasseur and Beaulieu, 2022).

Progress:

These indicators were used for additional tables of the Canadian Housing Statistics Program (CHSP). The Quality Secretariat keeps providing support on the code and the method. Some options were explored in collaboration with the CHSP team to improve the method and the consistency of the results.

For more information, please contact:
Martin Beaulieu (martin-j.beaulieu@statcan.gc.ca).

Reference

Gagnon, R., Qian, W., Yeung, A., Lebrasseur, D. and Beaulieu, M. (2022). Development of a composite quality indicator for statistical products derived from administrative sources. Statistics Canada, Available at: https://www150.statcan.gc.ca/n1/pub/46-28-0001/2022001/article/00001-eng.htm.

PROJECT: Quality Assurance Framework Update

The Quality Secretariat initiated a review of Statistics Canada's Quality Assurance Framework (QAF). The current version was released in 2017. While the content of the current version is still valid, the fast evolution of new data sources and new techniques used in the production of official statistics made this review relevant. The updated version will highlight the importance of data stewardship, data ethics principles and some considerations relative to new techniques used. The plan for the update was presented to the Advisory Committee on Statistical Methods in the Fall 2022 (Beaulieu, Yung and Rancourt, 2022).

For more information, please contact:
Martin Beaulieu (martin-j.beaulieu@statcan.gc.ca).

Reference

Beaulieu, M., Yung, W. and Rancourt, E. (2022). Data Quality and Official Statistics in a Modern World. Paper presented at the Advisory Committee on Statistical Methods, October 2022, Statistics Canada.

6 Other activities

6.1 Survey Methodology Journal

Survey Methodology is a free online peer-reviewed statistical journal published twice a year by Statistics Canada since 1975. The journal aims to publish innovative theoretical or applied research papers, and sometimes review papers, that provide new insights on statistical methods relevant to National Statistical Offices and other statistical organizations. Papers are published free of charge in both official languages and released at: www.statcan.gc.ca/surveymethodology. Its editorial board includes world-renowned leaders in survey methods from the government, academic and private sectors.

Progress:

The June and December 2023 issues (49-1 and 49-2) were released. The June 2023 issue contains eleven papers, which includes a special paper in memory of Professor Chris Skinner, the winner of the 2019 Waksberg Award, written by Natalie Shlomo, along with a tribute by Danny Pfeffermann and comments by J.N.K. Rao as well as Jae Kwang Kim and HaiYing Wang. Eighteen papers were published in the December 2023 issue, which featured the 2023 Waksberg paper by Ray Chambers entitled "The missing information principle – A paradigm for analysis of messy sample survey data", as well as a special paper by Pascal Ardilly, David Haziza, Pierre Lavallée and Yves Tillé on Jean-Claude Deville's contributions to survey theory and official statistics. The latter includes five discussions by Guillaume Chauvet, Marc Christine, Françoise Dupont, Camelia Goga and Anne Ruiz-Gazen, and Carl-Erik Särndal as well as a rejoinder by the authors. The December issue also features four invited papers that were presented at the 2021 Colloque francophone sur les sondages.

In 2023, 44 papers were submitted to the journal. The average number of days from submission to initial decision was 55. All submitted papers were reviewed within 130 days, except for one paper that was exceptionally reviewed in 188 days, and 80% of them were reviewed within 90 days. Among those 44 papers, 24 were rejected, 11 were accepted and 9 had not received a final decision (including papers that were not revised by the authors before the deadline) as of July 18, 2024. From April 2023 to March 2024, the *Survey Methodology* pages were viewed 52,046 times.

The June 2024 issue is devoted to three papers that were presented at the 2022 Morris Hansen Lecture event on the use of non-probability samples by Courtney Kennedy, Yan Li and Jean-François Beaumont. All three papers are discussed by international experts in the field, and discussions are followed by rejoinders. An introduction by Partha Lahiri, the Guest Editor for this special issue, precedes the papers. The June 2025 issue will be dedicated to the celebration of the 50th anniversary of *Survey Methodology*. It will include a special discussion paper by Carl-Erik Särndal, entitled "Progress in survey science: yesterday – today – tomorrow", along with discussions from eminent survey statisticians. In addition to

Särndal's paper and its discussion, it will contain several other invited papers by renowned experts in survey statistics and methodology.

For more information, please contact:
Jean-François Beaumont (jean-francois.beaumont@statcan.gc.ca).

6.2 Knowledge Transfer – Statistical Training

The Statistical Talent Development Working Group, whose primary mandate remains statistical training within the Field and the organization, had another busy and productive year. Several courses were offered this year including those related to time series, questionnaire design, sampling, record linkage, imputation, weighting, small area estimation, bootstrapping, modeling, introduction to supervised Machine Learning (ML), fairness and explainability in ML and programming in R.

In terms of new activities, the group continued to design and prioritize learning activities that can be developed in a timely manner and focus on active learning. This year, a new course on statistical analysis with survey data has been developed. This course explains in particular how to carry out statistical analysis under a complex sampling design. The course was offered in both official languages and other sessions are planned for next year. We have also developed an introductory course to programming with Python. This course was also offered in both official languages.

Concerning the next year, we will continue to offer the courses of the curriculum according to demand and the availability of teachers. In addition, a course on variance estimation will be developed and offered during the year 2024-25. A workshop on the design of simulation studies will also be developed and offered in 2024-25.

The Talent Development Working Group offers various types of training opportunities so that employees can enjoy flexibility in their professional development. In addition to the activities mentioned above, there are many opportunities for self-study and self-learning, including the DataCamp platform, as well as communities of practice.

For more information, please contact:
Keven Bosa (keven.bosa@statcan.gc.ca).

