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

Statistical Methodology Research and Development Program Achievements, 2024-2025

Release date: October 10, 2025

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

Email at infostats@statcan.gc.ca

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

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

Standards of service to the public

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

Note of appreciation

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

Statistical Methodology Research and Development Program

Achievements, 2024-2025

This report summarizes the 2024-2025 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, 2024-2025

Table of Contents

1 Data integration

1.1 Integration of probability and non-probability samples

PROJECT: Prediction inference for finite population totals or means with no or limited probability survey data

For a number of years, National Statistical Offices in several countries have been seeking to reduce data collection costs and the burden on respondents, while increasing the use of alternative data sources and modern prediction techniques. At Statistics Canada, an idea that is currently being explored to address these challenges is to reduce the frequency of probability surveys and replace missing survey data with predictions. For example, a probability survey could be conducted every other year, instead of every year, and missing survey data in non-survey years could be replaced with predictions, provided that auxiliary data are available in the non-survey years along with training data. The objective of this project is to study the following two questions: 1) How, and what assumptions are needed, to obtain approximately unbiased predictors of finite population totals or means when no or limited probability survey data is available? and 2) How to estimate the quality (prediction variance) of these predictors? This project is a follow-up of a previous project by DaSylva, Beaumont, Bosa and Maranda (2023).

Progress:

We considered two cases. In the first case, auxiliary data for a non-survey period are available in a probability sample, but no survey data is observed in that probability sample. We showed that (approximately) unbiased predictors of finite population totals or means can be obtained using a linear working model, even under deviations from the linearity assumption. This surprising property requires each observation in the training sample to be suitably weighted so as to ensure a certain calibration constraint is satisfied. We also developed a simple estimator of the prediction variance. We showed that this weighted linear predictor has properties similar to a nearest-neighbour predictor, which explains its robustness. A few simulation results have been obtained to illustrate the properties of predictors. We also started the development of a bootstrap strategy for estimating the prediction variance when machine learning predictions are used as an alternative to predictions from a linear working model or nearest-neighbour predictions.

In the second case, auxiliary data are available for the entire population along with survey data in a small probability sample. The availability of survey data allows us to make valid inferences without relying on model assumptions. A new approach, called prediction-powered inference, has recently been proposed to handle this case when observations are independent and identically distributed. It consists in eliminating the bias of machine learning predictions by leveraging the observed survey data. Prediction-powered inference is shown to be essentially equivalent to the well-known model-assisted approach under simple random sampling.

Our theoretical and empirical results will be presented at the 2025 Italian Conference on Survey Methodology in July 2025.

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

Reference

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.

1.2 Record linkage

PROJECT: Private linkage of international trade microdata with linkage errors

New insights could result from the linkage of international trade microdata across national statistical agencies. Privacy concerns currently stand in the way, which may be addressed by leveraging privacy enhancing technologies such as secure multiparty computation, private set intersection or secure enclaves. In previous work, a solution was proposed based on modifying an existing private set intersection protocol (Dasylva, De Cubellis, De Fausti and Franssen, 2025). This solution links two international datasets without a unique identifier, it assesses the rates of linkage errors with a model, and it computes summary statistics while adjusting for the linkage errors. However, it has many shortcomings, such as the limited linkage flexibility, the constraints on the possible statistical analysis and the lack of disclosure control. This work aims to address these issues by linking the datasets within a cloud-based secure enclave.

Progress:

A methodology was developed to privately link the datasets with the probabilistic method while controlling the linkage accuracy, within a cloud-based secure enclave. Using the linked data, a linear or logistic regression is performed, and some synthetic data is generated to allow some data exploration. Furthermore, all these operations are performed through the application of differential privacy techniques to ensure that all the statistical outputs are protected against statistical disclosure. The methodology was successfully tested on a secure enclave in the cloud. The methodology and tests have been described in an internal report (Dasylva, Santos, Franssen, De Cubellis, De Fausti, Pappagallo, Berrios and Fitzsimons, 2025), which has been accepted for publication in the Statistical Journal of the International Association for Official Statistics.

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

References

Dasylva, A., De Cubellis, M., De Fausti, F. and Franssen, L. (2025). Linking trade data from different national statistical offices through a private set intersection. *Journal of Official Statistics*, 41(2), 569-597, OnlineFirst, Special Issue, https://journals.sagepub.com/doi/10.1177/0282423X251329407.

Dasylva, A., Santos, B., Franssen, L., De Cubellis, M., De Fausti, F., Pappagallo, A., Berrios, N. and Fitzsimons, J. (2025). Private linkage of international trade microdata in a cloud-based secure enclave. To appear in the *Statistical Journal of the International Association for Official Statistics*.

PROJECT: Confidence Intervals for Record Linkage Error Rates Under Stratified Systematic Sampling

Social Data Linkage Environment (SDLE) Methodology estimates error rates for its linkages using the following procedure: The set of definite and rejected pairs is divided into strata based on percentiles of the distribution of the total pair weight, and within those strata, samples (usually size 100) are drawn using systematic sampling (wherein the pairs are ordered according to total pair weight). The sample of pairs thereby selected is sent to reviewers, and based on their decisions, the various pairs in the sample are classified as matched or unmatched. Using these classifications, we estimate the total number of false matches or missed matches in each stratum. The estimated totals are used to calculate false match/missed match rates for each stratum and for the overall linkage.

We have recently decided on a variance estimator for our error rates from a research project in the spring of 2023 (Millar and Loewen, 2023). A next step that we identified when deciding upon an appropriate variance estimator was to estimate confidence intervals. There are two factors that make it tricky to create confidence intervals of our error rates:

- One is the sampling method that we use (stratified systematic sampling). Complex sampling designs need complex methods to construct confidence intervals as simple methods often have coverage issues, containing the true population parameter at a rate less than the nominal value. There is no consensus on the best method to construct confidence intervals for samples drawn from a stratified systematic sample.
- The other is that our error rates (false match and missed match rates) are often close to 0 (especially the false match rate). It is well known that simple confidence intervals such as Wald have especially poor coverage for proportions that are close to 0 or 1 (Franco, Little, Louis and Slud, 2019, and Neusy and Mantel, 2016). A more complex method is needed to create confidence intervals so that the coverage of the interval is at least at the nominal level. Strata with estimated variance of 0 also pose a problem as some confidence intervals would have no length in that case, when using a Wald-type confidence interval calculation.

The problem that we are facing is finding an appropriate confidence interval to use to indicate the quality of SDLE linkage error rates.

Progress:

Confidence intervals have been tested based on coverage of the true population parameter and the length of the confidence interval. In the end, the Agresti-Coull interval was chosen because it met the coverage requirement, it was not significantly longer than any other interval that also met the coverage requirement, and it was the simplest of those intervals. This is in line with Brown, Cai and DasGupta (2001) who report that the Agresti-Coull interval works well when the sample size is greater than 40.

Furthermore, a report (Loewen and Jin, 2025a) and two presentations were produced (Loewen and Jin, 2025b, 2025c). Loewen and Jin (2025a) detail the context of our error rate calculations, the work that was done to find the best confidence interval, and the results of our study. The first presentation (Loewen and Jin, 2025b) was given to the Scientific Review Committee. It details the process of finding a good confidence interval method. The second presentation (Loewen and Jin, 2025c) is a divisional seminar detailing our work and the results of our study.

For more information, please contact:
Rylee Loewen (rylee.loewen@statcan.gc.ca).

References

Brown, L., Cai, T. and DasGupta, A. (2001). Interval estimation for a binomial proportion. *Statistical Science*, 16, 101-117.

Franco, C., Little, R., Louis, T. and Slud, E. (2019). Comparative study of confidence intervals for proportions in complex sample surveys. *Journal of Survey Statistics and Methodology*, 7, 334-364.

Loewen, R. and Jin, N. (2025a). Confidence intervals for Social Data Linkage Environment Error Rates. Internal document, Statistics Canada.

Loewen, R. and Jin, N. (2025b). Confidence intervals for Social Data Linkage Environment Error Rates. Slides for presentation to the Scientific Review Committee, May 23, 2025, Statistics Canada.

Loewen, R. and Jin, N. (2025c). Confidence intervals for Social Data Linkage Environment Error Rates. Slides for Divisional Seminar, June 12, 2025, Statistics Canada.

Millar, G. and Loewen, R. (2023). Variance estimation for record linkage error-rates obtained via clerical review of systematic samples of linked pairs. Internal document, Statistics Canada (Document Overview: Variance estimation for record linkage error.docx).

Neusy, E. and Mantel, H. (2016). Confidence intervals for proportions estimated from complex survey data. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, June 2016.

1.3 Small area estimation

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

The Fay-Herriot (FH) area level model is often used when direct estimators of population parameters are unstable because of small sample sizes. The idea behind the FH model and other Small Area Estimation (SAE) models is to borrow strength from other domains. However, when the number of domains is small, the FH and other SAE models do not usually perform well. The idea of this project is to borrow strength not only 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:

We studied prior specifications for regression parameters in the area level models for small area estimation. More specifically, we investigated the use of noninformative and informative priors through an LFS application and a simulation study. A research paper (You and Bosa, 2025) was completed and

submitted to the *Survey Methodology* Journal. The paper has been revised according to the suggestions from the Associate Editor and referees. The final version of the paper has been accepted by the journal and will be published in its December 2025 issue.

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. and Bosa, K. (2025). Performance of hierarchical Bayes small area estimators using noninformative and informative priors with an application to the Canadian Labor Force Survey. To appear in the December 2025 issue of *Survey Methodology*.

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: 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 strongly 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:

Previously, we developed and evaluated through empirical studies a non-parametric version of the Empirical Best (EB) predictor when random forests are used to replace fully parametric models. Our proposed 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.

In 2024-2025, this project was presented at the International Conference on Establishment Statistics in Glasgow in June 2024, and at the CANSSI conference in Ottawa in July 2024. We also wrote a paper (Bosa, Beaumont, Bocci and Sombo, 2025) that will be presented at Statistics Canada's Advisory Committee on Statistical Methods in June 2025, and we are currently developing an R program that implements the proposed methodology.

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

Reference

Bosa, K., Beaumont, J.-F., Bocci, C. and Sombo, S. (2025). The use of a random forest algorithm in small area estimation. Paper presented to the Advisory Committee on Statistical Methods, June 2025, Statistics Canada.

