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

Statistical Methodology Research and Development Program Achievements, 2022/2023

Release date: October 11, 2023

Statistics Canada Statistique Canada

Canada

How to obtain more information

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

You can also contact us by

Email at infostats@statcan.gc.ca

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

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

Standards of service to the public

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

Note of appreciation

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

Statistical Methodology Research and Development Program

Achievements, 2022/2023

This report summarizes the 2022/2023 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 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
(613-863-9024, jean-francois.beaumont@statcan.gc.ca)

Statistical Methodology Research and Development Program

Achievements, 2022/2023

Table of Contents

1 Data integration

1.1 Integration of probability and non-probability samples

PROJECT: Handling non-probability samples through inverse probability weighting

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

In this research, we focused on inverse probability weighting methods, which involve modelling the probability of participation in the non-probability sample. As a starting point, we considered the logistic model along with the pseudo maximum likelihood method of Chen, Li and Wu (2020). In previous years, we proposed a variable selection procedure based on a modified Akaike Information Criterion (AIC) that properly accounts for the data structure and the probability sampling design. We also proposed a simple rank-based method of forming homogeneous post-strata. In addition, we extended the Classification and Regression Trees (CART) algorithm to this data integration scenario, while again properly accounting for the probability sampling design. Our modified version of CART is called nppCART. Finally, we proposed a bootstrap variance estimator that reflects two sources of variability: the probability sampling design and the participation model.

We applied different inverse probability weighting methods to real probability and non-probability survey data collected by Statistics Canada. A main conclusion of our experiments is that inverse probability weighting methods are successful at bias reduction, but often some bias remains. We also observed the importance of forming homogeneous groups to stabilize estimates. The nppCART algorithm performed well with these data. Logistic regression with main effects only is also a reasonable option provided the estimated participation probabilities from the logistic model are used to form homogeneous groups.

Progress:

In the current year, we completed the writing of a paper that was submitted to *Survey Methodology*. After revision, taking into account the reviewers' comments, the paper was accepted for publication in the journal (Beaumont, Bosa, Brennan, Charlebois and Chu, 2023).

For more information, please contact:
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

References

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2023). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology* (accepted in 2023 and expected to appear in 2024).

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

PROJECT: An Approximate Bayesian Approach to Integrating Data from Probability and Non-Probability Samples

In this data integration project, we consider the scenario where the survey and auxiliary variables are observed in both a probability and non-probability sample. Our objective is to use data from the non-probability sample to improve the efficiency of survey-weighted estimates obtained from the probability sample. Recently, Sakshaug, Wiśniowski, Ruiz and Blom (2019) and Wiśniowski, Sakshaug, Ruiz and Blom (2020) proposed a Bayesian approach to integrating data from both samples for the estimation of model parameters. In their approach, the non-probability sample data are used to determine the prior distribution of model parameters, and the posterior distribution is obtained under the assumption that the probability sampling design is ignorable (or not informative). The goal of this project was to extend this Bayesian approach to the prediction of finite population parameters under a non-ignorable (or informative) probability sampling design.

In previous years, we proposed an approximate Bayesian procedure that accounts for the probability sampling design by conditioning on appropriate survey-weighted statistics, following Wang, Kim and Yang (2018), and conducted simulation experiments. The main conclusion of our experiments was that our Bayesian approach may yield efficiency gains over survey-weighted estimators, even in a situation where the non-probability sample is highly informative, provided the prior variance of model parameters is carefully chosen. However, it also led to efficiency losses in a scenario where the correlation between the survey and auxiliary variables was weak.

Progress:

This project was presented at the June 2022 meeting of Statistics Canada's Advisory Committee on Statistical Methods (ACSM) and at the 2022 Joint Statistical Meetings. Following ACSM advice, we conducted additional simulation experiments to strengthen our conclusions. We are also finalizing the writing of a paper that we plan to submit to a peer-reviewed statistical journal (You, DaSylva and Beaumont, 2023).

For more information, please contact:
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

References

Sakshaug, J.W., Wiśniowski, A., Ruiz, D.A.P. and Blom, A.G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics*, 35, 653-681.

Wang, Z., Kim, J.K. and Yang, S. (2018). Approximate Bayesian inference under informative sampling. *Biometrika*, 105, 91-102.

Wiśniowski, A., Sakshaug, J.W., Ruiz, D.A.P. and Blom, A.G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, 8, 120-147.

You, Y., DaSylva, A. and Beaumont, J.-F. (2023). An approximate Bayesian approach to estimation of population means by integrating data from probability and non-probability samples. Draft manuscript to be submitted to a peer-reviewed statistical journal.

PROJECT: Statistical data integration using a prediction approach

We investigated how a big non-probability database can be used to improve estimates from a small probability sample through data integration techniques. In the situation where the survey variable is observed in both data sources, Kim and Tam (2021) proposed two design-consistent estimators that can be justified through dual frame survey theory. In the previous year, we determined conditions ensuring that these estimators are more efficient than the Horvitz-Thompson estimator when the probability sample is selected using either Poisson sampling or simple random sampling without replacement. We also studied through a simulation study a class of predictors, following Särndal and Wright (1984), that handles the case where the non-probability database contains auxiliary variables but no survey variable. The probability survey collects the survey variables, and we assume that the non-probability database can be linked to the probability sample. This case is relevant to a survey on postal traffic conducted by La Poste in France. This project involves a collaboration with La Poste as well as the Toulouse School of Economics and the university of Besançon.

Progress:

The initial version of our paper was submitted to *Survey Methodology*. During the year, we revised our paper following reviewers' recommendations. The paper has recently been accepted in the journal (Medous, Goga, Ruiz-Gazen, Beaumont, Dessertaine and Puech, 2023).

For more information, please contact:
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

References

Kim, J.-K., and Tam, S.-M. (2021). Data integration by combining big data and survey sample data for finite population inference. *International Statistical Review*, 89, 382-401.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A. and Puech, P. (2023). QR prediction for statistical data integration. *Survey Methodology*, 49 (to appear).

Särndal, C.-E., and Wright, R. (1984). Cosmetic form of estimators in survey sampling. *Scandinavian Journal of Statistics*, 11, 146-156.

PROJECT: Disaggregation Plan for the Survey of Household Spending Food Data

The Survey of Household Spending (SHS) collects detailed information on household expenditures, using both a questionnaire and a one-week expenditure diary. SHS is conducted every other year since 2017 and data collection is continuous throughout the collection year to account for seasonal variations in spending. Food purchased from stores is collected via the SHS diary and it represents a fairly large portion of a household budget (11% in 2019). It is made up of about 250 different food categories, and since respondents only report their food purchases for a short 1-week period, it strongly limits the domains for which statistics can be produced (due to high CVs). Currently, the survey releases a biennial table of detailed food spending at the national and provincial level, and for other domains, it often limits the food information to the 8 highest-level food categories.

The objective of this research is to develop a disaggregation plan for the SHS Food data, with the aim that a similar approach could be used by other programs that use continuous collection throughout the year and having access to alternative data sources. Two alternative data sources now available at Statistics Canada were identified (retail scanner data, and non-probabilistic (volunteer) household panel spending data). The project aims to apply data integration methods to obtain adjusted estimates from these alternative data sources. Then, small area estimation methods would be used, with these adjusted estimates as auxiliary variables, to improve the precision of the released information with respect to three potential dimensions: time (monthly or quarterly instead of biennial), location (metropolitan areas and regions within provinces instead of national/provincial) and contents (more detailed food categories instead of only the 8 highest-level food categories).

Progress:

The progress made during 2022-2023 consists of the initial step of the evaluation of using non-probability inverse propensity weighting methods in order to improve the representativity of the food spending data from the non-probabilistic panel alternative data source. Modifications to inverse propensity weighting methods to account for non-probabilistic datasets, developed by Chen, Li and Wu (2020) and extended by Beaumont, Bosa, Brennan, Charlebois and Chu (2023) were adapted to the panel data. Auxiliary variables potentially able to explain frequent participation in the panel and that were available in both the panel and in a probability sample were identified. Logistic regression with variable selection methods was applied, with various model options including the allowance or not of pairwise interactions between auxiliary variables and the creation or not of homogenous classes based on the predicted propensities.

These methods did not succeed in notably reducing the bias in the 2019 panel data when compared to SHS 2019 overall and along the domains of province, income quintile, and household size. As the investigation of this panel data began it became clear there were significant limitations and that participation bias was not the main contributor to the bias observed; rather, widespread incomplete reporting and variations in reporting among different sub-types of food items (e.g., with a scannable bar code or without) were even more significant as far as suitability of this data source for the purposes envisaged. An internal report has been written (Charlebois, 2023) detailing the attempt to apply the data integration methods to the panel data source. The next steps for this project include investigating the second alternative data source, retail scanner data, with a view towards implementing small area estimates with the Fay-Herriot model using SHS data.

For more information, please contact:
Joanne Charlebois (613-875-5407, joanne.charlebois@statcan.gc.ca).

References

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2023). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology* (accepted in 2023 and expected to appear in 2024).

Charlebois, J. (2023). Nielsen Homescan Spending Data: Weighting by inverse propensity modelling of "frequent participation" in the Panel. Internal report, Social Survey Methods Division, Statistics Canada.

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

1.2 Record linkage

PROJECT: Identification of duplicate records

Duplicate records are records from the same unit in a data source, regardless of whether they are identical. Their identification is required when the source is used to produce official statistics, such as a sampling frame or a census. Fortini, Liseo, Nuccitelli and Scanu (2001), Tancredi and Liseo (2011), Sadinle (2017) and Steorts, Hall and Fienberg (2016) have described Bayesian models to perform this task in an automated manner, i.e., without clerical-reviews that are expensive. Yet, they involve computer-intensive procedures and tend to assume that the linkage variables are conditionally independent, when this is seldom the case in practice.

Progress:

A new model has been proposed for applications, where one can reasonably assume that each unit is associated with at most two records because duplication is rare, as in the private dwellings of the census of population. The duplication is modeled through the number of links from a given record as in a recent model of linkage errors (Dasylva and Goussanou, 2022a), while extending the latter to account for the multiplicity of false positives from some other units. The model has been presented at the Annual meeting of the Statistical Society of Canada, with a corresponding proceedings paper (Dasylva and Goussanou, 2022b), which includes Monte Carlo simulations based on public census data.

For more information, please contact:
Abel Dasylva (613-408-4850, abel.dasylva@statcan.gc.ca).

References

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

Dasylva, A., and Goussanou, A. (2022b). A new model for the automated identification of duplicate records. In *Proceedings of the Survey Methods Section*, Statistical Society of Canada. Available at https://ssc.ca/sites/default/files/imce/dasylva_ssc2022.pdf.

Fortini, M., Liseo, B., Nuccitelli, A. and Scanu, M. (2001). On Bayesian record linkage. *Research in Official Statistics*, 4, 185-198.

Sadinle, M. (2017). Bayesian estimation of bipartite matchings for record linkage. *Journal of the American Statistical Association*, 112, 600-612.

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

Tancredi, A., and Liseo, B. (2011). A hierarchical Bayesian approach to record linkage and population size problems. *Annals of Applied Statistics*, 5, 1553-1585.