6.3 Statistics Canada's International Methodology Symposium

Statistics Canada's 2024 International Methodology Symposium "Shaping the future of official statistics" will take place October 30, 31 and November 1, 2024. The Symposium will offer plenary sessions and parallel sessions that cover a variety of topics. Consistent with other statistical conferences worldwide, this year's Symposium will feature speakers delivering their presentations in person. Observers are offered the choice to attend in person or to join the sessions virtually.

Progress:

The memberships of the organizing, program and logistics committees have been formed, and the title and format of this year's Symposium have been confirmed. The logistics committee has begun

coordinating with Conference Services to ensure the provision of conference rooms, technical support delivery and simultaneous interpretation.

The program committee has identified topics and organizers for invited sessions and has started to identify topics and organizers for contributed sessions. Our plenary sessions will include an address by the 2024 winner of the Waksberg award, Richard Valliant, and another by our keynote speaker.

All relevant information will be made available on our website at:
https://www.statcan.gc.ca/eng/conferences/symposium2024/index.

For more information, please contact:
Peter Wright (peter.wright@statcan.gc.ca).

7 Research papers sponsored by the Methodology Research and Development Program

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2024). Authors' response to comments on "Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data": Some new developments on likelihood approaches to estimation of participation probabilities for non-probability samples. *Survey Methodology*, 50, 1, 123-141. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024001/article/00001-eng.pdf.

Brennan, A., and Summers, P. (2023a). Synthetic population modelling details. Internal report, Statistics Canada.

Brennan, A., and Summers, P. (2023b). Synthetic population overview. Internal report, Statistics Canada.

Chen, H., and Thomas, S. (2023). Assessing the Performance of the Open-Source Linear Programming Solver in Cell Suppression Problems. In *Proceedings of the Survey Research Methods Section,* American Statistical Association. Available at: https://doi.org/10.5281/zenodo.10359791.

Dasylva, A. (2024). Estimation of small area means with linked data. Internal report, Statistics Canada.

Dasylva, A., Beaumont, J.-F., Bosa, K. and Maranda, G. (2023). Measuring the accuracy of a prediction for a finite population total. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, May 2023.

Dasylva, A., and Goussanou, A. (2024a). Measuring the linkage accuracy when many files are linked to a spine. Internal report, Statistics Canada.

Dasylva, A., and Goussanou, A. (2024b). Making statistical inferences about linkage errors. *Japanese Journal of Statistics and Data Science*. Available at: https://doi.org/10.1007/s42081-023-00228-9.

Dasylva, A., Goussanou, A. and Nambeu, C.-O. (2024). Models of linkage error for capture-recapture estimation without clerical-reviews. *Survey Methodology*, 50, 2. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2024002/article/00007-eng.pdf.

Francis, J. (2023). *Canadian CPI Sample Design Methodology Review*. Internal document, Statistics Canada.

Francis, J. (2024). *Design-Based Estimates and Variance Estimation for SPPIs with Sequential Poisson PPS Samples*. Internal document, Statistics Canada.

Gray, D. (2023). Statistics Canada's Analytical Diversification Initiative – impact on Statistical Generalized Systems. Presented at the 77th meeting of the Advisory Committee on Statistical Methods, Statistics Canada.

Le Moullec, J., and Matthews, S. (2023). On the Path to Real-Time Economic Indicators: A use case in producing model-based flash estimates for monthly electricity generation: Simpler is better! Presented at the 76th meeting of the Advisory Committee on Statistical Methods, Statistics Canada.

Mather, A., Boulet, C. and Brennan, A. (2024). An Estimator for Concurrent Use of Full and Reduced Collection Effort on Random Subsamples. Paper presented to the Advisory Committee on Statistical Methods, 78, Statistics Canada.

Millar, G. (2024). Logistic regression on linked data from a secondary analyst perspective. Presentation at CANSSI-CRT Workshop on Modern Methods in Survey Sampling, University of Ottawa, July 8-10.

Miller, J. (2024). *Disclosure Risk of Parametric Regression Output*. Internal report, Statistics Canada.

Oyarzun, J., Wile, L. and Evans, J. (2023). Quality Control by Score. Paper presented at the Advisory Committee on Statistical Methods, October 2023, Statistics Canada.

Patak, Z., and Plunkett, K. (2023). Nowcasting monthly renovation activity expenditures. Presented at the April 28 meeting of the Scientific Review Committee of the Modern Statistical Methods and Data Science Branch. Internal document, Statistics Canada.

Santos, B. (2023). Multi-Party Privacy Preserving Record Linkage based on Circuit Private Set Intersection. Canadian Mathematical Society Winter Meeting, Montréal, Canada.

Statistics Canada (2024). *The Daily* – New retirement income microsimulation model now available. Available at: https://www150.statcan.gc.ca/n1/daily-quotidien/240423/dq240423c-eng.htm.

Stinner, M. (2024). *G-Sam user guide (draft)*. Internal document, Statistics Canada.

Toukal, J. (2023). Logistic regression in the context of record linkage. Internship report for l'École Nationale de la Statistique et de l'analyse de l'information.

Toupin, M.-H., and Martin, V. (2024). Améliorer la couverture des intervalles de confiance au niveau des degrés de liberté : application à l'estimation du questionnaire détaillé du recensement canadien. *Survey Methodology* (in revision).

You, Y. (2023). An empirical study of hierarchical Bayes small area estimators using different priors for model variances. *Statistics in Transition New Series*, 24, 169-178.

You, Y., and Bosa, K. (2024). Performance of hierarchical Bayes small area estimators using non-informative and informative priors with LFS application. Submitted to *Survey Methodology* (under revision).

You, Y., and Hidiroglou, M. (2024). Empirical study of confidence intervals for small area proportion estimation with LFS application. Internal report, Statistics Canada.