PROJECT: Reverse-Engineering a Hypothetical Raking for the Estimation of the Mean Squared Error of Raked Small Area Estimates

The area-level Fay-Herriot (FH) small area estimation (SAE) model assumes independence of the area direct estimates. It was observed in the Labour Force Survey (LFS) context for the estimation of total employment for a mapping of the provinces that direct estimates are very negatively correlated because of weight calibration. To support surveys in the same situation as the LFS, this project focuses on extensions of the basic FH model to the case of a strong dependence. In particular, it aims at obtaining a good estimate of the design-variance-covariance matrix of the direct estimates, which is assumed to be known in the FH model. To achieve this objective, a hypothetical raking process is reverse engineered to obtain the covariance terms given the (smoothed) variance terms and a variance-covariance matrix of aggregated direct estimates (provincial estimates in the case of the LFS). Because total estimates are good candidates for applying a raking process, both unraked and raked small area estimates are studied.

Progress:

The mathematics of the reverse-engineering problem and its solution have been made explicit, resulting in a clearly defined working variance-covariance matrix. Existing SAE software such as G-Est only applies the basic FH model. To make the most of existing software, under the assumed working variance-covariance matrix, Mean Squared Error (MSE) formulas of raked and unraked estimators resulting from fitting the basic model were obtained. The methodology was applied for the first official dissemination of LFS SAE estimates. Results were presented at the 2024 International Methodology Symposium (Verret and Walker, 2024).

The FH model extended for dependent sampling errors was then considered to obtain more efficient estimators. Corresponding MSE formulas were developed for the unraked and raked estimators. The extended model was programmed, as well as many features of G-Est (backward variable selection, diagnostics, outlier detection). A collinearity diagnostic and the first diagnostic of Lesage, Beaumont and

Bocci (2022) were also generalized to the extended model. The advantages of the extended model over the basic model were studied in the context of the LFS.

For more information, please contact:
François Verret (francois.verret@statcan.gc.ca) or
Braedan Walker (braedan.walker@statcan.gc.ca).

References

Lesage, É., Beaumont, J.-F. and Bocci, C. (2022). Two local diagnostics to evaluate the efficiency of the empirical best predictor under the Fay-Herriot model. *Survey Methodology*, 47(2), 279-297. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2021002/article/00001-eng.pdf.

Verret, F. and Walker, B. (2024). Reverse-engineering a hypothetical raking process for the estimation of mean squared error of raked small area estimates. *Proceedings: Symposium 2024, The Future of Official Statistics*, Statistics Canada, Ottawa, Canada.

2 Data science methods and applications

PROJECT: Large Language Models (LLMs) as a Post-Processor for Optical Character Recognition (OCR) engines

In this project we examine the challenges and advancements in Optical Character Recognition (OCR) technology, particularly in the context of handwriting recognition, by exploring how Large Language Models (LLMs) can enhance OCR outputs through post-processing. Despite notable improvements in OCR systems, handwritten text continues to pose significant challenges due to variability in handwriting styles, leading to errors that impact downstream tasks such as text summarization and Named Entity Recognition (NER).

Progress:

In this research we evaluated the performance of various OCR tools for handwritten text recognition, finding that Transformer-based Optical Character Recognition (TrOCR) delivered the best results against popular open-source OCR tools such as PaddleOCR and EasyOCR while proprietary solutions like GPT-4o and Document Intelligence also provided the best results with the lower character error rates. The study demonstrated how large language models could significantly enhance OCR output quality by correcting grammatical inconsistencies and spelling errors, with GPT-4o as a post-processor which substantially reduced these errors found in the text extracted by Tesseract (a free, open-source OCR engine) on sample extractions from the IAM Handwriting Dataset.

Additionally, we explored and concluded that automatic prompt engineering frameworks such as DSPy proved more effective than manual prompting strategies to generate context rich and detailed prompts. While the combination of specialized OCR tools with language models showed considerable promise for improving recognition accuracy, complex error combinations remained challenging even for advanced LLMs, which excelled at simple corrections but struggled with combined word substitutions and more complex issues.

We recognized the popularity of using Vision LLM's as OCR tools over the past few months. Thus, we also explored the performance for open-source tools such as Florence and Phi-3.5 by Microsoft and closed source (GPT-4o, Document Intelligence) as Handwritten text recognition (HTR) tools for extracting handwritten and printed text. Both types of tools provided acceptable levels of extracted text. We hope to continue exploring these solutions as viable OCR tools combined with prompt optimization are part of our future work.

For more information, please contact:
Oladayo Ogunnoiki (oladayo.ogunnoiki@statcan.gc.ca) or
Johan Fernandes (johan.fernandes@statcan.gc.ca).

PROJECT: International Trade Unit Value (UV) Error Detection

This work was to complete and publish a paper detailing the methodology behind a new machine learning-based system for editing and imputing trade data, which has been in production since October 2023. The International Accounts and Trade Division (IATD) processes millions of monthly import records, including variables like Harmonized System (HS) codes, value, and quantity, with Unit Value (UV) derived from the latter two. Due to the focus on verifying value for tariff purposes, quantity fields are often misreported, leading to errors in aggregate quantities and UVs–issues that historically required significant analyst time to address.

The previous UV-clipping system, which replaced extreme values with mid-range donors, had several limitations including overcorrection, persistent unaddressed errors, high manual workload, and delayed usability. The newly implemented machine learning system, reviewed in 2022 and launched in 2023, addresses these issues and has been well-received by analysts. Formalising the documentation of the methods and publishing the work will enhance transparency as well as facilitate sharing findings to increase collaboration and knowledge sharing, especially since the use of machine learning in official statistics remains a relatively new phenomenon.

Progress:

The project was completed with the following milestones:

- A complete draft of the paper has been prepared (Hatko and Sall, 2024).
- Peer and institutional review is underway.
- Next steps include reviewing the work for confidentiality and publication.

For more information, please contact:
Stan Hatko (stan.hatko@statcan.gc.ca).

Reference

Hatko, S. and Sall, A. (2024). International Trade Unit Value Error Detection and Correction. Internal Report, Statistics Canada.

PROJECT: Metrics for AI-Generated Text

This project investigates how the quality of Natural Language Generation (NLG) outputs is evaluated across different tasks, methods, and use cases. As NLG technologies, particularly Large Language Models

(LLMs), become more prevalent, their outputs demand evaluation methods that are robust, interpretable, multidimensional and task-specific. However, no standard framework currently exists for designing evaluation pipelines. To address this, the project reviews existing evaluation techniques, introduces a conceptual taxonomy grounded in functional decomposition, and proposes a reusable evaluation design framework. The primary focus is on unimodal text-to-text generation tasks (e.g., summarization, machine translation, dialogue), with the goal of providing both analytical clarity and practical tools for future evaluation design.

Progress:

The research project is finalized. We surveyed the NLG evaluation literature, identifying critical limitations in current practices (surface-level bias of rule-based metrics, inconsistency in human evaluation protocols) and synthesizing the findings into a structured and extensible framework. The literature review was conducted using a PRISMA-style protocol to analyze 385 research papers, targeting mostly recent literature while also revisiting foundational work. The review spans rule-based, human-based, and LLM-based evaluation methods, emphasizing the multidimensional nature of quality in NLG. We propose a modular taxonomy that distinguishes between evaluation components such as scoring functions, input transformations, and evaluation targets. We designed a practical Evaluation Pipeline Design Framework (Istrate and Robatian, 2025), including a phased evaluation workflow, stage-specific checklists, and best-practice recommendations to improve usability and reproducibility while offering practitioner-friendly tools and artifacts that do not require deep theoretical knowledge. The objective is to start experimenting with the framework in future projects involving LLM generated text, continuously incorporating lessons learned and best practices and improve the framework artifacts.

For more information please contact:
Alexandre Istrate (alexandre.istrate@statcan.gc.ca) or
Damoon Robatian (damoon.robatian@statcan.gc.ca).

Reference

Istrate, A. and Robatian, D. (2025). Evaluation of AI-Generated Text: Survey, Taxonomy and Practical Framework for Designing NLG Evaluation Pipelines. Internal report, Statistics Canada.

PROJECT: Quantifying Uncertainty in LLM Classification with Conformal Prediction

The aim of this project is to enhance the reliability and accuracy of Large Language Models (LLMs) used by Statistics Canada for classifying tasks. By using Conformal Prediction (CP) techniques, we are looking to introduce quantifiable confidence measures for the predictions made by these models with theoretical guarantees. This research helps improve the risk management in decision-making processes that rely on LLMs.

Progress:

Our implementation of conformal prediction for LLM classifiers demonstrated an optimal trade-off between higher accuracy and automation for LLM classification tasks, using a guaranteed coverage of 90%. The LLM achieved a baseline accuracy of 77%, however, by applying CP to the model predictions and automating the coding/classification to the LLM when it produced highly certain predictions (size-1 prediction sets), 70.2% of the data was automatically classified at a higher accuracy of 89.8%. The remaining data could be automatically coded as well or sent to a human-in-the-loop for manual

verification. The results outperformed baseline approaches that do not come with theoretical guarantees. The technique provides statistically valid uncertainty quantification, automatically processing high-confidence predictions while intelligently routing ambiguous cases for manual review. As a next step, this approach could be applied to the North American Industry Classification System (NAICS) code classification based on business descriptions.

For more information, please contact:
Rafik Chemli (rafik.chemli@statcan.gc.ca).

PROJECT: Synthetic Data for Official Statistics

This project addresses the significant challenges Statistics Canada faces due to declining survey response rates, the high cost of data acquisition, and the increasing need for detailed statistics on rare subpopulations. These factors make it difficult to produce robust and accurate statistical estimates. The proposed solution involves leveraging state-of-the-art Denoising Diffusion Probabilistic Models (DDPMs) to generate large volumes of synthetic tabular data that closely mimic real-world data distributions. Recognizing that synthetic data will introduce biases, the project will integrate a debiasing process, inspired by Prediction-Powered Inference (PPI) and the generalized difference estimator.

Progress:

The project was completed with the following milestones.

- Tabular diffusion models were trained and applied to three open-source tabular datasets.
- The synthetic data generated were analyzed and compared relative to the original data, exhibiting strong distributional similarities, while producing completely unique data points.
- Non-probability sampling based estimates using synthetic data produced estimates with lower relative bias with similar or slightly reduced coverage levels as compared to baselines.
- For the design-based probability sampling setting, a mathematical two-stage debiasing estimator was proposed. However, this estimator is not unbiased, and the variance estimator is not unbiased, and is also intractable, leaving this approach invalid for design-based estimation.
- We conclude that diffusion models are powerful generative AI tools that may be leveraged for dissemination of micro-level synthetic data. Experiments suggest they maintain statistical properties of the underlying true data distribution. However, finding an appropriate estimator for the probability design setting for population parameter estimation may be out of reach.