PROJECT: Capture-recapture with linkage errors

To reduce the response burden and costs, Statistics Canada is prioritizing the use of administrative sources for the production of official statistics (Rancourt, 2018). To do so, the Agency has developed a set of related quality indicators, including the under-coverage and over-coverage of a given source (Sirois, 2021), which are currently measured through costly field operations or clerical reviews (Oyarzun and Rodrigue, 2023). The capture-recapture method may offer a cost-effective alternative for measuring the under-coverage through a comparison to another source under standard assumptions, which include the independence of the sources and their perfect linkage. However, it must be modified when the linkage is imperfect, typically because there is no unique identifier that is common to the two sources. Many such adaptations have been proposed under the standard capture-recapture assumptions except for the imperfect linkage. Ding and Fienberg (1994), Di Consiglio and Tuoto (2015) and de Wolf, van der Laan and Zult (2019) have described solutions that still require clerical reviews. Racinskij, Smith and van der Heijden (2019) have instead proposed a solution that dispenses with clerical reviews, at the expense of making the strong assumption that the linkage variables are conditionally independent. Dasylva, Goussanou and Nambeu (2021) have proposed a solution, which builds on the error model described by Dasylva and Goussanou (2022), while dispensing with clerical reviews and allowing more general forms of dependence among the linkage variables. However, when the linkage does not have a perfect recall, the coverage must be estimated by fitting this model repeatedly based on a log-linear specification of the interactions in the matched pairs, which is inefficient.

Progress:

The model described by Dasylva and Goussanou (2022) has been generalized into a multivariate model. Whereas the original model is based on the number of links from a given record for a single linkage rule, the multivariate extension is based on a vector of such variables, each corresponding to a distinct linkage rule from a set of mutually exclusive rules, i.e., each record pair is linked by at most one rule. This extension is used to estimate the coverage based on a log-linear specification of the interactions in the matched pairs, and it leads to a more efficient estimator of the coverage than that obtained by fitting the univariate model repeatedly, when the recall is not perfect. The new model is described in a report (Dasylva, Goussanou and Nambeu, 2023) and it has been submitted to a peer-reviewed journal.

For more information, please contact:
Abel Dasylva (613-408-4850, abel.dasylva@statcan.gc.ca).

References

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

Dasylva, A., Goussanou, A. and Nambeu, C.-O. (2021). Measuring the undercoverage of two data sources with a nearly perfect coverage through capture and recapture in the presence of linkage errors. Proceedings: *Symposium 2021, Adopting Data Science in Official Statistics to Meet Society's Emerging Needs*, Statistics Canada. Available at https://www150.statcan.gc.ca/n1/pub/11-522-x/2021001/article/00006-eng.pdf.

Dasylva, A., Goussanou, A. and Nambeu, C.-O. (2023). Measuring the coverage of two data sources through capture-recapture with linkage errors. Internal report, Statistics Canada.

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

Di Consiglio, L., and Tuoto, T. (2015). Coverage evaluation on probabilistically linked data. *Journal of Official Statistics*, 31, 415-429.

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

Oyarzun, J., and Rodrigue, J.-F. (2023). Building key quality indicators for the Statistical Building Register. Presentation to the Advisory Committee on Statistical Methods, Statistics Canada, May 2023.

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

Rancourt, E. (2018). Admin-first as a statistical paradigm for Canadian official statistics: Meaning, challenges and opportunities. Proceedings: *Symposium 2018*, *Combine to Conquer: Innovations in the Use of Multiple Sources of Data*, Statistics Canada. Available at https://www.statcan.gc.ca/eng/conferences/symposium2018/program/03a2_rancourt-eng.pdf.

Sirois, M. (2021). Coverage quality indicators. Internal presentation, Statistics Canada, June 2021.

PROJECT: Probabilistic record linkage based on recursive partitioning without training data

The probabilistic method of record linkage (Fellegi and Sunter, 1969) is often preferred when linking records without a unique identifier (Enamorado, FiField and Imai, 2019; Bianchi Santiago, Colon Jordan and Valdes, 2020). For example, at Statistics Canada, it is intensively used in the Social Data Linkage Environment (Statistics Canada, 2022), which provides the linked data for many social, health and environmental studies. The probabilistic method is appealing because it provides a principled way of minimizing the linkage errors for a given set of features. However, it falls short of prescribing these features, which are still selected manually from experience; a labour-intensive process that does not guarantee the optimality of the chosen features. Another challenge is the estimation of the decision parameters because of the features interactions. Finally, the probabilistic method is often perceived as complex and non-intuitive by those familiar with the deterministic or hierarchical methods of record linkage. Recursive partitioning is ideally placed to address these limitations of the probabilistic method, but its application to record linkage has been hampered by the common lack of training data, with earlier attempts at using decision trees for record linkage by Elfeky, Verykios, Elmagarmid, Ghanem and Huwait (2003), and Feigenbaum (2016). Elfeky et al. (2003) train the tree on the result of an unsupervised two-means clustering procedure, which has its own limitations (Quadir and Bao, 2016; Quadir, 2017), while Feingenbaum trains the tree on a clerical review sample, which may be costly to source.

Progress:

A new record linkage methodology has been developed, which blends a particular form of recursive partitioning with the probabilistic method, while dispensing with training data or the ground truth (Chen, 2022; Dasylva and Chen, 2022). In this methodology, recursive partitioning is used to select the features and partition the record pairs into groups where the conditional match probability (i.e., the conditional probability that two records represent the same unit given the observed features) is as homogeneous as possible. The actual procedure is a variation of that proposed by Breiman, Friedman, Olshen and Stone (1984), where the tree cost function is estimated with the model described by Dasylva and Goussanou (2022), while implicitly accounting for all interactions among the selected features, at each node. The resulting tree is then used to establish optimal links, by assigning a weight to each leaf and linking the pairs with a positive probability where this weight is no less than a threshold. The methodology has been implemented in R with the Rpart package (Therneau, Atkinson, and Ripley, 2009) and a custom splitting function.

For more information, please contact:
Abel Dasylva (613-408-4850, abel.dasylva@statcan.gc.ca).

References

Bianchi Santiago, J., Colon Jordan, H. and Valdes, D. (2020). Record linkage of crashes with injuries and medical cost in Puerto Rico. *Transportation Research Record*, 2674(10), 739-748.

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984). *Classification and Regression Trees*. Wadsworth.

Chen, W. (2022). Optimal feature extraction for probabilistic record linkage with model-based trees. Internal report, Statistics Canada.

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

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

Elfeky, M., Verykios, V., Elmagarmid, A., Ghanem, T. and Huwait, A. (2003). Record linkage: A machine learning approach, a toolbox, and a digital government web service. Available at http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=504FD1F7CC64E71A806D75F14453621E?doi=10.1.1.11.1113&rep=rep1&type=pdf.

Enamorado, T., Fifield, B. and Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *American Political Science Review*, 113, 353-371.

Feigenbaum, J.J. (2016). A machine learning approach to Census record linking. Working paper, available at https://scholar.harvard.edu/files/jfeigenbaum/files/feigenbaum-censuslink.pdf.

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

Quadir, T., and Bao, C. (2016). Application of Machine Learning Algorithms in G-Link. Internal Report, Statistics Canada.

Quadir, T. (2017). Automated/semi-automated Estimation of Thresholds for Weights in G-Link. Internal Report, Statistics Canada.

Statistics Canada (2022). Social Data Linkage Environment. https://www.statcan.gc.ca/en/sdle/overview.

Therneau, T., Atkinson, B. and Ripley, B. (2009). Rpart: recursive partitioning and regression trees [Computer software manual]. Available at http://CRAN.R-project.org/package=rpart.

PROJECT: Private set intersection with linkage errors

Many important questions about international trade are not addressed currently because national statistical organizations cannot easily link their corresponding micro-data due to privacy concerns and regulatory requirements. For example, one could study the use of preferential tariffs by different types of Dutch exporters in the context of the Comprehensive Economic Trade Agreement (CETA) between Canada and the EU, if one could link Dutch exports to Canadian imports. Privacy enhancing techonologies such as secure multi-party computation can enable such a linkage, while mitigating the privacy risks (United Nations, 2023). Thus, they may be the difference between doing and not doing a study, under the appropriate governance and regulatory framework. In particular, the linkage may be based on a previously proposed three-party private set intersection protocol (Bruno, De Cubellis, De Fausti, Scannapieco, and Vaccari, 2021), and the target totals may be computed without any bias from the linked data set, if there is a unique identifier. However, without such an identifier, linkage errors may arise and bias the computed totals.

Progress:

A statistical methodology has been developed to assess the linkage errors and adjust the estimated totals accordingly, when the private set intersection is based on quasi-identifiers. This methodology estimates the rates of linkage error according to Dasylva and Goussanou (2022), and it adjusts the totals using a variation of the weight adjustment scheme proposed by Judson, Parker, and Larsen (2013), where the false positives are also accounted for. The methodology is described in the final report of the UNECE project on Input Privacy Preservation (UNECE, 2023), along with simulations that demonstrate the effectiveness of the proposed approach.

For more information, please contact:
Abel Dasylva (613-408-4850, abel.dasylva@statcan.gc.ca).

References

Bruno, M., De Cubellis, M., De Fausti, F., Scannapieco, M. and Vaccari, C. (2021). Privacy set intersection with analytics - An experimental protocol (PSI De Cristofaro). UNECE-IPP presentation.

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

Judson, D.H., Parker, J. and Larsen, M.D. (2013). Adjusting sample weights for linkage-eligibility using SUDAAN. National Center for Health Statistics. Available at https://www.cdc.gov/nchs/data/datalinkage/adjusting_sample_weights_for_linkage_eligibility_using_sudaan.pdf.

United Nations (2023). *United Nations Guide on Privacy-Enhancing Technologies for Official Statistics*. United Nations Committee of Experts on Big Data and Data Science for Official Statistics, New York. Available at https://unstats.un.org/bigdata.

UNECE (2023). UNECE project on input privacy preservation: Final report. United Nations Economic Commission for Europe. Available at https://statswiki.unece.org/display/hlgbas/Input+Privacy-Preservation+for+Official+Statistics+Project+outcome.

PROJECT: Variance estimation for record linkage error-rates obtained via clerical review of stratified systematic samples of linked pairs

A common method of estimating record linkage error-rates is to use a manual, clerical review process. A sample of confirmed and rejected pairs from the linkage is sent to independent clerical reviewers. The reviewers then make decisions about each pair in the sample and their decisions are compared to the outcomes from the linkage process to estimate false match and missed-match rates. The Social Data Linkage Environment (SDLE) Methodology Unit at Statistics Canada currently makes use of a variant of this scheme in which the sample is drawn using a stratified systematic sampling design. This is also the method currently programmed into the clerical review tool of Statistics Canada's Generalized System for Record Linkage (G-Link). Our unit does not currently provide estimates of design variance for our error-rate estimates. The goal of our project was to find a method of producing such estimates. This problem is interesting since there exists no unbiased estimator for the design variance of the Horvitz-Thompson estimator under systematic sampling.

Progress:

In order to find a suitable estimator, we employed methodology outlined by Kirk Wolter (1984) in a paper about variance estimation under systematic sampling. We considered a list of potential estimators, meant to be broadly representative of the variance estimators for systematic sampling available in the literature. We conducted a simulation study to evaluate the performance of these estimators. Our population of interest was the set of all pairs in a stratum from a typical SDLE linkage, together with the decisions clerical reviewers would make about those pairs. We created artificial versions of SDLE linkage strata by sampling from actual SDLE linkages, and we simulated clerical review decisions for these pairs using two different methods. One method was based on the Fellegi-Sunter record linkage model (Fellegi and Sunter, 1969), and the other was based on a decision tree trained using clerical review data (see Chen (2022) for a discussion of the use of decision trees as linkage models). Our estimators were evaluated in terms of bias, mean square error, and confidence probability. Through this comparison, we were able to identify one estimator that seems to perform better than the others for our purposes: the "overlapping strata" estimator discussed in (Yates (1981) page 231). In addition to our main line of investigation, we also computed the intra-class correlation coefficients for our artificial populations to compare the efficiency of systematic sampling and simple random sampling in this context, and we used our variance estimates

to derive optimal sample allocations for our clerical review samples. Finally, we have incorporated variance estimates into one of the standard programs used in the SDLE production process. A detailed account of our work can be found in Loewen and Millar (2023).