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

PROJECT: Prompt Engineering Guide for Effective Use of Large Language Models (LLMs)

The Prompt Engineering Guide focuses on the development of a comprehensive guide on prompt engineering to support data scientists, researchers, and analysts in effectively using Large Language Models (LLMs). The goal is to demystify prompt engineering and provide best practices grounded in both empirical insights and operational needs within the federal data science community. It serves as a practical resource to enhance the reliability, transparency, and efficiency of LLM-driven workflows in support of federal programs, policymaking, and public service delivery.

The guide covers fundamental concepts, the anatomy and structure of prompts, and introduces key prompting techniques such as instruction prompting, role prompting, and few-shot prompting.

Progress:

The document has been created, and the content has been structured into key sections, including: (1) Fundamentals of Prompting, (2) Prompt Structure, (3) Prompting Techniques (Instruction, Role, Few-shot), and (4) Best Practices and Pitfalls. Each section is supported with practical examples and visual illustrations to aid understanding and usability. Feedback is currently being collected from multiple internal groups, including Technical, Responsible AI, and AI Capacity and Program Development teams. The guide will be revised and updated in the coming months based on this feedback to ensure broad applicability and alignment with organizational priorities.

For more information, please contact:
Saptarshi Dutta Gupta (saptarshi.dutta-gupta@statcan.gc.ca).

PROJECT: Creating Structured Dataset from Unstructured Data using Large Language Models

This project explores the use of Large Language Models (LLMs) to extract structured data from unstructured PDF documents. It evaluates two major approaches: (1) fine-tuning LLMs for direct data extraction, and (2) using Retrieval Augmented Generation (RAG) to retrieve relevant text snippets before structured extraction. The goal is to create high-quality datasets that enable further machine learning applications and statistical analysis.

The research aims to compare the performance of fine-tuned and RAG-based methods using the CORD-19 dataset (Wang, Lo, Chandrasekhar, Reas, Yang, Burdick, Eide, Funk, Katsis, Kinney, Li, Liu, Merrill, Mooney, Murdick, Rishi, Sheehan, Shen, Stilson, Wade, Wang, Wang, Wilhelm, Xie, Raymond, Weld, Etzioni and Kohlmeier, 2020)–a corpus of over 1,000,000 scholarly articles related to COVID-19. Example queries (e.g., “What was the proportion of all COVID-19 patients who were asymptomatic?”) are used to evaluate extraction accuracy against known ground truth.

Progress:

Methodology: The CORD-19 dataset has been utilized to develop and test both fine-tuned and RAG-based approaches for structured data extraction. Ground-truth questions were designed to enable performance evaluation. Fine-tuning was conducted using models such as gpt-4o-mini and gpt-4.1-mini, while custom RAG pipelines were implemented in parallel.

Results: Evaluation shows that the fine-tuned models produce more consistent and quantitative responses focused on specific outcomes. In contrast, the RAG outputs provide broader, more descriptive responses, sometimes mixing qualitative and quantitative details. Although both methods deliver comparable answers in many cases, the variance in response quality depends on the query type. The next step involves testing the framework across a more diverse set of unstructured sources to validate the generalizability and robustness of the results.

For more information, please contact**:**
Saptarshi Dutta Gupta (saptarshi.dutta.gupta@statcan.gc.ca).

Reference

Wang, L.L., Lo, K., Chandrasekhar, Y., Reas, R., Yang, J., Burdick, D., Eide, D., Funk, K., Katsis, Y., Kinney, R., Li, Y., Liu, Z., Merrill, W., Mooney, P., Murdick, D., Rishi, D., Sheehan, J., Shen, Z., Stilson, B., Wade, A., Wang, K., Wang, N.X.R., Wilhelm, C., Xie, B., Raymond, D., Weld, D.S., Etzioni, O. and Kohlmeier, S. (2020). CORD-19: The COVID-19 Open Research Dataset. *ArXiv preprint*, https://doi.org/10.48550/arXiv.2004.10706.

PROJECT: Input/Output Privacy Framework – Privacy-Preserving Record Linkage

The Input/Output Privacy Framework project aims to design and implement robust, end-to-end privacy-preserving pipelines for privacy-sensitive data analysis and record linkage. The central goal is to allow two or more parties to collaboratively analyze and link datasets (such as personal or organizational records) without leaking any sensitive individual-level information. The project focuses on privacy both at the input/ingest stage (ensuring data is protected during transfer and computation) and at the output stage (ensuring no disclosure occurs in published results), leveraging Privacy Enhancing Technologies (PETs), especially Homomorphic Encryption (HE), and traditional Statistical Disclosure Control (SDC) methods.

Two main technological streams are explored: a cryptography-based pipeline using homomorphic encryption and secure multi-party computation, and an enclave-based pipeline deploying secure hardware environments. The primary use case demonstrating these technologies is Privacy-Preserving Record Linkage (PPRL), where statistical summaries can be computed on the securely linked dataset. This framework is highly relevant for settings such as healthcare, finance, or official statistics, where organizations must comply with stringent privacy regulations.

Progress:

Significant progress was achieved last year on the design, prototype implementation, and experimental validation of the cryptography-based pipeline. An end-to-end privacy-preserving record linkage protocol was developed, combining advanced homomorphic encryption techniques with output privacy safeguards.

A demo prototype was implemented that allows a client and server to securely determine record overlaps (exact or fuzzy matches) and compute encrypted statistical summaries (e.g., average by strata) without either party learning the other's sensitive data. Critical subprotocols ensure that intersection sizes below a set threshold are suppressed to prevent privacy attacks, and summary outputs remain safe as dictated by disclosure control guidelines. Experimental evaluation on synthetic datasets demonstrated the system’s feasibility for moderate-sized datasets, accurate statistical output, and reasonable computational performance. Additionally, this framework is designed to be modular and extensible for future applications and core outputs. Documentation and a technical report (Shukla, Rossiter and Santos, 2024), as well as the demo implementation, are available for further exploration and benchmarking.

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

Reference

Shukla, A., Rossiter, E. and Santos B. (2024). Input/Output Privacy Framework: A Privacy-Preserving Record Linkage System. Internal Report, Statistics Canada.

PROJECT: Synthetic Data Release with Differential Privacy Guarantees – An Alternative to Public Use Microdata Files

This project explores the generation of synthetic data with differential privacy guarantees as a robust alternative to traditional Public Use Microdata Files (PUMFs) for data dissemination. The objective is to enable researchers, data scientists, and the public to conduct meaningful analyses on realistic synthetic data while rigorously protecting individual confidentiality. This initiative aims to compare the

effectiveness of synthetic data with differential privacy guarantees against conventional disclosure control techniques currently employed in PUMFs. While synthetic data is recognized as a safer mechanism for data sharing, recent research indicates that only differential privacy provides strong, provable protections against reidentification on synthetic data and potentially against emerging threats, such as those posed by large language models.

The framework focuses on widely used Canadian PUMFs released by Statistics Canada under the Data Liberation Initiative. By leveraging advanced generative models–including diffusion models and generative adversarial networks–and robust statistical and privacy evaluation methodologies, the project aims to produce differentially private, protected synthetic datasets with high analytical validity. A crucial component is the development of an open-source Python package, inspired by the *Folktables* toolkit (Ding, Hardt, Miller and Schmidt, 2021), to provide synthetic data for benchmarking machine learning models and facilitating reproducible research.

Progress:

The project has completed a thorough literature review, identified key PUMFs for experimentation, and surveyed state-of-the-art methods for both differential privacy and generative synthetic data. The team has initiated the reimplementation of generative models (including diffusion models and generative adversarial networks) integrated with differentially private mechanisms, alongside the parallel development of an open-source Python toolkit to supply synthetic datasets from PUMFs for benchmarking and analysis. Early-stage evaluations focus on key utility and privacy metrics, examining the trade-off between analytical usefulness and confidentiality, while addressing the ongoing challenge of ensuring the generated synthetic data maintains validity for typical demographic and socioeconomic machine learning tasks.

A fair comparison of privacy-utility between PUMFs and synthetic data will require training synthetic data generators using deidentified microdata, a step that will be completed once access is secured. Forthcoming deliverables include a technical report and an open Python software package for public use and benchmarking.

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

Reference

Ding, F., Hardt, M., Miller, J. and Schmidt, L. (2021). Retiring Adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

PROJECT: Disclosure Control: Limitations and Alternatives

Statistics Canada in its role as the national statistical agency acts as an intermediary to collect and/or aggregate data from various sources and compile them into summary tables as well as more detailed datasets. There is always a risk of disclosing personal or sensitive business information while providing these tables. To prevent leaking sensitive information the agency deploys programs (like G-Confid) to identify and supress sensitive cells in a contingency table or more detailed datasets like Public Use Microdata Files (PUMF) or census microdata files. The purpose of this project was to identify the vulnerabilities in these methods given that the computational landscape has changed. This is not only due

to the arrival of large language models and low bar of entry tools, but also due to rise in computational power making certain reverse computations not too challenging.

Progress:

This project is completed, and a report has been submitted. The project mostly addressed the Common Output Data Repository tables and how to access supressed values. We have summarized various attack methods ranging on the expertise of the adversary. In addition, we have provided some recommendations to address these vulnerabilities. As for PUMF files, we have identified some initial attacks but would require permission to scrape the internet for cross-referencing individual information on public sites such as Facebook and LinkedIn. Thus, we have only provided a methodology (no result) for PUMF datasets.

For more information, please contact:
Abhishek Shukla (abhishek.shukla@statcan.gc.ca).

3 Estimation issues in surveys

PROJECT: Calculation of theoretical degrees of freedom for modified Wilson confidence intervals in the context of the long-form questionnaire of the Canadian population census

For users of census population data in Canada, confidence intervals can be used to measure the accuracy of an estimate and for decision-making. To produce these intervals, a certain distribution is assumed, in our case, the Student's t-distribution. This distribution has a degree of freedom parameter, which is normally approximated by the number of replicate weights in the sample. This approximation can lead to undercoverage of confidence intervals, usually observed in small domains. The objective of the project is to implement the theoretical formula that was developed in the article by Toupin and Martin (2025), that is, to find an approximation or an estimator based on the theoretical developments, which would allow implementation in an automated system used in the various outputs of the census program.

Progress:

In the past year, algebraic developments of several approximations have been made in order to evaluate their performance. A Monte Carlo simulation environment was developed in the SAS software to assess the different approximations. It was found that several of them improved the confidence interval coverage for estimating counts from the long-form questionnaire, compared to the fixed value currently implemented in the system, but rarely enough to reach the targeted nominal rate. The best performing approximation offers gains that are too limited to justify the complexity of its integration into the dissemination system.

For more information, please contact:
Samer Farfour (samer.farfour@statcan.gc.ca).

Reference

Toupin, M.-H. and Martin, V. (2025). 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* (under revision).

PROJECT: Bootstrap estimation of variance for calibration estimators with constraints on the range of weights

Survey estimates are often calibrated to match known population totals, called control totals. The Generalized Regression (GR) estimator is a commonly used calibration estimator that assumes a linear relationship between the survey variable and the auxiliary variables. The theory underlying the GR estimator is well established, and there exist linearization variance estimators. Deville and Särndal (1992) showed that for a family of calibration estimators, all members of the family are asymptotically equivalent to the GR estimator. This suggests the use of the variance estimator of the GR estimator, or an adaptation of it, for other calibration estimators. This type of variance estimator is generally implemented in practice.

Certain calibration methods, such as ridge calibration, have constraints on the range of weights. Linearization in the presence of bounds on the weights is not always straightforward, and the standard variance estimator of the GR estimator may not be suitable. The objective of this research was to find another variance estimator in this situation.

This project started during the 2023-2024 year where we developed a bootstrap variance estimator that properly takes into account the bounds on weights and the control totals.

Progress:

We developed and completed a new simulation study in which we compare four bootstrap approaches to estimate the variance of a ridge calibration estimator. We considered two different populations and four sample sizes. The results of this study highlight the importance of using adjusted constraints and properly accounting for control totals to obtain the bootstrap version of the ridge calibration estimator, and thus approximately unbiased estimators of the variance of the ridge calibration estimator.

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: Study on the effects of variable selection methods and parameter tuning on the cubist method with and without pre-processing

In the Economic Statistics Methods Division (ESMD), and particularly in the Integrated Business Statistics Program (IBSP), classical imputation methods are currently used to impute both item and unit nonresponse. Despite individual methods being simple, such as donor or ratio imputation, tens or sometimes hundreds of these methods are overlayed to create complex edit and imputation strategies that become difficult to implement, maintain and troubleshoot. In 2023-24, under the Methodological Accelerator Program (MethAx2023-24), we carried out exploratory research where the performance of linear regression, random forest and cubist methods was investigated following the results of simulation studies carried out by Dagdoug, Goga and Haziza (2023), where the authors reported, among other results, that the cubist method (Kuhn and Johnson, 2016; and Kuhn, 2022) worked well in a variety of situations, and the random forest method performed well in high dimensional settings. Our MethAx2023-24 research also indicated that cubist methods performed well among the above-mentioned methods. In

this research, we compared the performance of the cubist imputation method under different scenarios to identify processes that may improve its performance.

Progress:

To understand the performance of the cubist imputation method, we carried out a study to understand the effects of feature selection methods such as removing near-zero variance variables, variables that are linear combination of others, highly correlated variables and variables that are susceptible to multicollinearity, which gave four different scenarios. For each scenario, we studied the impact of parameter tuning with and without feature selection and with and without tolerance for improvement. This resulted in 16 different scenarios (configurations). Additionally, we fitted the cubist method using two different R packages, namely caret and tidymodels, which gave 32 different configurations. Furthermore, to understand whether data pre-processing improves the performance, we applied these scenarios on non-standardized and standardized datasets, where all the variables are standardized, and datasets that contained only the important variables outputted by the previously fitted cubist algorithm.

We selected a few target variables from three different Statistics Canada's economic surveys of different sizes. Predictors are variables on the frame in the current year and all the variables from the previous year. Final dataset contained the respondents from the current and the previous year. Results were obtained under 10-fold cross validation. Our reproducible R program outputs predicted values, root mean square error, bias, variance, mean and median absolute error, mean absolute error percentage, coefficient of determination, computing time, number of committees and neighbors, where applicable by configuration and fold under different pre-processed datasets. Our results indicate that carrying out some data cleaning activities using the variable selection methods listed above, in combination with tuning of the parameters (but not the variation in tolerances), and/or the data pre-processing activities mentioned above improved the performance of the cubist models for some variables.

For more information, please contact:
Ahalya Sivathayalan (ahalya.sivathayalan@statcan.gc.ca).

References

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

Kuhn, M. and Johnson, K. (2016). *Applied Predictive Modelling*. Springer, 69-71.

Kuhn, M. (2022). caret: Classification and Regression Training. R package version 6.0-93, https://cran.r-project.org/package=caret.

PROJECT: Bayesian Nonparametric Methods for Survey Data

Application of Bayesian nonparametric methods is an important area of research for statistical agencies. In our work, we study the application of techniques based on Dirichlet processes (DP) to complex survey data. We propose a novel approach that combines existing methods, used by Statistics Canada, for the derivation of survey weights for stratified, clustered, unequal probability sampling designs, and the flexibility of nonparametric Bayesian methods based on DP.

We have developed software and applied such models to some of the datasets collected during Cycle 6 of the Canadian Health Measures Survey (CHMS) and conducted several simulation studies to show the robustness of our models under various scenarios as well as applicability of the models to a possible production of official statistics.

Progress:

The project was completed. The obtained results indicate that the proposed method has a significant potential for further research. Software was developed and applied to practical models. A paper (Volkov, 2025) was published in The Survey_Statistician.

For more information, please contact:
Oleksii Volkov (oleksii.volkov@statcan.gc.ca).

Reference

Volkov, O. (2025). Bayesian nonparametric methods for survey data. *The Survey Statistician*, 91, 34-43, https://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2025_January_N91.pdf.

PROJECT: Subsampling for Non-Response Follow-Up in Social Surveys

In the context of the challenges of both falling response rates and the high cost of interviewer-facilitated follow-up, it is important to consider how much non-response follow-up is required and how to use it efficiently. Selecting only a subsample of a larger main sample for deeper ("full") non-response follow-up while doing lesser ("reduced") follow-up on the remainder offers the opportunity to reduce the amount of money spent on interviewer-facilitated follow-up and then to reallocate this budget, or at least part of it, to more thorough non-response follow-up on the full follow-up part of the sample. This can help address the increased potential for bias associated with low response rates, or can aim to be done without compromising the quality of estimates. This type of survey design requires adjustments to the weighting process; the Concurrent Multi-Mode Estimator (CMME) has been developed specifically for this context (Mather, Boulet and Brennan, 2024). This project expands on previous theoretical work by exploring its properties empirically via pilot studies and evaluating potential future applications through simulations.

Progress:

The approach of selecting a random subsample for full follow-up and then applying the CMME was investigated using both an empirical trial and simulations (Boulet and Mather, 2025). A pilot was run on the Canadian Social Survey. A usual wave of this survey has a sample size of 20,000 units, all of which are eligible for non-response follow-up by telephone; in the pilot, the number of units eligible for this follow-up was reduced to 17,000, but the overall sample size was increased to 34,000. As in a usual wave, all units were eligible for non-response follow-up through mailed reminders, emails, and SMS messages. By design, the overall collection cost of this trial was approximately the same as a usual wave. The result was an increase in precision compared to a usual wave (larger effective sample size, lower standard error on common variables, and an increase in the number of domains that meet dissemination guidelines). Further, although there were signs that the approach of random subsampling and applying the CMME introduces some bias, the magnitude of this potential bias is small and is deemed to be within acceptable ranges, particularly when considering the gains in precision.

Additionally, a series of planning worksheets were developed to model the costs and yields of various survey contexts. Based on assumptions coming from empirical results and simulations, the model accounts for the relationships between expected return rates for the subsamples with full and reduced follow-up, the relative costs associated with full and reduced follow-up, and the increase in design effect due to subsampling and applying the CMME. The results suggest that this approach could be effective in certain contexts to either increase precision for the same collection costs or reduce collection costs for the same level of precision. The approach seems particularly applicable for surveys with moderate but not extreme differences in both cost and response rate between full and reduced follow-up. This is the case for many Statistic Canada surveys, such as those in which reduced follow-up consists of mail and email and full follow-up adds telephone attempts. The model and associated planning worksheets are ready to be used to assess the applicability of the subsampling/CMME approach to specific surveys interested in implementing the idea.

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

References

Boulet, C. and Mather, A. (2025). Methodology Acceleration Initiative Research Project: Subsampling for Non-Response Follow-Up in Social Surveys. Internal document, 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.

PROJECT: Probabilistic Web Panels at Statistics Canada

In 2020, Statistics Canada started to use probabilistic web panels as an alternate method of collecting official statistics. In a web panel, respondents to large probabilistic social surveys (“recruitment surveys”) are invited to provide contact information to participate in future short surveys. The web panels can be used to collect information quickly and are therefore a useful tool for providing timely data. However, the low cumulative response rates associated with this method limit the precision of web panel estimates and raise concerns about potential bias. This project reviewed five years of panel experiences at Statistics Canada for insights into panel recruitment, collection, and estimates.

Progress:

The first component of the project investigated ways in which to effectively recruit panel participants and subsequently collect data using panel surveys. The manner in which recruitment questions are presented can result in very different rates of participation. A test of alternative question presentations showed that the presence of an explicit opt-in question had a large impact, with 26% of respondents providing their contact information (email and/or phone) when they were explicitly asked on a first screen to opt-in to future panels prior to having the chance to provide contact information, compared to 88% who gave contact information without the opt-in question. Moreover, the wealth of auxiliary information available on the recruitment survey can be used to actively manage panel collection operations, by predicting the probability of response and using this information to target follow-up efforts. When employed, this approach made it possible to decrease the range of response rates between groups with initially lower

response rates (for example, younger age groups, or those with lower educational attainment) and higher-response groups. For more detail on this work, see MacIsaac, Boulet and Thomas (2024).