For more information, please contact:
Goldwyn Millar (343-553-3930, goldwyn.millar@statcan.gc.ca).

References

Chen, W. (2022). Optimal feature extraction for probabilistic record linkage with model-based trees. Statistics Canada Coop student report, research supervised by Abel Dasylva, Statistics Canada, Ottawa.

Fellegi, I.P., and Sunter, A.B. (1969). A theory for record linkage. *Journal of the American Statistical Association*, Vol. 64, No. 328, 1183-1210.

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

Wolter, K. (1984). An investigation of some estimators of variance for systematic sampling. *Journal of the American Statistical Association*, Vol. 79, No. 388, 781-790.

Yates, F. (1981). *Sampling Methods for Censuses and Surveys,* 4th edition, New York, Macmillan Publishing Co.

1.3 Small area estimation

PROJECT: The use of random forests in small area estimation

When domain sample sizes are small, design-consistent direct estimators of population parameters can 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 could 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. We are particularly interested in random forests for two reasons: i) they can be easily applied to the case of a mixture of categorical and continuous auxiliary variables, and ii) 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:

We have developed a few non-parametric versions of the Empirical Best (EB) predictor when random forests are used to replace parametric models. We have evaluated the properties of our proposed EB predictors using real data and through simulation studies. Our results show that random forests are promising but further studies are needed before making firm conclusions. This project was presented at the 2023 Colloque francophone sur les sondages in Paris and at the 2023 annual conference of the Statistical Society of Canada. In the next year, we plan to complete our empirical studies and write a paper to be submitted to a peer-reviewed statistical journal.

For more information, please contact:
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

PROJECT: Sample allocation under the Fay-Herriot small area model

Small area estimation procedures gained popularity in the last two decades because of the growing appetite for more granular estimates. Surveys are usually not designed to meet all users' needs, which is why models such as the well-known Fay-Herriot model are used to compensate for small sample sizes in some domains. Moreover, a theoretical framework exists for many small area procedures (see Rao and Molina, 2015), which contributes to democratize this methodology.

The main objective of this project is to determine an effective method of allocating a sample to domains such that the resulting small area estimates, under the Fay-Herriot model, have sufficient precision for the largest number of domains possible. In other words, we focus on the optimization of the sample allocation when the Fay-Herriot model is used at the estimation stage. We consider the case where domains coincide with strata, as in Longford (2006). We also consider the theoretical scenario where the Fay-Herriot model parameters are known. This allows us to compute the best predictor and the variance of its prediction error, which is known as the g1 variance term in the literature. An important difference with standard design-based sample allocation is the substitution of the sampling variance of the direct estimator with the g1 variance term. This allows us to optimize, at the allocation stage, the precision of final domain estimates (or small area estimates).

Progress:

We proposed a new simple allocation method, which aims to reach target g1 variances for the most important domains. This method was tested using data from the Canadian Labour Force Survey and compared with the method proposed by Longford (2006) as well as more traditional sample allocation methods. The main conclusion is that sample allocation has more impact on direct survey estimates than small area estimates. Another conclusion is that our proposed method allows us by design to reach the desired precision for more domains than the alternative methods considered.

This project was presented at the 2023 Statistical Society of Canada Annual Meeting in Ottawa (Bosa and Beaumont, 2023). We are planning to write a paper summarizing our findings that will be submitted to the proceedings of the conference and/or to a peer-reviewed statistical journal.

For more information, please contact:
Keven Bosa (613-863-8964, keven.bosa@statcan.gc.ca).

References

Bosa, K., and Beaumont, J.-F. (2023). How to allocate the sample to maximize benefits from small area estimation techniques? Presentation at the Statistical Society of Canada Annual Meeting, May 2023.

Longford, N.T. (2006). Sample size calculation for small-area estimation. *Survey Methodology*, 32, 1, 87-96. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006001/article/9259-eng.pdf.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation* (2nd edition). NJ: John Wiley & Sons, Inc.

PROJECT: Relative efficiency of area-level and unit-level small-area estimators when unit level auxiliary data is available

Small-area estimation is widely used in many statistical agencies to produce reliable statistics using a model-based approach using either area-level or unit-level models. This theory is well documented in many leading sources such as Rao and Molina (2015). In this research, we compare the two different approaches to modeling under the situation where we have complete auxiliary information for all the units in the population. This follows up on the work by Hidiroglou and You (2016) and Fay (2018).

We examine the unit-level empirical best linear unbiased predictor (EBLUP), and we compare it to two area-level model estimators under different scenarios for generating the population of interest. The first area-level model estimator uses direct estimates using the simple design-weighted estimator, and the second uses direct estimates from the survey regression estimator which is constructed from the auxiliary information.

Progress:

We conducted various simulations by generating populations under different linear models of the auxiliary variables. The simulations tend to support the results noted by Hidiroglou and You (2016) and Fay (2018). When the auxiliary variables in the unit-level model provide a sufficiently good linear approximation to the survey variables, then the unit-level model produces very efficient small area estimates. We can also produce efficient estimates using an area-level model by first producing direct estimates using the survey regression estimator with the same auxiliary variables. These direct estimates when used as inputs to the area-level model produce an area-level estimator with very similar properties as the EBLUP estimator from the unit-level model. There appears to be no advantage in using one estimator over the other. We also found that both estimators are more efficient than the area-level estimator produced by using the direct estimates from the simple expansion or design-weighted estimator. We also developed theory that supports these empirical findings.

These results are the basis of a presentation by J.N.K. Rao (Rao, Estevao, Beaumont and Bosa, 2023) at the 2023 Annual Meeting of the Statistical Society of Canada at Carleton University in Ottawa.

For more information, please contact:
Victor Estevao (613-863-9038, victor.estevao@statcan.gc.ca).

References

Fay, R.E. (2018). Further comparisons of unit- and area-level small area estimators. In *Proceedings of the Survey Research Methods Section*, 2018 Joint Statistical Meetings, Vancouver, Canada.

Hidiroglou, M.A., and You, Y. (2016). Comparison of unit level and area level small area estimators. *Survey Methodology*, 42, 1, 41-61. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2016001/article/14540-eng.pdf.

Rao, J.N.K., Estevao, V., Beaumont, J.-F. and Bosa, K. (2023). Relative efficiency of area-level and unit-level small-area estimators when unit level auxiliary data is available. Presentation at the Statistical Society of Canada Annual Meeting, May 2023, Ottawa, Canada.

Rao, J.N.K., and Molina, I. (2015). *Small Area Estimation*. New York: John Wiley & Sons, Inc.

PROJECT: Sampling variance smoothing methods for small area proportion estimation

In this project, we consider sampling variance smoothing methods including the use of design effects and the generalized variance function (GVF) for small area estimation. In particular, we propose different methods including the average smoothed estimator and weighted average estimator for sampling variance smoothing. The proposed smoothing methods can be used in the small area estimation as a standard approach and simplify the smoothing procedure for model-based small area estimation.

Progress:

The proposed smoothing methods perform very well for small area proportion estimation. We presented a paper at the Symposium 2022 conference, a conference paper has been written and submitted (You and Hidiroglou, 2022). We completed Labor Force Survey (LFS) application and more simulation study. A modified version of the paper has been submitted to Journal of Official Statistics (JOS) and the paper has been accepted by JOS for publication (You and Hidiroglou, 2023).

For more information, please contact:
Yong You (613-863-9263, yong.you@statcan.gc.ca).

References

You, Y., and Hidiroglou, M. (2022). Application of sampling variance smoothing methods for small area proportion estimation. Proceedings: *Symposium 2022*, *Data Disaggregation: Building a more-representative data portrait of society*, Statistics Canada, Ottawa, Canada (to appear).

You, Y., and Hidiroglou, M. (2023). Application of sampling variance smoothing methods for small area proportion estimation. *Journal of Official Statistics*, to appear in the December issue 2023.

PROJECT: HB inference for small area estimation using different priors for variance components

Hierarchical Bayes (HB) modeling is very popular in small area estimation and prior specification is very important in this approach. In this project, we study the impact of priors on variance components for small area estimation based on the HB models of You and Chapman (2006) and You (2021). Particularly we investigate the flat prior and inverse gamma priors for the variance components through simulation study and real data analysis.

Progress:

We have studied prior specifications for variance components in the HB models of You and Chapman (2006) and You (2021). We conducted a simulation study and applied the models to LFS data application. Our results indicate that the use of inverse gamma prior for variance component can be very effective in

the HB models. A research paper has been submitted and will be published by Statistics in Transition new series (You, 2023).

For more information, please contact:
Yong You (613-863-9263, yong.you@statcan.gc.ca).

References

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

You, Y. (2023). An empirical study of hierarchical Bayes small area estimators using different priors for model variances. *Statistics in Transition* new series, to appear.

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: Estimation of the poverty measures for small areas under a two-fold nested error linear regression model: comparison of two methods

Demand for reliable statistics at a local area (small area) level has greatly increased in recent years. Traditional area-specific estimators based on probability samples are not adequate because of small sample size or even zero sample size in a local area. As a result, methods based on models linking the areas are widely used. The World Bank has focused on estimating poverty measures, in particular poverty incidence and poverty gap called FGT measures (named for Foster, Greer and Thorbecke (1984)), using a simulated census method, called ELL (named for Elbers, Lanjouw and Lanjouw (2001)), based on a one-fold nested error model for a suitable transformation of the welfare variable. Modified ELL methods leading to significant gain in efficiency over ELL also have been proposed under the one-fold model. An advantage of ELL and modified ELL methods is that distributional assumptions on the random effects in the model are not needed.

In this research, we focused on two-fold random effect models involving area and subarea random effects. We extended ELL and modified ELL to two-fold nested error models to estimate poverty indicators for areas and subareas.

Progress:

We developed extensions of the ELL method and proposed two modifications methods to estimate FGT poverty measures in small areas under two-fold nested error linear regression model for unit-level data that includes area and subarea effects. Our simulation results indicated that the modified ELL estimators lead to large efficiency gains over ELL at the area level and subarea level. Further, a modified ELL method retaining both area and subarea estimated effects in the model (called MELL2) performs significantly better in terms of mean squared error (MSE) for sampled subareas than the modified ELL retaining only estimated area effects in the model (called MELL1). We have written a paper detailing the results of our research and submitted it to a peer-reviewed statistical journal (Sohrabi and Rao, 2023).

For more information, please contact:
Maryam Sohrabi (343-553-4529, maryam.sohrabi@statcan.gc.ca).

References

Elbers, C., Lanjouw, J.O. and Lanjouw, P. (2001). *Welfare in Villages and Towns: Micro-Level Estimation of Poverty and Inequality*. Unpublished manuscript, The World Bank.

Foster, J., Greer, J. and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica: Journal of the Econometric Society,* 52(3), 761-766.

Sohrabi, M., and Rao, J.N.K. (2023). Estimation of the poverty measures for small areas under a two-fold nested error linear regression model: Comparison of two methods. Draft manuscript submitted to a peer-reviewed statistical journal. arXiv:2306.04907 [stat].