The auxiliary information from the recruitment surveys can also be used to develop rich models that can be used in the weighting process for panels. The second component of this project studied how effective this is in reducing potential bias associated with the low response rates achieved on panels. Estimates of recruitment survey variables were calculated using both recruitment survey weights and web panel weights, and these were compared; differences signal the possibility of residual bias that was not corrected by the web panel weighting process. This investigation found more significant differences than would be expected if the web panel estimator fully corrected for the bias resulting from the web panel response process. Questions related to certain topics such as politics and voting, sense of belonging, and media consumption were found to have the most significant differences between web panel estimates and recruitment survey estimates. For more details on this work, see Mather and Boulet (2024) and Mather, Boulet and Ra (2024).

The results of this project were presented at the 2024 International Methodology Symposium on "The Future of Official Statistics".

For more information, please contact:
Cilanne Boulet (cilanne.boulet@statcan.gc.ca).

References

MacIsaac, K., Boulet, C. and Thomas, M. (2024). Recruitment and collection of Web panels at Statistics Canada. *Proceedings: Symposium 2024, The Future of Official Statistics*, Statistics Canada, Ottawa, Canada.

Mather, A. and Boulet, C. (2024). *Analysis of Residual Bias on Web Panels*. Methodology Branch Working Paper (SSMD-2024-002E), Statistics Canada.

Mather, A., Boulet, C. and Ra, Y. (2024). A bias evaluation for probabilistic Web panels at Statistics Canada. *Proceedings: Symposium 2024, The Future of Official Statistics*, Statistics Canada, Ottawa, Canada.

PROJECT: The Evolution of the Quantity and Quality of the Data Collected for the Survey of Household Spending Through the Years

Social survey response rates have declined significantly over the past 25 years, a trend that was further accelerated by the COVID-19 pandemic and the shift away from in-person interviews. While this decline is well-documented, recent initiatives have primarily focused on understanding and addressing response rates. However, the impact of growing public disengagement with statistical programs may extend beyond just lower participation–it could also affect the quantity and quality of information provided by those who still respond and compromise the validity and accuracy of published results if existing methodological safeguards are not robust enough.

Unlike overall non-response, this aspect of disengagement is more difficult to measure, requiring a deeper investigation. This research aims to assess changes in the quantity and quality of data provided in household surveys, with a specific focus on the Survey of Household Spending (SHS). By comparing key indicators over time, this study hopes to establish a framework that other household surveys can adopt, ultimately leading to broader insights into the evolving challenges of survey data collection and processing.

Progress:

The quantity and quality of respondent-provided information were assessed using various indicators from multiple cycles of the Survey of Household Spending (2015-2023). Where possible, indicators were broken down by collection mode to evaluate the impact of changes in collection strategy.

This study found no clear trend of increase or decline in data quantity and quality over the last five SHS cycles. While growing respondent disengagement has impacted survey non-response, it has not obviously affected item non-response or the completeness and accuracy of responses. Overall, indicators suggest respondents have maintained consistent answering patterns over time. However, these findings may apply specifically to the SHS, and other social surveys should conduct their own analyses, using this study as a template, to determine whether respondent disengagement has influenced their own data quality.

For more information, please contact:
Yves Lafortune (yves.lafortune@statcan.gc.ca).

Reference

Lafortune, Y. and Mayer, E. (2025). The Evolution of the Quantity and Quality of the Data Collected for the Survey of Household Spending Through the Years. Research Paper, 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

As indicated in the previous report, 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.

Progress:

A paper (Tang, 2024) outlining the simulation process and discussing the justifications for proposed confidentiality rules was presented at the 2024 Methodology Symposium and will be published as part of the proceedings.

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

Reference

Tang, C. (2024). Statistical disclosure control analysis for small area estimation. *Proceedings: Symposium 2024, The Future of Official Statistics*, Statistics Canada, Ottawa, Canada.

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:

Based on the success of releasing the Synthetic database for the PASSAGES dynamic microsimulation model, a draft version of an update to internal guidelines for the creation of synthetic data files was prepared (Gauvin, 2025). This will replace the existing documentation used to guide developers in the creation of synthetic data.

Gauvin (2024) presented her work on developing the synthetic data for the PASSAGES model at the 2024 Statistical Society of Canada's Annual Meeting.

Yu (2024) presented a review on synthetic data disclosure risk assessment as part of the 2024 International Methodology Symposium.

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

References

Gauvin, H. (2025). Practical Guidelines for the Creation of Synthetic Data Files. Internal document, Statistics Canada.

Gauvin, H. (2024). Creating a synthetic version of a longitudinal and structured file: Challenges and lessons. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, St. Johns, NF, https://ssc.ca/sites/default/files/imce/gauvin_ssc2024.pdf.

Yu, Z. (2024). Synthetic data disclosure risk assessment. *Proceedings: Symposium 2024, The Future of Official Statistics,* Statistics Canada, Ottawa, Canada.

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.

Progress:

Statistics Canada has successfully built a beta, Python-based solution of G-Confid. That includes an optimal additive rounding solution, a tool for the calculation of sensitivity measures, the creation of an optimal CCS pattern, and an audit program to ensure that patterns are valid. Open-source solutions available through the PuLP Python package were studied as potential replacements and with careful implementation applied on the most complex cases at Statistics Canada.

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

PROJECT: A Utility-Disclosure Risk Framework for Comparing Statistical Disclosure Control Mechanisms

As a National Statistical Organization (NSO), Statistics Canada has been tasked with finding ways to release data at more detailed levels. There is a particular interest in granting statistical analyses on small subpopulations at finer geographies. This is referred to as the Disaggregated Data Action Plan (DDAP) initiative. However, data dissemination must always be done in a manner that complies with the confidentiality provisions of the Statistics Act. A key component of releasing confidentiality compliant statistical output is to apply Statistical Disclosure Control (SDC) to statistical output prior to its release.

Progress:

To better suit the needs of users, an analysis of the various SDC methods is being researched, with the goal of measuring the utility and disclosure risk associated with them. A utility-disclosure risk trade-off framework is being proposed for comparing the presented SDC methods. By applying the framework, a data disseminator can select an SDC method that optimally balances disclosure risk and utility in their given situation. An internal working paper is in development and will be considered for external publication.

For more information, please contact:
Joshua Miller (joshua.miller@statcan.gc.ca).

PROJECT: A Utility-Disclosure Risk-Based Impact Assessment of the Implementation of Differential Privacy in the Context of Tabular Data Dissemination

Differential Privacy (DP) is a framework for Statistical Disclosure Control (SDC) that seeks to constrain the level of private information leakage that would be incurred based on any single individual's contribution to a private database after statistical output is released. This framework was first studied in Dwork, McSherry, Nissim and Smith (2006). A DP compliant SDC framework has been proposed by academics and government agencies. Specifically, methods that lend themselves to the framework have been employed as alternatives to traditional SDC methods such as rounding and cell suppression. For example, the United States (U.S.) Census Bureau has implemented DP algorithms as a method to publish results for their 2020 Census as highlighted in Abowd, Ashmead, Cumings-Menon, Garfinkel, Heineck, Heiss, Johns, Kifer, Leclerc, Machanavajjhala, Moran, Sexton, Spence and Zhuravlev (2022). It is therefore of utmost interest for Statistics Canada to fully understand the DP framework and invest in areas where it could translate into great improvements for its data dissemination ecosystem.

Although DP has been known to provide strong privacy guarantees, it is not without its criticisms. A common criticism of DP is the impact that the methods have on the utility of the published output. Moreover, delineating an appropriate trade-off between utility and disclosure risk can be difficult. In the DP context, this translates to determining an appropriate privacy budget. Furthermore, if improperly implemented, DP can fail to protect sensitive cells. Thus, an assessment on the utility of DP compliant output as well as a closer study of the challenges of DP is warranted. This also entails a comparison of DP compliant methods to traditional SDC methods.

Progress:

A detailed paper was written that summarized the practical challenges that are faced when adopting a DP framework (Miller, 2025). This paper includes a summary of DP and analysis of its core properties. A comparison of the different definitions of DP is also included. Moreover, an intuitive explanation of what

guarantees a DP framework provides is discussed. Additionally, traditional SDC methods, such as random rounding, are examined through the lens of DP. A major practical barrier that disseminators face when considering the adoption of DP is the selection of an appropriate privacy parameter. This issue is discussed at length in the paper with some suggestions proposed.

A DP compliant SDC algorithm was tested on 2021 Canadian Census data to demonstrate what the adoption of DP could look like in practice. The algorithm that was implemented was based off the U.S. Census Bureau TopDown algorithm. The TopDown algorithm was first applied to 2020 American Census data as discussed in Abowd et al. (2022). A key finding of the Canadian Census case study was that the utility of statistical output can be significantly compromised if the goal of the disseminator is to provide meaningful privacy guarantees.

The DP paper by Miller (2025) is under peer review and will be officialised in the 2025-2026 fiscal year as a branch working paper. The research results were also shared in the form of a methodology research seminar.

For more information, please contact:
Joshua Miller (joshua.miller@statcan.gc.ca).

References

Abowd, J.M., Ashmead, R., Cumings-Menon, R., Garfinkel, S., Heineck, M., Heiss, C., Johns, R., Kifer, D., Leclerc, P., Machanavajjhala, A., Moran, B., Sexton, W., Spence, M. and Zhuravlev, P. (2022). The 2020 Census disclosure avoidance system TopDown algorithm. *Harvard Data Science Review*, (Special Issue 2).

Dwork, C., McSherry, F., Nissim, K. and Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. *Journal of Privacy and Confidentiality*, 7(3), 17-51.

Miller, J. (2025). A Privacy-Utility Based Study of Differential Privacy. Internal report, Statistics Canada.

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, benchmarking, and reconciliation during the year to internal and external participants through the Statistics Canada training centre (Statistics Canada, 2024). In particular, the introductory course on time series components and seasonal adjustment was offered to members of the Lao Statistics Bureau. The Centre also participated in outreach and ad hoc training to other groups in Statistics Canada on time series topics (methodology branch seminar series for recruits, data navigator course and seasonal adjustment dashboard demonstrations).

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 quarterly and monthly Gross Domestic Product programs, balance of payments, international trade in services, and securities transactions. 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. Work on producing guidelines on time series continuity continued and was integrated into a larger project on updating and expanding guidelines on time series methods for Statistics Canada’s programs. 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.

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 reconciliation (raking and balancing) (Statistics Canada, 2016; Ferland, 2025), the Time Series Processing System (Ferland, 2022), and the Seasonal Adjustment Dashboard (Verret, 2021).