PROJECT: Guiding Principles: Using the 2021 Census of Population Data to Produce Statistics on DDAP Groups of Interest

The gain in momentum of movements for Indigenous rights, racial justice, and economic equality have changed the data that must be collected. To help address the change in data needs, Statistics Canada implemented the Disaggregated Data Action Plan (DDAP) to better understand and highlight the challenges faced by population groups such as women, Indigenous peoples, racialized populations, and those living with daily limitations. As a result, more data centred around these diverse population groups at more levels of geography will become available for public use, with the goal to promote fairness and inclusion in decision making. However, as surveys designs change to account for the new requirements, questions related to the data sources to be used, ethical implications, and the appropriateness of various sampling methods have arisen. In an effort to address this, the Census Operations DDAP Research Project was funded to help document sampling methods to consider in DDAP contexts, as well as ethical and practical considerations The research project also focuses on the role of the 2021 Canadian Census of Population in identifying some of the targeted DDAP subgroups and provides a link between the theory and practice while accounting for factors such as the respondent's burden and privacy.

Progress:

In order to fulfill the previously mentioned goals, the first outcome of the research project was a suite of tables of population counts for various DDAP subgroups of interest, based on the 2021 Census data, along with distributions of these populations when cross-tabulated by socio-demographic variables such as age, gender, and province (intersectionality variables). These population tables are used to gain clarity on the sampling possibilities of various subpopulations, and to guide survey designs by allowing appropriate sampling methods to be chosen based on the subpopulation's size.

The second outcome of the research project was the document "*Guiding Principles: Using the 2021 Census of Populations Data to Produce Statistics on DDAP Groups of Interest*" (Pearce, Sallier and Laperrière, 2023) consisting of three chapters. The first chapter presents the organizational context of DDAP at Statistics Canada; the second chapter covers the various existing data sources for DDAP initiatives as well as ethical considerations. Lastly, the third chapter is the result of a literature review which aimed at listing sampling methods that can be used for DDAP initiatives. Indeed, to accompany population size tables in determining an appropriate sampling method, population characteristics like socially connectedness, frequenting known locations, and hiddenness also need to be considered. Therefore, chapter three considers these various factors, in theory as well as in practice, and lists pros and cons of the methods presented. This chapter also provides concrete examples of applications of these methods at Statistics

Canada, as well as a section on practical considerations. This document of guiding principles and recommendations is currently being revised to be released internally and aims to centralize the available information to promote consistency and comparability within Statistics Canada, while also providing a set of coherent guiding principles moving forward for decision makers working at all levels.

For more information, please contact:
Kenza Sallier (343-998-8623, kenza.sallier@statcan.gc.ca).

Reference

Pearce, A., Sallier, K. and Laperrière, C. (2023). Guiding Principles: Using the 2021 Census of Populations Data to Produce Statistics on DDAP Groups of Interest. Internal Document, May 2023, Statistics Canada.

2 Data science methods and applications

PROJECT: Anonymization of training text data and its effect on the performance of NLP models

There is an increasing demand inside Statistics Canada and other agencies for Natural Language Processing (NLP) and text classification projects, such as the projects that were conducted on the Census comments and Immigration, Refugees and Citizenship Canada enquires. Those projects used text provided freely by the public to classify into operationally useful categories by using the latest Transformers NLP models. This kind of data may contain information that is protected. NLP models have been shown to memorize private information and base their decisions on it. Therefore, any bias that could arise from personal data in the text should be prevented as much as possible with adequate anonymization techniques to ensure that privacy is respected and that the model isn't biased against certain categories of people.

There are currently various proposed anonymization techniques to solve this problem, but as far as we know there are very few impact studies on how anonymization techniques influence down the line the performance of NLP models, and what strategy is optimal to ensure both anonymized data and good predictive performance. Once the project is completed, the teams working on NLP projects will have access to the pros and cons of different anonymization techniques, a recommended approach and a custom Python package to be used to anonymize the text before entering their NLP pipeline.

Progress:

On comparing the effects of different de-identification techniques like replacement of Personal Identifiable Information (PII) with entity tag, replacement of PII with synthetic data and replacement of PII with static string on the performance of a multi-label text classification task, it was observed that all techniques performed closely to each other and hence we cannot recommend a single technique that works best overall. It was demonstrated over a series of experiments that the F1 performance (the harmonic mean of precision and recall) of the text classification model is only slightly decreased (between 0.08% to 0.34%) when training with de-identified data instead of the original data containing all PII. Therefore, it shows that anonymizing training data is a worthwhile preprocessing step to protect data privacy without significantly affecting the performance of the NLP model. A description of the above experiments is provided in Istrate and Mashhadi (2023).

In addition to the experiments explained above, a python package was developed extending and improving on the existing open-source anonymization tool – Microsoft Presidio. The new package called Canonym uses transformers for named-entity recognition. It supports both English and French text. Canonym can accurately identify and mask various Canadian-specific PII like the Social Insurance Number, the Canadian Passport number, Canadian addresses, Canada postal codes, etc. in free-form text. It is easy to use and allows for various customizations to be made. This package has already benefitted another project in the Data Science Division.

For more information, please contact:
Alexandre Istrate (alexandre.istrate@statcan.gc.ca) or
Sayema Mashhadi (sayema.mashhadi@statcan.gc.ca).

Reference

Istrate, A., and Mashhadi, S. (2023). Anonymization of Training Text Data and its Effect on the Performance of NLP Models. Internal report, Statistics Canada, Ottawa.

PROJECT: On-device (Lightweight) Machine Learning models for Mobile/Web Applications

All state-of-the-art Artificial Intelligence are deep learning models with high performance capabilities and high resource requirements due to their size. The bigger the model, the more resources it requires for hosting and inference, making it difficult to deploy since not all projects can support or justify such high resource requirements. Larger models also have longer inference times and more energy consumption during inference. Due to these challenges, model compression techniques have become increasingly popular to reduce the size of the model while maintaining model accuracy and improving inference speed. These smaller models can run on a Central Processing Unit, a mobile device, within a web browser or on edge devices instead of requiring graphical processing units (GPUs).

Many inference acceleration and model compression techniques have been developed over the years. Some of the techniques are as follows:

- **Quantization** refers to reducing the precision of numerical values to optimize storage and computational efficiency without significant loss of model accuracy. Specific quantization techniques are:
  - Post-training Quantization, Quantization Aware Training, Dynamic Quantization.
- **Pruning** involves removing unnecessary connections or parameters from the model to reduce its size and computational complexity while maintaining performance. Specific pruning techniques are:
  - Weight pruning, Neuron pruning, Layer pruning
- **Knowledge Distillation** refers to training a smaller model (student model) by transferring knowledge from a larger, more complex model (teacher model) to achieve similar performance, effectively compressing the knowledge into a more compact form. Specific knowledge distillation techniques are:
  - Offline distillation, Online distillation, Self-distillation

For this research project, we explored existing open-source tools for model compression and ran experiments to reduce the size and increase inference speed of XLM-RoBERTa model, fine-tuned on the Standford Sentiment Treebank version 2 (SST2) dataset, while maintaining baseline accuracy.

Progress:

We conducted multiple experiments where the goal is to reduce inference time of a fine-tuned XLM-RoBERTa, with a minimal amount of loss in accuracy. Two baselines were first conducted: inference with a GPU and using only the CPU. Then for our experiments we applied the following different techniques from Hugging Face Optimum:

- Better Transformer: Pytorch optimisation method to achieve faster inference. 10% improvement in speed over baseline with no reduction in accuracy.
- ONNX: Converting the model from a Pytorch format to ONNX and running it in ONNX runtime showed significant performance increase. This can be further increased by quantizing and optimizing the model, and at virtually no cost on accuracy, roughly halving the size of the model. The inference speed of the model was multiplied by 3 with no reduction in accuracy.
- Neural compressor: The experiment failed to produce any significant results in performance improvement. However, it reduced the size of the model from 2.1GB to 500MB.
- Pruning the model: Followed by quantization and optimization, it showed the biggest improvement in speed and latency (almost x4), however at the cost of a reduced accuracy (-4%).

The model compression tools used in our experiments can be found at https://huggingface.co/docs/optimum/index.

For more information, please contact:
Alexandre Istrate (alexandre.istrate@statcan.gc.ca) or
Sayema Mashhadi (sayema.mashhadi@statcan.gc.ca).

PROJECT: XAI Labelling Assistant

The Explainable Artificial Intelligence (XAI) Labelling Assistant project is a research study designed to enhance the interaction between human expertise and machine learning models using "local explanations". As machine learning algorithms become more prevalent in the operations and priorities of Statistics Canada, the demand for high-quality labelled data has increased significantly. However, creating and maintaining this data can be labour-intensive and costly, especially for complex tasks that require the input of subject matter experts.

This research is primarily designed to assist experts in the Consumer Price Index (CPI) program to continuously evaluate model performance and identify relevant changes to the training data that affect model performance. Although the study is centred on CPI, its findings and methods could be beneficial for any program at Statistics Canada that requires ongoing data labelling.

This study aims to test three key hypotheses:

- Providing local explanations for a model's predictions can significantly enhance an annotator's confidence in these predictions, leading to a higher level of trust in the model.
- The use of local explanations can enhance the annotator's work experience, resulting in better quality annotations.
- The level of agreement between the model and the annotator in terms of the importance assigned to different factors (known as feature ranking agreement) can vary across different categories.

Progress:

In winter of 2023, realized progress includes: (i) The development of a testing process involving human annotators and 300 chosen data points. These data points come from six specific CPI product categories. (ii) The creation of a user-friendly dashboard that shows the model's SHAP-based local explanations (**SHapley Additive exPlanations**), helping the annotators understand the model's decision-making process. The next step is executing the process of collecting feedback from the annotators about their experiences.

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

PROJECT: Literature Review of Fair Machine Learning

As machine learning algorithms gain more widespread use, steps must be taken to minimize potential biases generated by model outputs. The objectives of this literature review were to understand different criteria for evaluating fairness, describe methods for incorporating fairness into machine learning methods, and discuss the benefits and drawbacks of these criteria and methods.

Progress:

The literature review summarized and compared various individual and group fairness metrics and introduced algorithm-specific methods of incorporating fairness. Briefly, individual fairness metrics focus on attaining fairness by analyzing decisions made on an observation level, whereas group fairness considers fairness and equality on a class basis (Barocas, Hardt and Narayanan, 2019). Fair methods in the classification, clustering, adversarial learning, and recommender systems settings were covered (Chierichetti, Kumar, Lattanzi and Vassilvitskii, 2017; Zhang, Lemoine and Mitchell, 2018; Yang and Stoyanovich, 2017). Guiding suggestions for determining which metrics to use were also discussed. The final deliverable was an internal document (Wang-Lin, 2023) that can serve as an introductory reference for Fair Machine Learning within the Data Science Division, which can be applied in various use cases to generate fair outputs such as imputing for demographic variables in surveys. This document falls under the Respect for People pillar of the Framework for Responsible Machine Learning as detailed in "Responsible use of machine learning at Statistics Canada" (Bosa, 2021), and can be used to assist in fairness assessments of projects at Statistics Canada.

For more information, please contact:
Angela Wang-Lin (angela.wang-lin@statcan.gc.ca).

References

Barocas, S., Hardt, M. and Narayanan, A. (2019). *Fairness and Machine Learning*. fairmlbook.org.

Bosa, K. (2021). Responsible use of machine learning at Statistics Canada. Retrieved from https://www.statcan.gc.ca/en/data-science/network/machine-learning.

Chierichetti, F., Kumar, R., Lattanzi, S. and Vassilvitskii, S. (2017). Fair clustering through fairlets. *Advances in Neural Information Processing Systems*, 30.

Wang-Lin, A. (2023). Literature Review of Fairness in Machine Learning. Unpublished internal Statistics Canada document.

Yang, K., and Stoyanovich, J. (2017). Measuring fairness in ranked outputs. In *Proceedings of the 29th International Conference on Scientific and Statistical Database Management,* pp. 1-6.

Zhang, B.H., Lemoine, B. and Mitchell, M. (2018). Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society,* pp. 335-340.

PROJECT: Multimodal Classification in Deep Learning

This research addresses how the standard classification approach used in deep learning (softmax activation function with cross-entropy loss) is implicitly unimodal in nature as it involves mapping all inputs of a given class, no matter how complex the distribution over inputs may be, to a single region in the embedding space determined by the class weight. Subsequently, this research proposes to use multiple class centres to improve the performance of multimodal classification in deep learning by increasing the diversity of prototypical class representations. This research outlines approaches that Statistics Canada can use to improve classification in a multimodal setting.

Progress:

The report first demonstrates how softmax function with cross-entropy loss-based classifiers implicitly represents class embeddings as unimodal distributions within the embedding space, centered about a prototypical class representation. Alternative methods in literature that address this limitation, such as SoftTriple loss and Prototypical Parts Network, are discussed (Qian, Shang, Sun, Hu, Li and Jin, 2019; Chen, Li, Tao, Barnett, Rudin and Su, 2019). Next, the report details a proposed approach for multimodal classification in deep learning called K subcentres via K-means clustering: first, train a standard softmax with cross-entropy classifier, then find K distinct representative instances for each class via their embeddings which are subsequently used for classification. Two variations of this method, interpretable K subcentres via K-means clustering and learned K subcentres via K-means clustering initialization, are also introduced. Experiment results from SoftTriple, softmax and cross entropy, and the three aforementioned K subcentres variant classifiers trained on CIFAR-10 dataset show that the learned K subcentres via K-means clustering initialization classifier performed the best. To solidify this finding, further experimentation on additional datasets, model architectures and replicates is encouraged.

For more information, please contact:
Nicholas Denis (613-618-9948, nicholas.denis2@statcan.gc.ca).

References

Chen, C., Li, O., Tao, D., Barnett, A., Rudin, C. and Su, J.K. (2019). This looks like that: Deep learning for interpretable image recognition. *Advances in Neural Information Processing Systems*, *32*.

Qian, Q., Shang, L., Sun, B., Hu, J., Li, H. and Jin, R. (2019). Softtriple loss: Deep metric learning without triplet sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision,* pp. 6450-6458.

PROJECT: Modern Processes to Predict Building's Daily Energy and Capital Costing - BTAP

Natural Resources Canada (NRCan) performs complex modelling to derive the hourly energy usage and capital costing values for buildings. Statistics Canada has collaborated with NRCan to research how to utilize Machine Learning models to act as surrogate models for predicting the daily energy usage and capital costing values for building data. Utilizing these predictions, a basic analysis can be performed on buildings at a significantly faster rate when compared to the current approach.

Progress:

Following an initial research phase which explored existing surrogate modelling solutions, such as the BESOS platform (Faure, Christiaanse, Evins and Baasch, 2019), the project has now been extended and operationalized as an open-source proof-of-concept which can be run on various platforms. The work being done within this project has been brought to the building surrogate model research community and will continue to be enhanced while considering the research of other groups. The program can train Machine Learning models dynamically from any valid dataset and utilize a trained model to output the daily energy and capital costing predictions. This will work for any building type and climate zone, with the program providing outputs to allow the user to analyze the performance of the Machine Learning model.

The program can output the total daily energy and capital costing values with higher precision or can output breakdowns of the total values with less precision (such as the daily electricity and gas usage). The efficiency of the approach indicates that the use of surrogate models is good for allowing analysts to quickly perform a basic analysis of a building before performing more complex simulations. The effectiveness of the predictions will vary for the data used as input.

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

Reference

Faure, G., Christiaanse, T., Evins, R. and Baasch, G.M. (2019). BESOS: A collaborative building and energy simulation platform. In *Proceedings of the 6th ACM International Conference on Systems for Energy-Efficient Buildings, Cities, and Transportation,* pp. 350-351.

PROJECT: Building a Generalized Crowdsourcing Application

Generalized crowdsourcing is a concept that involves leveraging a large, decentralized group of people to accomplish a task or address a problem. It has been applied in various fields such as data annotation, content creation, and problem-solving. This research aims to investigate the implementation of a crowdsourcing application using open-source technologies. It aims to explore the advantages of a generalized crowdsourcing system and analyze the benefits and obstacles associated with this approach.

Progress:

In the initial phase of this research, our focus was to investigate existing work in the field of crowdsourcing. We examined examples such as the OpenStreetMap (OSM) crowdsourcing pilot, the COVID-19

crowdsourcing initiative by the Government of Canada, and the Crowdsourcing-Cannabis project also conducted by the Government of Canada (Statistics Canada, 2020; Statistics Canada, 2022).

Additionally, we utilized our expertise to design an architecture and create a schema for storing data in SQLite Database. We developed demonstration user interface pages to showcase our ideas and progress in subsequent phases of the research. Our work was summarized and published in the Data Science Network, specifically addressing the reduction of data gaps in training machine learning algorithms through the use of a generalized crowdsourcing application (Manda and Widhani, 2023).

For more information, please contact:
Nikhil Widhani (nikhil.widhani@statcan.gc.ca),
Chatana Mandava (chatana.mandava@statcan.gc.ca) or
Ekramul Hoque (ekram.hoque@statcan.gc.ca).

References

Mandava, C., and Widhani, N. (2023). Reducing data gaps for training machine learning algorithms using a generalized crowdsourcing application. Statistics Canada. https://www.statcan.gc.ca/en/data-science/network/reducing-data-gaps.

Statistics Canada (2020). Crowdsourcing: Impacts of COVID-19 on Canadians' experiences of discrimination public use microdata file. https://www150.statcan.gc.ca/n1/en/catalogue/45250008.

Statistics Canada (2022). Statistics Canada data strategy. https://www.statcan.gc.ca/en/about/datastrategy.

PROJECT: UNECE Exploration of Membership Inference Attacks and Differential Privacy

Within the United Nations Economic commission for Europe (UNECE) different research teams from international National Statistics Offices have collaborated on various projects. One such project is the exploration of how differential privacy can be used to protect a Machine Learning model from Membership Inference Attacks, which aim to identify data which has originally been used to train the model.

Progress:

The completed project has identified that the use of differential privacy successfully reduces the effectiveness of Membership Inference Attacks on deep learning models. One Membership Inference Attack against Machine Learning models that has been tested uses shadow models and synthesized data to build an attack model which predicts whether a datapoint is or is not part of the target model's training dataset (Shokri, Stronati, Song and Shmatikov, 2017). When applying differential privacy to the training process, it has been observed that significantly less data can accurately be identified as being within the original training set. Furthermore, the tests highlight how the application of differential privacy to the input data itself or to the training process will reduce the effectiveness of the trained model. Thus, it is important to identify how to balance the trade-off between privacy and utility, while considering from which attacks a Machine Learning model must be protected.

The results from the three different research tracks within the UNECE group have been published online in a report and provide detailed results and explanations (United Nations Economic Commission for

Europe, 2023). These include the results from the above work within the Private Machine Learning track and from both the Private Set Intersection and Open Consultation tracks.

For more information, please contact:
Julian Templeton (julian.templeton@statcan.gc.ca) or
Benjamin Santos (438-459-7721, benjamin.santos@statcan.gc.ca).

References

Shokri, R., Stronati, M., Song, C. and Shmatikov, V. (2017). Membership inference attacks against machine learning models. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 3-18, May 2017.

United Nations Economic Commission for Europe (2023). *UNECE Project on Input Privacy Preservation*. https://statswiki.unece.org/x/mQCQFw.

PROJECT: A new method to choose the number of clusters of a mixed dataset under Kproto clustering, Part I: Introduction

Identifying homogenous subgroups without prior knowledge on the number of these subgroups aids clinicians and policy makers in tailoring better strategies to introduce interventions according to the characteristics of these homogenous subgroups. Clustering is traditionally used to identify these subgroups. Partition clustering has been used extensively due to its efficiency over other clustering methods but requires the number of clusters to be known in advance. While there exist many measures to estimate the number of clusters for datasets with only numeric variables, no approaches exist in the literature to estimate the number of clusters for mixed datasets (i.e., those that contain both numeric and categorical variables). In the first part of this research, we introduced and demonstrated a new method to choose the number of clusters in a mixed dataset so that the clusters are stable, have maximized and stable categorical variable contribution, while using the extensively studied, Kproto clustering (Huang, 1997 and 1998; Szepannek, 2018; Szepannek and Aschenbruck, 2019).

Progress:

In the first part of this research, this new method was applied on a dataset which was large enough to avoid the problems arising from low or zero counts in cells created by cross-classifying categorical variables and clusters when the dataset is clustered using Kproto clustering. The Canadian Social Survey COVID-19 and Well-being (CSS-CW) Wave 1 dataset (9,278 instances with three numeric variables and seven categorical variables) was used. The number of clusters of this mixed dataset was chosen by combining the results of a stability analysis of cluster assignment and the results of categorical variables contribution to clusters. The first step was carried out using adjusted mutual information (Vinh and Epps, 2009; Vinh and Bailey, 2010; and Chiquet, Rigail and Dervieux, 2019) over varying number of clusters under completely randomized design. The second step that utilized the Pearson Chi-Square (CS) contribution, assured maximized and stable categorical variables contribution to clusters. Combining these results with subject matter knowledge helped identify the most important categorical variables and choose the number of clusters in the CSS-CW dataset. A reproducible R-program was written to carry out this part of the research (Sivathayalan, Chu and Le Moullec, 2023).

For more information, please contact:
Ahalya Sivathayalan (613-302-6647, ahalya.sivathayalan@statcan.gc.ca).

References

Chiquet, J., Rigail, G. and Dervieux, V. (2019). Efficient Computations of Standard Clustering Comparison - package aricode; CRAN.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In Proceedings of the 1st Pacific Asia Knowledge Discovery and Data Mining Conference. Singapore: World Scientific.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data Mining Knowledge Discovery*, vol 2, 283-304.

Sivathayalan, A., Chu, K. and Le Moullec, J. (2023). A new method to choose the number of clusters of a mixed dataset under Kproto clustering, Part I: Introduction; working paper, Statistics Canada.

Szepannek, G. (2018). clustMixType: User-Friendly Clustering of Mixed-Type Data in R; The R journal Vol. 10/2.

Szepannek, G., and Aschenbruck, R. (2019). k-Prototypes Clustering for Mixed Variables-Type Data: Package clustMixType; CRAN.

Vinh, N.X., and Epps, J. (2009). A Novel Approach for Automatic Number of Clusters Detection in Microarray Data based on Consensus Clustering. Ninth IEEE International Conference on Bioinformatics and Bioengineering.

Vinh, N.X., and Bailey, J. (2010). Information theoretic measures for clustering comparison: Is a correction for chance necessary? *Proceedings of the 26th International Conference on Machine Learning*, Montreal, Canada.

PROJECT: A new method to choose the number of clusters of a mixed dataset under Kproto clustering, Part II: Evaluation

Partition clustering has been traditionally used to identify homogenous subgroups without prior knowledge of the number of these subgroups due to its efficiency over other clustering methods. There exist many clustering methods and measures in the literature to obtain the number of clusters of datasets with only numeric variables and to cluster them accordingly, but not for datasets with both numeric and categorical variables. In the first part of this research (Sivathayalan, Chu and Le Moullec, 2023), a new method was introduced to choose the number of clusters of a mixed dataset so that clusters are stable, have maximized and stable categorical variable contribution, while applying Kproto clustering (Huang, 1997 and 1998; Szepannek, 2018; Szepannek and Aschenbruck, 2019). In this second part of the research, we applied this method on five different datasets of varying types, quality and sizes ranging from approximately 150 to 5,000 instances to understand the issues that may be faced while applying this new method under Kproto clustering.