The remaining work on a new, open-source, version of G-Series in R was completed, with the addition of the balancing functionality, utility functions, and expanded bilingual documentation. The new version of G-Series is developed and maintained in Statistics Canada's internal GitLab environment using Continuous Integration/Continuous Delivery (CI/CD) best practices. It underwent alpha and beta testing, as well as a code review and a cybersecurity application vulnerability assessment. The new G-Series will be released officially as an R package on GitHub and CRAN in May 2025.

The Time Series Processing System (TSPS) 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. Just like G-Series, the TSPS will be redeveloped in R. The work is planned to start next year and will provide the flexibility to incorporate tools and new techniques available from open-source software.

The Seasonal Adjustment Dashboard is now hosted on Statistics Canada's internal GitLab and reconfigured to make it easier to install. Other minor improvements to the dashboard include reducing dependencies and decreasing the load time. The dashboard was implemented for additional two labour programs this year, during which a minor revision was done to incorporate a program-specific calendar effect. The Centre also provided training to subject-matter analysts.

Time Series modelling and forecasting

The Centre completed a project on producing early indications of structural breaks using state-space models. An R tool was created to identify and model shocks (sudden changes) in time series, either in measurement equations (additive outliers) or state equations (more permanent effects).

The Labour Force Survey calibrates some of its estimates to population counts. The non-permanent resident portion of the population varies seasonally (in part due to international students and temporary workers). To stabilize calibration, a simple methodology is currently used to smooth out these fluctuations (12-month moving average). The Centre started a study to determine whether seasonal adjustment and trend-cycle estimation techniques could provide an improved methodology, especially in times of rapid changes in that segment of the population. The investigation is ongoing.

The COVID-19 pandemic period heavily affected many economic time series during the 2020-2022 period. With data now available for 2023 and 2024, the impact of the pandemic on seasonal patterns, the trend-cycle, and volatility can start to be assessed. The Centre has begun a preliminary investigation using representative time series from various statistical programs (labour, trade, manufacturing, building permits and investment in building construction, tourism, and consumer price index) to determine whether these series have returned to pre-pandemic patterns or have shifted to a new post-pandemic regime. The investigation will continue to be refined as more post-pandemic data becomes available.

The Centre also contributed to a nowcasting project as part of Statistics Canada's Methodological Acceleration Initiative, an initiative started to address the need to make a significant leap forward in the Agency's ability to efficiently and quickly implement modern solutions to data challenges. The Centre developed, tested, and validated one scenario of macro-estimate nowcasting for the Retail Commodity Survey, using ARIMA models and Neural Networks. Additionally, one scenario of micro-modeling, at the enterprise level, was developed, tested, and validated.

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.

This unit developed and tested a sampling protocol for quality control of machine learning classifiers used in large scale multi-class classification problems and longitudinal collection. For the last few years, the CPI has incorporated point-of-sale scanner data, giving a census of transactions from a few major retailers each month. Product text descriptions need to be classified into CPI's hierarchical aggregation structure of over 700 commodity classes. A linear support vector machine classifier has been trained to do this classification. However, it makes too many errors to be acceptable for the CPI, requiring additional review by human annotators. Cost of human annotation has been the limiting step in expanding the use of scanner data. Three sampling strategies were developed to select cases for human annotation and provide estimates of misclassification error rates with modified Wilson score confidence intervals. The Centre collaborated with data scientists from the Consumer Prices Division to implement these sampling strategies in Python and test them on simulated cases of food sales, with the goal of making code publicly available for other national statistical offices and researchers. Simulation results favoured a hybrid between two strategies: 1) a stratified simple random sample for new products each month to estimate misclassification rates and monitor model performance, and 2) a probability-proportional-to-size sample among unreviewed units from previous months, targeted at cases most likely to cause misclassification bias in the CPI estimate, based on an influence measure derived for price indices (Spackman, Francis and Goussev, 2025). The strategy and simulation results were presented at an April 2025 United Nations Economic Commission for Europe (UNECE) meeting on consumer price index methods in Geneva, Switzerland.

The Centre also consulted on the development of a set of quality indicators for the CPI. The first phase of this project produced a set of quantitative indicators and an aggregate measure for potential collection and processing issues. As well, a preliminary set of qualitative indicators covering Statistics Canada's six dimensions of quality were developed (Francis, Carrillo-Garcia, Yélou and Rassart, 2025).

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

References

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

Ferland, M. (2025). gseries: Improve the Coherence of Your Time Series Data. R package version 3.0.2, https://StatCan.github.io/gensol-gseries/en/.

Francis, J., Carrillo-Garcia, I., Yélou, C. and Rassart, E. (2025). Developing CPI Quality Indicators: Project Update. Internal document, Statistics Canada.

Spackman, W., Francis, J. and Goussev, S. (2025). Optimal use of machine learning at scale: Designing quality control of machine learning classification and mitigating misclassification error. *Proceedings of the 17th Meeting of the Group of Experts on Consumer Price Indices*, UNECE, Geneva, Switzerland. Available at: https://unece.org/sites/default/files/2025-04/Spackman%20et%20al%20%282025%29%20-%20QC%20for%20mitigating%20misclassification%20bias%20-%20paper.pdf.

Statistics Canada (2024). Training. Available at: https://www.statcan.gc.ca/eng/wtc/training.

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

Verret, F. (2021). Statistics Canada's seasonal adjustment dashboard. *Proceedings: Symposium 2021, Adopting Data Science in Official Statistics to Meet Society's Emerging Needs*, Statistics Canada, Ottawa, Canada.

5.2 Resource Centre for Economic Statistical Tools and Innovation

The Economic Generalized Systems unit has been reorganized as a section called the Resource Centre for Economic Statistical Tools and Innovation. This new Centre is responsible, among other functions, for the support and development of three generalized systems: G-Sam (generalized sampling system), Banff (generalized system for statistical data editing), and G-Est (generalized system for estimation).

Progress:

Projects in the Centre can generally be categorized as support (for users of the system), research and development.

The Centre handled a typical volume of support cases for G-Sam, Banff, and G-Est. These included support cases for both the current SAS-based versions of the systems, and support for migration to the Python version of Banff. While some cases are resolved by a few email exchanges, several others required more in-depth involvement and included recommendations on appropriate implementation of the system to achieve statistical goals. A summary of these cases is provided below.

The Banff team:

- Continued to collaborate with the Job Vacancy and Wage Survey to implement changes in their Edit and Imputation step. Additionally, they provided an open-source version of the process, calling the Python version of Banff, excluding custom SAS steps.
- Worked with the tax section on an imputation strategy for the T1 redesign project, using the Python version of Banff.
- Consulted with several surveys, on both the economic and household sides, exploring the use of the new Python version. This included demonstrations of the Banff Processor and recommendations on open-source methods that could be integrated within Banff.

The G-Sam team:

- Provided support to the Prepared Food and Beverage Sales Survey, Canada's Core Public Infrastructure (CCPI), and Canadian Survey on Interprovincial Trade on optimizing allocation under various constraints–sample size, precision, and coordination–with additional input on rotation and re-stratification.
- Presented the G-Sam allocation functionality to CCPI, highlighting its features and potential applications.
- Delivered a detailed explanation to the Singapore Department of Statistics on the Lavallée-Hidiroglou stratification algorithm, including its theoretical foundation and practical application in survey design.
- Resolved numerical issues with allocation for the Fruits and Vegetables Survey, ensuring effective implementation of sampling strategies.

Most of the research and development in the fiscal year was devoted to Statistics Canada's Shift to Open Source (SOS) initiative. The team took an active role in updating users, both inside and outside the branch,

on the progress for the system. This included presentations to the Generalized Systems Steering Committee, Branch Seminars, Field Planning Board Meetings, and an external article (Gray and Pierre, 2025). The Centre also met with representatives from the Federal Statistical Office of Germany (Destatis) to discuss transitioning Banff to open source and to initiate a collaborative project on developing a framework and tool for assessing imputation methods.

The Banff modernization project was successfully completed within the defined scope, budget, and timeline. Banff is a modular Statistical Data Editing (SDE) system for identifying and treating reporting errors and non-response. In January 2025, the Python-based version was released, enhancing flexibility, supporting advanced imputation methods, and facilitating international collaboration. Banff includes nine procedures performing various SDE functions including outlier detection, error localization, imputation and prorating, and methods for reviewing and imputing data constrained by linear relationships. The Banff Processor is a metadata-driven tool that executes data editing in production, calling built-in Banff procedures alongside custom modules in a user-specified process flow. Users are encouraged to share custom Banff-compatible modules in the Banff Plugin Repository, open to all users, to foster collaboration and reduce duplication. This Python version brings the system in line with the Generic Statistical Data Editing Model (UNECE, 2019). A keynote presentation discussing the new system was given at the United Nations Economic Commission for Europe (UNECE) Expert Meeting on Statistical Data Editing (Gray, 2024). To promote the system, the Banff team delivered a launch seminar (Gray and Seffal, 2025a) and shared lessons learned during the Modern Statistical Methods and Data Science Branch Learning week (Gray and Seffal, 2025b).

The R version of G-Sam has been rebuilt from scratch, retaining core methods from the SAS version while significantly improving performance and expanding functionality. New allocation and selection functions handle a broader range of problems and outperform their SAS counterparts with less code and fewer inputs. Dependencies are minimal, and input structures have been redesigned for better usability. To ease the transition, the team is developing migration tools and guidance. Outputs remain largely consistent, with only minor differences due to the change in optimization engines. A *beta* version is scheduled for release in June 2025, with the production release planned for September 2025.

The R version of G-Est will offer weighting and variance estimation functions for complex survey designs and include most of the same features as the SAS version. It will support bootstrap replicates, calibration, nonresponse adjustments, small area estimation, and variance due to imputation (via the System for Estimation of the VAriance due to Nonresponse and Imputation, SEVANI, module). Like G-Sam, the implementation has been built from scratch, optimized for R with minimal external dependencies and the potential to add methods in the future. Alpha versions of calibration and one-phase variance estimation are completed, and initial tests show significant performance improvements relative to the SAS version. For small area estimation, the transition team is reviewing an existing package for suitability, while development of the bootstrap replication functions and SEVANI modules are scheduled to begin this spring. A *beta* version is scheduled for release in December 2025, with the production release planned for December 2026.

For more information, please contact:
Fritz Pierre (fritz.pierre@statcan.gc.ca).

References

Gray, D. (2024). Building the new Banff: An open-source data editing system based on GSDEM concepts. UNECE Expert Meeting on Statistical Data Editing, 7-9 October 2024, Vienna.