Progress:

Previously (Sivathayalan, Chu and Le Moullec, 2023), the new method was applied on a large enough dataset such that there were no zero or few counts while verifying the categorical contribution to clustering and while carrying out stability analysis of cluster assignment. This second part of the research

addressed one of the future works stated in the previous part. Here, we applied this new method on datasets of varying number of categorical variables and numeric variables, quality and sizes ranging from 150 to 5,000 to understand the problems that may arise. This evaluation was carried out by making use of the reproducible R-program created during the part I of this research (Sivathayalan, Chu and Le Moullec, part I, 2023). The following five datasets were used: i. Opioid overdose dataset (4,196 instances with two numeric and 12 categorical variables) ii. German Credit dataset (1,000 instances with seven numeric and 13 categorical variables), iii. Credit Approval dataset (690 instances with six numeric and nine categorical variables), iv. Heart dataset (303 instances with six numeric and seven categorical variables), v. Lymphography dataset (148 instances with two numeric and 16 categorical variables). Note that the first data set is from Statistics Canada and the last four data sets are public datasets (Dua and Graff, 2019). A suitable number of clusters for each dataset, different issues faced in choosing the number of clusters, and steps to followed to overcome these issues were reported. Further, the importance of combining the chi-square contribution of categorical variables with subject matter knowledge was stated while identifying the important variables in clustering (Sivathayalan and Le Moullec, 2023).

For more information, please contact:
Ahalya Sivathayalan (613-302-6647, ahalya.sivathayalan@statcan.gc.ca).

References

Dua, D., and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

Huang, Z. (1997). Clustering large data sets with mixed numeric and categorical values. In *Proceedings of the 1st Pacific Asia Knowledge Discovery and Data Mining Conference*, Singapore: World Scientific.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large datasets with categorical values. *Data Mining Knowledge Discovery*, vol 2. No. 3.

Sivathayalan, A., Chu, K. and Le Moullec, J. (2023). A new method to choose the number of clusters of a mixed dataset under Kproto clustering, Part I: Introduction. Working paper, Statistics Canada.

Sivathayalan, A., and Le Moullec, J. (2023). A new method to choose number of clusters of a mixed dataset under Kproto clustering, Part II: Evaluation. Working paper, Statistics Canada.

Szepannek, G. (2018). clustMixType: User-Friendly Clustering of Mixed-Type Data in R. *The R journal*, Vol. 10/2.

Szepannek, G., and Aschenbruck, R. (2019). k-Prototypes Clustering for Mixed Variables-Type Data: Package clustMixType. CRAN.

3 Estimation issues in surveys

PROJECT: Accuracy of machine learning predictions

To reduce costs and speed up the production of statistics, a specific survey occasion of a repeated survey may be replaced by predictions of the finite population totals or means of interest. Each predicted total or mean is obtained by making a prediction for each population unit, based on available data, and computing the total or mean of these predictions across all the units. For example, one may predict the yield of different crops at a future date based on fixed covariates, which comprise remote sensing and agro-climatic variables, as well as historical data, where the latter includes past responses and covariates (Statistics Canada, 2020; National Academies of Sciences, Engineering, and Medicine, 2023, Chapter 8.3). Increasingly, the predictions are made with machine learning methods (Chu, 2022) and must be published with a measure of their accuracy (Yung, Tam, Buelens, Chipman, Dumpert, Ascari, Rocci, Burger and Choi, 2022). Unfortunately, cross-validation and similar techniques (Hastie, Tibshirani and Friedman, 2001, chapter 7) are inadequate because they provide a measure of the unconditional error, treating the covariates as random. Instead, we are more interested in a conditional measure. Also, our focus is on the uncertainty of the predicted total instead of that of a unit prediction, which is commonly measured by the test error.

Progress:

A bootstrap methodology has been developed for a general setup with two finite population of units. In the first population, which provides the historical data, the fixed covariates and responses are given for each unit. In the second population, only the fixed covariates are observed for each unit. All the responses are assumed to be mutually independent and such that a response is the sum of a mean function of the covariates and a random error with a zero mean and a variance given by another function of the covariates. The mean and variance functions are the same across the two populations, with each random error following the same distribution (i.e., the distribution of the scaled error is the same for both populations, e.g., the standard normal distribution). The methodology was presented at the annual conference of the Statistical Society of Canada (Dasylva, Beaumont, Bosa and Maranda, 2023).

For more information, please contact:
Abel Dasylva (613-408-4850, abel.dasylva@statcan.gc.ca).

References

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

Dasylva, A., Beaumont, J.-F., Bosa, K. and Maranda, G. (2023). Measuring the accuracy of a prediction for a finite population total. Presentation at the 2023 Annual Conference of the Statistical Society of Canada, May 2023.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The Elements of Statistical Learning*. New York: Springer.

National Academies of Sciences, Engineering, and Medicine (2023). *Toward a 21st century national data infrastructure: Enhancing survey programs by using multiple data sources*. Washington, DC: The National Academies Press. Available at https://doi.org/10.17226/26804.

Statistics Canada (2020). Integrated crop yield modelling using remote sensing, agroclimatic data and survey data. Available at https://www.statcan.gc.ca/en/statistical-programs/document/5225_D1_T9_V1.

Yung, W., Tam, S.-M., Buelens, B., Chipman, H., Dumpert, F., Ascari, G., Rocci, F., Burger, J. and Choi, I. (2022). A quality framework for statistical algorithms. *Statistical Journal of the IAOS*, 38, 291-308. Available at https://content.iospress.com/download/statistical-journal-of-the-iaos/sji210875?id=statistical-journal-of-the-iaos%2Fsji210875.

PROJECT: Modeling exercise to predict water quantity for non-surveyed years for two environmental surveys

The Industrial Water Survey (IWS) is a biennial survey that has collected, since 2005, data on water use in industries in the mining, manufacturing and thermoelectric sectors. The survey was not carried out in 2019 because of the Covid-19 pandemic. Data from this survey are required for the calculation of the Canadian Environmental Sustainability Indicators published by Environment and Climate Change Canada. The Survey of Drinking Water Plants (DKWP) is also a biennial survey and since 2007 has been collecting data on the production of drinking water. These data are used to track the state of water reserves on a regional basis in Canada. The modeling project consists of designing models that would predict the amount of water used in industries for the first survey and the amount of drinking water produced for the second survey for the years when these surveys are not carried out.

To predict the amount of water used by industry and the amount of drinking water produced, the project explored several modeling methods, including linear regression (Hastie, Tibshirani and Friedman, 2009), smoothing spline regression (Silverman, 1985), exponential smoothing (Hyndman, Koehler, Ord and Snyder, 2008), multiple imputation by sequential regression (Van Buuren, 2018, and Wang, Akande, Poulos and Li, 2022), regression by “eXtreme Gradient Boosting (XGBoost)” (Chen and Guestrin, 2016 and Boehmke and Greenwell, 2020), etc.

Progress:

The modeling project made it possible to predict the amount of water used in the various industries of the manufacturing, mining and thermoelectric sectors for the years not surveyed of the IWS (the even years from 2006 to 2018 and the year 2019). It also predicted the amount of drinking water produced by water treatment plants for the non-surveyed years of the DKWP survey (the year 2009 and the even years from 2008 to 2020). Some predictions from the project, including those from 2019 from the IWS, have already been used in Canada’s system of environmental-economic accounting for the publication on water use in 2019 (Statistics Canada, 2022a and 2022b). Modeling work continues in order to improve the models as new data are collected and also to extend the modeling to other key variables in these surveys. The project will end with the writing of an article in the coming months.

For more information, please contact:
Martin Hamel (613-854-2827, martin.hamel@statcan.gc.ca).

References

Boehmke, B., and Greenwell, M. (2020). *Hands-On Machine Learning with R*. Chapman & Hall.

Chen, T., and Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794. New York, NY, USA: ACM. https://doi.org/10.1145/2939672.2939785.

Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. New York: Springer.

Hyndman, R.J., Koehler, A.B., Ord, J.K. and Snyder, R.D. (2008). *Forecasting with Exponential Smoothing: The State Space Approach*. Springer.

Silverman, B.W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting (with discussion). *Journal of the Royal Statistical Society Series B*, 47, 1-52.

Statistics Canada (2022a). Canadian System of Environmental–Economic Accounts: Water use, 2019. https://www150.statcan.gc.ca/n1/daily-quotidien/221219/dq221219d-eng.htm.

Statistics Canada (2022b). Physical flow accounts: Water Use, 2019. https://www150.statcan.gc.ca/n1/pub/11-627-m/11-627-m2022087-eng.htm.

Van Buuren, S. (2018). *Flexible Imputation of Missing Data*. 2nd Edition. Chapman & Hall/CRC Interdisciplinary Statistics Series. https://stefvanbuuren.name/fimd/.

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

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

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

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

Progress:

The writing of an article (Toupin and Martin, 2023) for submission to a scientific journal is complete. It is currently being reviewed internally.

For more information, please contact:
Marie-Hélène Toupin (343-573-1872, marie-helene.toupin@statcan.gc.ca).

References

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

Toupin, M.-H., and Martin, V. (2023). Degrés de liberté liés à l'estimation du questionnaire détaillé du recensement. Internal report (will be submitted to a scientific journal), Statistics Canada.

PROJECT: Generalization and application of a simple, exact and optimal method of sample allocation

In 2017, in order to assess the stability of sample allocation in the annual economic surveys of the Integrated Business Statistics Program (IBSP), a colleague found in the literature a very simple algorithm obtaining exactly (integer stratum sample sizes, n_h) the Neyman allocation. The algorithm is detailed in the articles cited in references (Wright 2014, 2016, 2017 and 2020). The aim was to implement the method to compare it with the more complex one integrated into the IBSP (power allocation, establishment clusters, sampling cells, etc.). We found the method developed by the author very elegant in its simplicity and efficiency. After the work carried out on the project, we always considered the possibility of generalizing the method to power allocation (and not just Neyman), and above all, to see how we could integrate stratification into the method, as well as the other components used in IBSP. For example, the author of articles on the topic never dealt with stratification or sampling cells.

The relevance of the project stems from the important link between the quality of estimates from a probability source (an establishment survey, for example) and efficient sampling. Obtaining a Neyman (or power) allocation in an exact and easy way, while allowing one to define a more appropriate stratification to reduce the targeted variance, is precisely a way to improve the sampling efficiency, at less cost. The application of the project is mainly related to surveys using stratification according to a measure of size (income, sales, for example) and a relatively small population (with small strata).

Progress:

The latest articles on the topic have been read and we still found no trace of complexities related to sampling for generalizing the method to a practical problem. Only the allocation of units within pre-determined strata is covered. Thus, the relevance of integrating the allocation method into a complete sampling methodology, such as that of the IBSP, was confirmed.

A SAS program was developed to apply the method starting with known strata and sample size. The generalization to power allocation has been completed and added to this program. We also wanted to be able to add a component called "sampling cells" to follow the IBSP concepts, which was done recently. It will thus be possible to apply this program to some surveys integrated into the IBSP for comparison

purposes with respect to the objective function of a sample allocation (Neyman or power). Finally, some simulations were conducted in connection with the addition of a stratification component, but more work is required.

For more information, please contact:
Pierre-Olivier Julien (613-716-6174, pierre-olivier.julien@statcan.gc.ca).

References

Wright, T. (2014). A Simple Method of Exact Optimal Sample Allocation under Stratification with Any Mixed Constraint Patterns. Research Report Series of US Census Bureau, https://www.census.gov/library/working-papers/2014/adrm/rrs2014-07.html.

Wright, T. (2016). Two Optimal Exact Sample Allocation Algorithms: Sampling Variance Decomposition is Key. Research Report Series of US Census Bureau, https://www.census.gov/content/dam/Census/library/working-papers/2016/adrm/rrs2016-03.pdf.

Wright, T. (2017). Exact Optimal Sample Allocation: More Efficient than Neyman. *Statistics and Probability Letters*, 129, 50-57.

Wright, T. (2020). A general exact optimal sample allocation algorithm: With bounded cost and bounded sample sizes. *Statistics and Probability Letters*, 165, 108829.

4 Confidentiality

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 two projects described below. The Centre for Confidentiality and Access group 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.

PROJECT: Synthetic data

Working towards more options for access by 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.

Progress:

In 2022, a synthetic version of a cross-sectional census-based database linked to administrative data was developed for the new Canadian Retirement Income System Model (CRISM). A first version of the synthetic database was made available to external researchers. The next challenge was to develop strategies to synthesize the longitudinal aspect of the database. A donor-imputation technique was applied to add life-trajectory information to family units and then wages and employment earnings were synthesized at the individual-level for the 1966 to 2015 period using a Classification and Regression Tree

(CART) approach. The goal for next year is to synthesize additional variables related to the Canadian Pension Plan and release an augmented version of the synthetic database with the synthetic historical information.

The High-level Group for the Modernization of Official Statistics from UNECE published its Starter Guide on synthetic data in 2023 (United Nations, 2023). Statistics Canada contributed to multiple parts of the guide.

For more information, please contact:
Heloise Gauvin (343-597-3490, heloise.gauvin@statcan.gc.ca) or
Steven Thomas (613-882-0851, steven.thomas@statcan.gc.ca).

Reference

United Nations (2023). Synthetic Data for Official Statistics: A Starter Guide | UNECE. Geneva: United Nations. Available at https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide.

PROJECT: Assessment of reconstruction attack risk using Statistics Canada Census data

The publication of more detailed disaggregated data can increase transparency and provide important information on underrepresented groups. Developing more varied and readily available access options increases the amount of information available to and produced by researchers. Increasing the breadth and depth of the information released allows for a better representation of the Canadian population, but also puts a greater responsibility on Statistics Canada to do this in a way that respects the privacy of individuals and protects the confidentiality of their data. It is helpful to develop tools which allow Statistics Canada to evaluate the risk from the additional data granularity.

One strategy to help identify specific situations where there is a risk of disclosure is through a database reconstruction. This evaluation tool creates a microdata file consistent with a collection of given statistics, say, those published online. Risks are identified when observations are uniquely and consistently linked with certainty to individuals in the dataset. The objective of this project is to develop the tools necessary to evaluate the risks through a database reconstruction and to use these tools to quantify the reconstruction risk to Canadian Census data.

Progress:

A methodology was developed to reconstruct the Canadian Census database based on published tables following a procedure outlined by Garfinkel, Abowd and Martindale (2019). Using synthetic data allowed us to mount various attacks on the Canadian Census data under various conditions and to evaluate how successful they were, and hence, how much of a risk those attacks pose. It was found that mounting a database reconstruction takes significant effort, technical expertise, and computing power suggesting that the probability of such an attack was low. They study also suggested that our standard statistical disclosure control (SDC) techniques (e.g., rounding) are reasonably good at hindering disclosure risks in the conditions that were evaluated. Measures of risk including match rate between reconstructed and true data were developed, but work remains to be done in interpreting these measures and their limitations.

These results were presented at Statistics Canada's 2022 International Methodology Symposium and will be published in the conference proceedings (Abado and Stefan, 2022).

For more information, please contact:
Mathew Abado (519-868-4792, mathew.abado@statcan.gc.ca) or
Steven Thomas (613-882-0851, steven.thomas@statcan.gc.ca).

References

Abado, M., and Stefan, G. (2022). Reconstruction attack risk using Statistics Canada census data. Proceedings: *Symposium 2022*, *Data Disaggregation: Building a more-representative data portrait of society*, Statistics Canada, Ottawa (to appear).

Garfinkel, S., Abowd, J.M. and Martindale, C. (2019). Understanding database reconstruction attacks on public data. *Communications of the ACM*, 62, 46-53.

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,
- development and support for seasonal adjustment and trend-cycle estimation, and
- time series modelling and forecasting, particularly in the context of real-time estimation.

PROJECT: Consultation and training in time series methods

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

Progress:

A course on seasonal adjustment was delivered during the year through the Statistics Canada training centre (Statistics Canada, 2022). Courses on time series modelling and forecasting as well as on seasonal adjustment and reconciliation techniques were delivered to participants from external agencies via a remote format. Members of the centre also participated in outreach and training to other groups in

Statistics Canada on time series topics as part of training for recent recruits (methodology branch seminar series for recruits and the data navigator course).

The centre has also offered consultation to various internal programs (autoregressive and regARIMA modelling, trend estimation, calendarization, backcasting). In particular, the centre provided time series support to the System of National Accounts in a number of areas, including the monthly Gross Domestic Product. Representatives from the centre also periodically attend a weekly analyst forum to maintain a presence in the analyst community.

The centre regularly consults on back-casting to preserve or restore comparability across time, and has worked to finalize a directive on time series continuity for Statistics Canada's programs. This is a joint initiative with the System of National Accounts and the recent work involved presenting the document to upper management for comments and approval, and working with dissemination to finalize the document for public release.

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

For more information, please contact:
Etienne Rassart (613-410-3640, etienne.rassart@statcan.gc.ca).

Reference

Statistics Canada (2022). Workshops, training and references. https://www.statcan.gc.ca/eng/wtc/training.

PROJECT: Support and enhancement of 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 Time Series Processing System (Ferland, 2022),
- the Generalized System G-Series, for benchmarking and raking/reconciliation/balancing (Statistics Canada, 2016),
- the Seasonal Adjustment Dashboard (Verret, 2021).

Progress:

The Time Series Processing System is a customizable SAS based application to apply time series techniques including seasonal adjustment, benchmarking and reconciliation, used extensively in the production of seasonally adjusted estimates for sub-annual programs within Statistics Canada (many of them mission critical). The system is in a mature and stable state. However, it requires updating on an ongoing basis to broaden functionality and address new needs of programs in the agency. For the longer term, a new version of the system to allow flexibility to incorporate tools and new techniques available from open-source software is being considered.

Investigations were made to evaluate tools available to apply benchmarking, raking and seasonal adjustment through open-source tools, including those available in R, Python, and Java-based programs developed and published by the European Commission. Prototype R functions were completed with functionality equivalent to the benchmarking and raking procedures in G-Series and are currently being evaluated for internal use. Statistics Canada continued its involvement in the Seasonal Adjustment Centre of Excellence of Eurostat as a partner organization to participate in discussions and development of related tools.

A number of improvements to the Seasonal Adjustment Dashboard were implemented this year. In particular, the dashboard was adapted to quarterly data, new graphs were added, and minor bugs were resolved. The dashboard is in the process of being deployed for an additional labour program.

For more information, please contact:
Etienne Rassart (613-410-3640, etienne.rassart@statcan.gc.ca).

References

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

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.

PROJECT: Development and support for seasonal adjustment and trend-cycle estimation

The objective of this project is to conduct analyses and evaluations of new methods and techniques for seasonal adjustment and trend-cycle estimation as well as consultation and centralization of expertise in applying seasonal adjustment.

Progress:

The Time Series Research and Analysis Centre provided extensive quality assurance for seasonal adjustment as the effects of the COVID 19 pandemic continued. Exchanges were held with representatives from other national statistical offices, via emails as well as conferences and virtual meetings. The exchanges included members from Eurostat, the Office for National Statistics, the United States Census Bureau, the United States Bureau of Labor Statistics, and others. These exchanges were extremely valuable to compare and contrast approaches for the ongoing treatment of pandemic effects. The centre shared findings and recommendations from these consultations with relevant contacts within the agency through ad hoc consultations to ensure that the approach used was generally consistent across programs.

A strategy to reduce the increased validation applied during the initial pandemic shocks was developed to reduce support on an ongoing basis. As each economic indicator stabilizes, a strategy is being developed to gradually transition to the quality assurance practices that were in place prior to the pandemic without causing adverse effects to the seasonally adjusted estimates in real-time and retrospectively. State-space models were investigated to produce early indications of structural breaks and this work will continue going forward.

The Time Series Research and Analysis Centre also provided extensive consultations on seasonal adjustment within the agency for programs not formally supported by the team. Given the challenges for

seasonal adjustment that are introduced by the pandemic and its economic shocks periodic consultations were held to offer advice and consultation on practical problems with several of the programs in the national accounts.

As well, trend-cycle estimation methods were evaluated in the context of the COVID-19 pandemic. In particular, given the severe nature of the economic shocks, the trend-cycle currently published for selected programs at Statistics Canada (based on the cascade linear filter proposed in Dagum and Luati, 2009) may present an overly smooth impression of the economy so a number of alternative measures were identified including breaking the series at a particular point, and introducing outlier effects to model shocks more directly. These methods have been applied to the results for several programs and presented externally to determine if the adjustments to the methodology should be adopted (Matthews, 2022a and Matthews, 2022b).

For more information, please contact:
Etienne Rassart (613-410-3640, etienne.rassart@statcan.gc.ca).

References

Dagum, E.B., and Luati, A. (2009). A cascade linear filter to reduce revisions and false turning points for real time trend-cycle estimation. *Econometric Reviews*, 28, 40-59.

Matthews, S. (2022a). Estimating the Trend-Cycle in Topsy-Turvy Times. Seasonal Adjustment Practitioners Workshop, United States Census Bureau.

Matthews, S. (2022b). Estimating the Trend-Cycle in Topsy-Turvy Times. 2nd Workshop on Time Series Methods for Official Statistics, Eurostat.

PROJECT: Time Series modelling and forecasting, particularly in the context of real-time estimation

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

Progress:

Evaluation of statistical models in this context continued, comparing machine learning models with other time series approaches with ARIMA-X models and state space models being the most appealing candidates. State space models include an application of dynamic factor models, which combine dimensionality reduction, mixed frequency data and other practical issues and are increasingly being used in nowcasting applications internationally. The evaluation of dynamic factor models was continued in partnership with Dr. Rafal Kulik from the University of Ottawa.

A further evaluation of this approach was conducted as a proof-of-concept exercise in the context of producing advanced indicators of energy statistics. This work involved comparisons of aggregate time series modelling with a microdata-based approach involving prediction or imputation of non-responding units. The work suggested that the microdata-based approach provided opportunity to use early respondents, so efforts were centered around development of efficient imputation models. This work will be presented at the upcoming meeting of the Statistics Canada Advisory Committee on Statistical Methods (Le Moullec and Matthews, 2023). An invited session was organized on the topic of real-time estimation at the International Statistical Institute's World Statistics Congress in July of 2023.

The centre conducted a nowcasting project to develop a more accurate method of estimating renovation activity expenditures without sacrificing quality, as part of the monthly Investment in Building Construction Program, a key economic indicator that quantifies the state of building construction investment in the economy. In this program – a primary input to monthly and quarterly Gross Domestic Product–, outputs are modelled by using building permit data to estimate levels of construction investment. A data rich approach combined with regression with ARMA errors was used to achieve this goal (Patak and Plunkett, 2023).

A draft document was prepared to outline a framework for advance indicators, to promote standardization of terminology for producing advance indicators (including model-based nowcasts), to outline criteria to support the decision to publish new advance indicators, and provide guidance on appropriate methods for nowcasting along with advantages and disadvantages (Matthews, 2022).

For more information, please contact:
Etienne Rassart (613-410-3640, etienne.rassart@statcan.gc.ca).