Gray, D. and Pierre, F. (2025). De SAS vers les sources libres : conversion des systèmes généralisés de Statistique Canada. To be published in the next issue of *Convergence*, a journal of the Association des statisticiennes et statisticiens du Québec.

Gray, D. and Seffal, M. (2025a). The New Banff: A Modern Edit and Imputation Platform for Everyone. Internal presentation, Statistics Canada.

Gray, D. and Seffal, M. (2025b). From SAS to Open-Source: What we learned from the Banff project. Internal presentation, Modern Statistical Methods and Data Science Branch Learning Week, Statistics Canada.

United Nations Economic Commission for Europe (UNECE) (2019). *Generic Statistical Data Editing Model (GSDEM)*, 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. The RLRC also offered support to internal and external G-Link users who requested assistance, provided comments or submitted suggestions through requests to the G-Link_info mailbox.

During the year, most of the methodological work focused on maintenance 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.

Development work focused on an update of tools for the creation of synthetic linkage-ready data for testing and training purposes. The work was completed in R and Python as part of the move to open-source tools in the agency.

The RLRC has also worked on a variety of other probabilistic linkages 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 work for external clients on Census data linked to multiple longitudinal administrative files.

Members of the team offered formal courses with Statistics Canada's Training Centre, as well as prepared materials for a workshop at the 2025 Canadian Research Data Centres Network conference and updated reference and strategy documentation for users of probabilistic record linkage techniques.

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, 2024 and March 31, 2025, DARC responded to approximately 50 requests. The questions varied in complexity and included topics such as logistic regression with survey data, variance estimation with bootstrap weights, confidence intervals for small proportions, comparison of dependent subpopulations, chi-square tests, calculation of degrees of freedom for survey data and quantile regression. DARC also helped clients with the implementation of statistical methods in R, SAS, SUDAAN and STATA software. In addition, DARC reviewed analytical papers for scientific journals, conferences and for internal publication.

Provision of Training

DARC presented, in English, the internal course 0438A "Statistical Analysis of Survey Data – Module 1". This six-day course is a mix of theory and practice. Exercises and examples were presented using R, SUDAAN and SAS code.

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 with complex survey data of the Statistical Modelling Course at Statistics Canada, in English. DARC also gave the seminar for recruits on analysis of data from a complex survey.

DARC, in collaboration with the Health Analysis Division (HAD), developed and delivered a 12-hour course titled "Surviving the Transition to R for Survey Analysts." This hands-on training was provided to survey analysts within HAD and included practical examples and exercises focused on data wrangling, data analysis using the R survey package, and data visualization.

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 2019, 2020 and 2022 cycles 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). This year, DARC estimated the variance of the factor scores, enabling the calculation of coefficients of variation and confidence intervals for all domains presented in the dashboard.

For more information, please contact:
Pierre-Olivier Julien (pierre-olivier.julien@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.-S., Mach, L. and Houle, S. (2021). Measuring workplace psychosocial factors in the federal government. *Health Reports*, 32, 12.

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.

Anonymization

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.

Statistics Canada has contributed to a set of fact sheets that will be published by the Information and Privacy Commissioner of Ontario.

The Centre for Confidentiality and Access continues to offer governance, guidance and strategic advice in producing open datasets. This year, over 15 new Public Use Microdata Files were reviewed and made available on the Statistics Canada website.

External consultation

Statistics Canada has continued to offer its expertise to several groups internationally and within Canada. Internationally, Statistics Canada continued to participate in a Pilot study between Canada (StatCan), France (Institut national du cancer), and the US (National Cancer Institute) meant to study the ability of sharing cancer data between nations.

Domestically, we are continuing to work with Health Canada in providing non-disclosive tables of daily mortality counts for input into their Air Quality Health Index Project (AQHI). Rather than conventional rounding techniques, noise addition methods are being proposed that aim to give better utility for the same level of disclosure risk.

Statistics Canada has offered its expertise with disclosure control strategies with Correction Services Canada, analysts for the National Inuit Health Survey, Health Canada/Public Health Agency of Canada in their release of information from Canada Vigilance Adverse Reaction Online Database, and Bank of Canada with their Canadian Survey of Consumer Expectations.

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

5.6 Support and Research Activities in Artificial Intelligence

The Methodology Research and Development Program (MRDP) at Statistics Canada has supported many activities for the Centre of Artificial Intelligence, Research and Excellence (CAIRE) and the Artificial Intelligence and Methods Division (AIMD). The support from the MRDP has enabled many in depth research, prototype, community, centre of expertise and guidelines that can profit the agency.

Activities, Mandates and Products:

The funding enabled 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 of 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.

Another community that was funded by the MRDP deals with Privacy Enhancing Technologies (PET). The support enabled the launching of the community of practice which is still ongoing. In this era of privacy, the public servant that will require PET will be able to seek and find information on the latest techniques, reach out to the PET team at Statistics Canada, and eventually find out lectures and articles as the site evolves. The support of the MRDP also enabled the launching of the Responsible PET team, which will develop guidelines and create a review committee for any internal or external PET project (work is in progress).

Aside from the communities and centres, the funding enabled a continuation of an in-depth study that involves leveraging state-of-the-art Denoising Diffusion Probabilistic Models (DDPMs) to generate large volumes of synthetic tabular data that closely mimic real-world data distributions. The project integrates a debiasing process, inspired by Prediction-Powered Inference (PPI) and the generalized difference estimator. This proposed solution is now being tested for implementation, which could save time, efforts, and money when conducting sample surveys.

The second promising research project funded by the committee is an investigative study aimed at creating evaluation metrics for generative Artificial Intelligence (AI), i.e., a framework for Large Language Models. Unlike classical problems such as classification or regression with established metrics (such as the Root Mean Squared Error, RMSE, or F1 score), generative AI lacks standardized evaluation methods; the ones that are used in practice vary based on specific tasks and have reliability issues. A designed evaluation pipeline framework was created to improve usability and reproducibility while offering practitioner-friendly tools and artifacts that do not require deep theoretical knowledge.

Other research was enabled through the MRDP, which highlights its importance in supporting AI research, modernization of methodology and the government AI community at large.

For more information, please contact:
Marie-Eve Bedard (marie-eve.bedard@statcan.gc.ca) or
Nabila Ould-Brahim (nabila.ould-brahim@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 Public Works and Government Service Canada, Canada Energy Regulator, Public Services and Procurement Canada and others.

The group also contributed to various corporate consultation initiatives.

For more information, please contact:
Jeremy Solomon (jeremy.solomon@statcan.gc.ca).

5.8 Quality Assurance Resource Centre

The Quality Assurance Resource Centre (QARC) is committed to advancing research and development in statistical methods that enhance quality assurance and control processes. Our primary objective is to raise the standards of survey data collection and processing operations within the bureau. To achieve this objective, we explore a range of methodologies with a particular focus on improving the outgoing quality of data.

At the core of our efforts is 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 spans a wide array of quality assurance and control practices, addressing challenges related to efficiency and automation. The insights gained from this research not only benefit our internal operations but also have broad applicability across various stages of survey processes.

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 been dedicated to implementing an innovative methodology called "Quality Control by Score" to enhance the quality control (QC) of machine learning (ML) text coding processes. As ML technology becomes increasingly integral to data processing, maintaining the quality of generated codes is more important than ever. To meet this challenge, Statistics Canada has actively

pursued a strategy for determining optimal QC sampling rates, leveraging scores derived from the ML process. This methodology promotes a responsible approach to data classification while facilitating the broader adoption of machine learning. Our goal is to apply this approach to QC across various classifications within key surveys, including the Labour Force Survey (LFS), Job Vacancy and Wage Survey (JVWS), Canadian Community Health Survey (CCHS), and the Statistical Business Register (SBR). A paper detailing this methodology was presented at the Advisory Committee on Statistical Methods (Oyarzun, Wile and Evans, 2023).

As well, QARC assisted LFS in updating its data to align with the latest industry and occupational classifications, which involved significant structural changes. Traditionally, split-off classes were handled through manual recoding and random allocation. A new hybrid framework, combining machine learning (fastText) with linear programming, was developed to improve efficiency while maintaining consistency with traditional estimates. This work was presented at the Statistics Canada Symposium 2024 (Evans and Wile, 2024).

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

References

Evans, J. and Wile, L. (2024). Life in the FastText Lane: Harnessing linear programming constrained machine learning for classifications revision. *Proceedings: Symposium 2024, The Future of Official Statistics*, Statistics Canada, Ottawa, Canada.

Oyarzun, J., Wile, L. and Evans, J. (2023). Quality control by score. Paper presented to 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 work has recently been expanded to include reviews related to Machine Learning and Artificial Intelligence. 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, Quality, 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 over 150 ethical reviews over the past year alone, members of the Data Ethics Secretariat have given numerous presentations to inform internal partners, and colleagues from other federal departments and international organizations about Statistics Canada's approach to assessing data ethics. The team gathers information to remain up to date on topics which could be perceived as sensitive by the public. This is done by conducting literature reviews on some targeted topics and having informal discussions with internal partners, such as Communications and the Questionnaire Design Resource Centre, counterparts from other federal departments or National Statistical Offices around the world.

In addition to its internal activities, the team is also active internationally, playing a leadership role in the United Nations Economic Commission for Europe (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 was completed in 2025.

For more information, please contact:
Ryan Chepita (ryan.chepita@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, a presentation on quality management practices in relation to the Generic Statistical Business Process Model was made to the First Nations Information Governance Centre.

Discussions occurred within the Government of Canada Enterprise Data Community of Practice Data Quality Working Group. This working group, co-chaired by Statistics Canada, released an abridged data quality called Guidance on Data Quality in January 2024. Since its relaunch in autumn 2024, the group has aimed to identify gaps in quality governance by conducting an environmental scan on topics such as metadata, open data, and quality reporting.

At the international level, the Quality Secretariat met with the Abu Dhabi Statistics Centre to provide counsel on data quality dashboards, while continuing its involvement with the United Nations Expert Group on National Quality Assurance Frameworks. Statistics Canada served as the co-chair of the Subgroup on administrative and other data sources, whose purpose was to prepare a module for quality assurance when using administrative and other data sources to produce Official Statistics. This module, released in early 2025, aims to provide practical and concise guidance as well as best practices for statistical agencies in assuring the quality of official statistics when alternative data sources are used to produce official statistics; it is to be used as a complement to the United Nations National Quality Assurance Framework Manual for Official Statistics (United Nations, 2019).

PROJECT: Triennial Program Reviews

As part of Statistics Canada's Triennial Program Reviews, the Quality Secretariat designed a self-assessment questionnaire which was completed by the seven programs that were in scope for the 2024-2025 reviews. The questionnaire allowed each program to gauge their quality preparedness and current

best practices as they pertain to their culture and Statistics Canada's six dimensions of quality. The Quality Secretariat will continue to play a key role in these reviews as their scope expands in 2025-2026.

For more information, please contact:
Ryan Chepita (ryan.chepita@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/.

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 2024 and December 2024 issues (50-1 and 50-2) were released. The June 2024 issue is a special issue for papers presented at the 29th 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. Sixteen papers were published in the December 2024 issue, which featured the 2024 Waksberg invited paper by Richard Valliant entitled "Sample design using models".

In 2024, 62 papers were submitted to the journal. The average number of days from submission to initial decision was 43. All submitted papers were reviewed within 125 days, and 77% of them were reviewed within 90 days. Among those 62 papers, 37 were rejected, 17 were accepted, and 8 had not received a final decision (including papers that were not revised by the authors before the deadline) as of June 9, 2025. From April 2024 to March 2025, the *Survey Methodology* pages were viewed 46,472 times.

The June 2025 issue will be dedicated to the celebration of the 50th anniversary of *Survey Methodology*. It will feature two discussion papers. The first discussion paper, by Carl-Erik Särndal, is entitled "Progress in survey science: yesterday – today – tomorrow" whereas the second discussion paper, by J.N.K. Rao and Sharon Lohr, is entitled "Trends and directions in sample survey theory and methods". Both papers are followed by discussions from eminent survey statisticians and a rejoinder. With this special issue, we are pleased to announce that we are launching a new section of the journal for interview papers. These papers will typically contain conversations with survey statisticians or methodologists that had a significant impact in the survey sampling field. The special June 2025 issue includes two interview papers, one

interview with Ivan P. Fellegi and the other with Geoffrey Hole, to start this new section off on the right foot. It also features seven invited papers by renowned experts in survey statistics and methodology.

We are also currently planning two forthcoming special issues in 2026 or 2027. The first of these special issues will feature selected papers presented at the seventh International Conference on Establishment Statistics in Glasgow in June 2024. Paul Smith and Mojca Bavdaz are the guest editors for this special issue. The second special issue will be devoted to the theme "Shaping the future of survey statistics in the data-driven era". Maria Rosaria Ferrante and Natalie Shlomo are the guest editors for this special issue.

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

6.2 Knowledge Transfer – Statistical Training

The primary mandate of the Statistical Talent Development Working Group is statistical training within the Field and the organization. Several courses have been offered this year, including those related to time series, questionnaire design, sampling, record linkage, imputation, weighting, modeling, and statistical analysis with survey data, as well as an introductory workshop on simulations.

As for the new activities, the group continued to design and prioritize learning activities that can be developed in a timely manner and focused on active learning. This year, work to redesign the second statistical analysis course with survey data continued. We also developed an introductory course on simulations. This course was offered in both official languages. We also worked to update an old course on variance estimation.

Regarding the upcoming year, we will continue to offer the courses in the curriculum based on demand and the availability of teachers. Additionally, the development of the course on variance estimation will be completed and the course will be offered in both official languages.

The Talent Development Working Group offers various types of training opportunities so that employees can enjoy some flexibility in their professional development. In addition to the activities mentioned earlier, there are many self-training and self-learning opportunities, including the DataCamp and Onyxia platforms, 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 "The future of official statistics" took place October 30, 31 and November 1, 2024. The Symposium offered plenary sessions and parallel sessions that covered a variety of topics. Consistent with other statistical conferences worldwide, the 2024 Symposium featured speakers delivering their presentations in person. Observers benefited from the choice to attend in person or to join the sessions virtually.

Progress:

The 2024 Symposium included about 325 participants from the Modern Statistical Methods and Data Science Branch, and a further 150 participants who work in other branches within the agency or who represented other organizations. Nine countries were represented among the attendees.

Articles for the *Proceedings* were collected, edited and in the process of translation. The *Proceedings* are expected to be published in the summer of 2025.

Information on the 2024 Symposium is 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

Bosa, K., Beaumont, J.-F., Bocci, C. and Sombo, S. (2025). The use of a random forest algorithm in small area estimation. Paper presented to the Advisory Committee on Statistical Methods, June 2025, Statistics Canada.

Boulet, C. and Mather, A. (2025). Methodology Acceleration Initiative Research Project: Subsampling for Non-Response Follow-Up in Social Surveys. Internal document, Statistics Canada.

Dasylva, A., De Cubellis, M., De Fausti, F. and Franssen, L. (2025). Linking trade data from different national statistical offices through a private set intersection. *Journal of Official Statistics*, 41(2), 569-597, OnlineFirst, Special Issue, https://journals.sagepub.com/doi/10.1177/0282423X251329407.

Dasylva, A., Santos, B., Franssen, L., De Cubellis, M., De Fausti, F., Pappagallo, A., Berrios, N. and Fitzsimons, J. (2025). Private linkage of international trade microdata in a cloud-based secure enclave. To appear in the *Statistical Journal of the International Association for Official Statistics*.

Evans, J. and Wile, L. (2024). Life in the FastText lane: Harnessing linear programming constrained machine learning for classifications revision. *Proceedings: Symposium 2024, The Future of Official Statistics*, Statistics Canada, Ottawa, Canada.

Ferland, M. (2025). gseries: Improve the Coherence of Your Time Series Data. R package version 3.0.2, https://StatCan.github.io/gensol-gseries/en/.

Francis, J., Carrillo-Garcia, I., Yélou, C. and Rassart, E. (2025). Developing CPI Quality Indicators: Project Update. Internal document, Statistics Canada.

Gauvin, H. (2025). Practical Guidelines for the Creation of Synthetic Data Files. Internal document, Statistics Canada.

Gauvin, H. (2024). Creating a synthetic version of a longitudinal and structured file: Challenges and lessons. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, St. Johns, NF, https://ssc.ca/sites/default/files/imce/gauvin_ssc2024.pdf.

Gray, D. (2024). Building the new Banff: An open-source data editing system based on GSDEM concepts. UNECE Expert Meeting on Statistical Data Editing, 7-9 October 2024, Vienna.

Gray, D. and Pierre, F. (2025). De SAS vers les sources libres : conversion des systèmes généralisés de Statistique Canada. To be published in the next issue of *Convergence*, une revue de l'Association des statisticiennes et statisticiens du Québec.

Gray, D. and Seffal, M. (2025a). The New Banff: A Modern Edit and Imputation Platform for Everyone. Internal presentation, Statistics Canada.

Gray, D. and Seffal, M. (2025b). From SAS to Open-Source: What we learned from the Banff project. Internal presentation, Modern Statistical Methods and Data Science Branch Learning Week, Statistics Canada.

Hatko, S. and Sall, A. (2024). International Trade Unit Value Error Detection and Correction. Internal Report, Statistics Canada.

Istrate, A. and Robatian, D. (2025). Evaluation of AI-Generated Text: Survey, Taxonomy and Practical Framework for Designing NLG Evaluation Pipelines. Internal report, Statistics Canada.

Lafortune, Y. and Mayer, E. (2025). The Evolution of the Quantity and Quality of the Data Collected for the Survey of Household Spending Through the Years. Research Paper, Statistics Canada.

Loewen, R. and Jin, N. (2025a). Confidence intervals for Social Data Linkage Environment Error Rates. Internal document, Statistics Canada.

Loewen, R. and Jin, N. (2025b). Confidence intervals for Social Data Linkage Environment Error Rates. Slides for presentation to the Scientific Review Committee, May 23, 2025, Statistics Canada.

Loewen, R. and Jin, N. (2025c). Confidence intervals for Social Data Linkage Environment Error Rates. Slides for Divisional Seminar, June 12, 2025, Statistics Canada.

MacIsaac, K., Boulet, C. and Thomas, M. (2024). Recruitment and collection of Web panels at Statistics Canada. *Proceedings: Symposium 2024, The Future of Official Statistics*, Statistics Canada, Ottawa, Canada.

Mather, A. and Boulet, C. (2024). *Analysis of Residual Bias on Web Panels*. Methodology Branch Working Paper (SSMD-2024-002E), Statistics Canada.

Mather, A., Boulet, C. and Ra, Y. (2024). A bias evaluation for probabilistic Web panels at Statistics Canada. *Proceedings: Symposium 2024, The Future of Official Statistics*, Statistics Canada, Ottawa, Canada.

Miller, J. (2025). A Privacy-Utility Based Study of Differential Privacy. Internal report, Statistics Canada.

Shukla, A., Rossiter, E. and Santos B. (2024). Input/Output Privacy Framework: A Privacy-Preserving Record Linkage System. Internal Report, Statistics Canada.

Spackman, W., Francis, J. and Goussev, S. (2025). Optimal use of machine learning at scale: Designing quality control of machine learning classification and mitigating misclassification error. *Proceedings of the 17th Meeting of the Group of Experts on Consumer Price Indices*, UNECE, Geneva, Switzerland. Available at: https://unece.org/sites/default/files/2025-04/Spackman%20et%20al%20%282025%29%20-%20QC%20for%20mitigating%20misclassification%20bias%20-%20paper.pdf.

Tang, C. (2024). Statistical disclosure control analysis for small area estimation. *Proceedings: Symposium 2024, The Future of Official Statistics*, Statistics Canada, Ottawa, Canada.

Toupin, M.-H. and Martin, V. (2025). 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* (under revision).

Verret, F. and Walker, B. (2024). Reverse-engineering a hypothetical raking process for the estimation of mean squared error of raked small area estimates. *Proceedings: Symposium 2024, The Future of Official Statistics,* Statistics Canada, Ottawa, Canada.

Volkov, O. (2025). Bayesian nonparametric methods for survey data. *The Survey Statistician*, 91, 34-43, https://isi-iass.org/home/wp-content/uploads/Survey_Statistician_2025_January_N91.pdf.

You, Y. and Bosa, K. (2025). Performance of hierarchical Bayes small area estimators using noninformative and informative priors with an application to the Canadian Labor Force Survey. To appear in the December 2025 issue of *Survey Methodology*.

Yu, Z. (2024). Synthetic data disclosure risk assessment. *Proceedings: Symposium 2024, The Future of Official Statistics*, Statistics Canada, Ottawa, Canada.