References

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

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

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

5.2 Economic Generalized Systems

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

Progress:

A typical volume of support cases for G-Sam, BANFF and G-Est was processed by the project team. Most of these were resolved with suggestions on how to apply the systems in practical terms, however several required more involvement.

Extensive testing of the three economic generalized systems was performed to ensure a smooth migration to Statistics Canada's new cloud infrastructure. Testing revealed issues due to changes to the numerical optimization procedure in recent versions of SAS (in particular the version installed in the cloud). These issues were also encountered with the generalized system G-Series (time series) and G-Confid (disclosure control). Members of the unit participated in a working group to investigate these issues, and a summary

of findings was presented to the Generalized Systems Steering Committee (Statistics Canada, 2023). For G-Est, cases are currently treated on an ad-hoc basis, though further research is planned to determine if an automated scaling fix similar to one developed for G-Series could be applied.

Progress continued on the modernization initiative for the economic generalized systems – leveraging open-source tools and including new leading-edge method. The modernization of BANFF was selected as an investment proposal by the agency this year and the project is underway (software launch planned for December 2024 and project completed in March 2025). The evaluation of options for G-Sam and G-Est will be completed in the coming year, with resulting recommendations presented to the Statistical Generalized Systems Steering Committee.

The Banff team has begun working with generalized systems programmers on the Banff Modernization Project, which consists of porting existing Banff procedures from SAS to Python, and the development of a new Banff Processor. A discussion of this initiative was presented at the United Nations Economic Commission for Europe (UNECE) Expert Meeting on Statistical Data Editing (Gray, 2022). Concurrently, plans for new Banff modules are in the early stages of research and development, including a generalized pro-rating and rounding system using open-source optimizers (Baillargeon, 2022). Work continues on ImpACT (Imputation Assessment and Comparison Tool), with prototypes for two out of the three modules completed.

Work has begun on G-Sam version 1.04, featuring several new features (Stinner, 2023). These include the introduction of a domain-level size constraint, new calculations for sampling variance and probabilistic bounds, as well as several new diagnostic outputs and bug fixes. These changes will better align G-Sam functionality with survey needs, increase the accuracy of the allocation model, and reduce both user and support troubleshooting costs. The release of version 1.04 in the next year will be accompanied by a new user guide, offering a comprehensive guide to the context, functionality, and application of the software. In addition, a research project exploring stratification via hierarchical clustering (McGrouther and Baillargeon, 2022) showed promise; additional research will be performed in the next year.

The use of G-Est's SEVANI module for estimating variance due to imputation has steadily increased, in particular for IBSP surveys, both to improve quality measures and as an input for the application of Random Tabular Adjustment (RTA) disclosure control. This has resulted in an increased number of support cases related to the SEVANI module. G-Est version 2.3.4 was released in November 2022, featuring improved nonresponse variance estimation performance, and a number of minor bug fixes (Statistics Canada, 2022). Additional minor improvements to SEVANI, to address rare and unusual bugs, will be included in the next G-Est release. A prototype for adapting bootstrap calibration range bounds has been written and tested. Additionally, Wilson confidence intervals for proportions have been reviewed and are under consideration for a future release.

Members of the team participated in training through formal courses with Statistics Canada's training centre, as well as seminars for recently recruited statisticians and other ad hoc presentations to analysts and other organisations. The team also contributed to the organisation of the UNECE Expert Meeting on Statistical Data Editing (October 2022).

For more information, please contact:
Etienne Rassart (613-410-3640, etienne.rassart@statcan.gc.ca).

References

Baillargeon, J. (2022). Bidirectional Pro-rating. Internal document, Statistics Canada.

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

McGrouther, S., and Baillargeon, J. (2022). Stratification via Hierarchical Clustering. Internal presentation, Statistics Canada.

Statistics Canada (2022). G-Est 2.03.004 Release Notes. Internal document, Statistics Canada.

Statistics Canada (2023). Generalized Systems Report on Achievements and Objectives. Internal document, Statistics Canada.

Stinner, M. (2023). G-Sam 1.04 Notes on Allocation. Internal document, Statistics Canada.

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 applications to interested parties inside and outside Statistics Canada.

Progress:

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

During the year, most of the methodological work focused on the maintenance, development and support for users of version 3.5 of G-Link on SAS servers in cloud computing. The development consisted of integrating blocking and pair status classification procedures from the splink software (developed at the U.K. Office for National Statistics) using python code.

The RLRC has also worked on a variety of other probabilistic linkage projects and hosted meetings with international colleagues. 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. The RLRC has also directed research on the evaluation of the quality of probabilistic linkage which has led to international exchanges with the National Institute of Statistics and Economic Studies of France, the National Bureau of Statistics of the

United Kingdom and the Italian statistical agency (ISTAT). This resulted in the organization of an invited session at the 2022 Statistics Canada Symposium.

For more information, please contact:
Abdelnasser Saïdi (613-863-7863, 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 Centres (RDCs).

Progress:

Consultations

Consultation services were provided as requested by internal and external clients. Between April 1, 2022 and March 31, 2023, DARC responded to 34 requests. The questions varied in complexity and included topics such as pooling survey cycles and adjusting weights for pooled data, guidelines for publishing model parameters, quality guidelines for data dissemination, analyses with linked data, and different types of regression analyses for complex surveys. DARC also helped clients with the implementation of statistical methods in SUDAAN, SAS, STATA, and R software.

Provision of Training and Training Material

DARC gave a presentation at the 2023 RDCs Annual Analyst Conference about working with linked data. Topics included linkages at Statistics Canada, the Social Data Linkage Environment, and linked products with linkage-adjusted weights.

DARC presented at Statistics Canada's internal Data Interpretation Workshop on data analysis with complex survey data and on the use of descriptive statistics.

DARC again presented the linear regression with complex survey data session of the Statistical Modelling Course at Statistics Canada, in English and French, in addition to the logistic regression session of the same course, in French only. DARC also gave the seminar for recruits on analysis of data from a complex survey.

Collaboration

DARC collaborated in developing measurement strategies for the Workplace Mental Health Performance Measurement Project with the Treasury Board Secretariat (TBS). This project used data from the 2019 and 2020 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 launched on 2022-05-17: Federal Public Service Workplace Mental Health Strategy. The measurement models were developed using factor analysis and structural equation modelling as discussed by Blais, Mach, Michaud and Simard (2020) and Blais, Michaud, Simard, Mach and Houle (2021).

For further information, please contact:
Fritz Pierre (613-720-4318, fritz.pierre@statcan.gc.ca) or
Isabelle Michaud (613-314-8971, 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 Data Ethics Secretariat

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

Progress:

In addition to conducting roughly 150 ethical reviews, members of the Data Ethics Secretariat have given numerous presentations to inform internal partners, colleagues from other federal departments as well as from international organizations on Statistics Canada's approach on data ethics. Two short introduction videos on data ethics were produced as part of the Data Literacy Training Initiative. Documentation on the foundations of ethical reviews in the statistical context outlining six guiding principles was completed and is available internally. This document is expected to be available on Statistics Canada's website later in 2023. Finally, the team gathers information to remain up to date on topics perceived as sensitive by the public. This is done by conducting literature reviews on some targeted topics, informal discussions with internal partners such as Communications and the Questionnaire Design Resource Centre or counterparts from other federal departments or foreign National Statistical Offices.

For more information, please contact:
Martin Beaulieu (613-854-2406, martin-j.beaulieu@statcan.gc.ca).

5.6 Quality Secretariat

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

PROJECT: Capacity building with internal, national and international partners

The Quality Secretariat's objective is to provide advice and undertake capacity-building measures internally, with national partners (other departments or other organizations) and international partners,

primarily by giving a general overview of Statistics Canada's quality management practices and official quality-related documents (the Quality Assurance Framework and the Quality Guidelines) and by providing quality management support services.

Progress:

The Quality Secretariat undertook capacity building for many partners during the reporting period. Internally, training was offered through various courses for staff. At the national partner level, formal presentations on quality management practices were made to two organizations, in addition to holding workshops and seminars. Materials on data quality and good quality management practices were provided to Statistics Canada's Data Literacy Training Initiative. Discussions occurred within the Government of Canada Enterprise Data Community of Practice Data Quality Working Group. This working group, co-chaired by Statistics Canada, aims to define a data quality framework applicable to all Government of Canada organizations as part of the implementation of the Data Strategy. A draft Data Quality Framework is available for partners from other federal departments and formal approval is pending. At the international level, involvement with the United Nations Expert Group on National Quality Assurance Frameworks continued in preparation for the implementation of the United Nations National Quality Assurance Framework Manual for Official Statistics (United Nations, 2019). A presentation on Statistics Canada's practices was given as part of the UN Expert Group on National Quality Assurance Frameworks' global seminar on building a culture of quality in national statistical offices.

For more information, please contact:
Martin Beaulieu (613-854-2406, martin-j.beaulieu@statcan.gc.ca).

Reference

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

PROJECT: Quality indicators for statistics from integrated data

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

Progress:

These indicators were used for additional tables of the Canadian Housing Statistics Program. The code was improved to be more efficient and transferred to the team responsible for CHSP. The Quality Secretariat keeps providing support on the code and the method. This project was presented at the Federal Committee on Statistical Methodology 2022 Research and Policy Conference in October 2022.

For more information, please contact:
Martin Beaulieu (613-854-2406, martin-j.beaulieu@statcan.gc.ca).

Reference

Gagnon, R., Qian, W., Yeung, A., Lebrasseur, D. and Beaulieu, M. (2022). Development of a Composite Quality indicator for Statistical Products Derived from Administrative Sources. Statistics Canada, https://www150.statcan.gc.ca/n1/pub/46-28-0001/2022001/article/00001-eng.htm.

PROJECT: Quality Assurance Framework Update

The Quality Secretariat initiated a review of Statistics Canada's Quality Assurance Framework (QAF). The current version was released in 2017. While the content of the current version is still valid, the fast evolution of new data sources and new techniques used in the production of official statistics made this review relevant. The updated version will highlight the importance of data stewardship, data ethics principles and some considerations relative to new techniques used. The plan for the update was presented to the Advisory Committee on Statistical Methods in the Fall 2022 (Beaulieu, Yung and Rancourt, 2022). The new version is expected to be released early in 2024.

For more information, please contact:
Martin Beaulieu (613-854-2406, martin-j.beaulieu@statcan.gc.ca).

Reference

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

5.7 Quality Assurance Resource Centre

The objective of the Quality Assurance Resource Centre (QARC) is to conduct research and development activities on statistical methods of quality assurance and control with the goal to improve the outgoing quality of survey data collection and processing operations within the bureau. This includes offering methodological services for G-Code which is used at Statistics Canada to create coding databases for data processing. Research on quality assurance and control is often generic in nature and involves issues of efficiencies and automation that are frequently applied to many steps of survey operations.

Progress:

The methodological support team helped the G-Code development team and tracked user inputs to help identify ideas for potential improvements for G-Code. The QARC also provided internal and external G-Code users with support when help/comments/suggestions regarding G-Code was needed.

During the year, work revolved around the implementation of a new version of G-Code (Version 3.3), which included the addition of machine learning capabilities (XgBoost, FastText and Pytorch). More specifically, the QARC team has been involved in a coding and classification proof of concept looking at the integration of the FastText algorithm into G-CODE. The new algorithms have been widely used to code various classifications for the Labour Force Survey (LFS), Job Vacancy and Wage Survey (JVWS), Canadian Community Health Survey (CCHS), Building Permits Survey (BPER), Postsecondary Student Information System (PSIS), Statistical Business Register (SBR) and the Census of Population. Additionally, these new

functionalities have been presented to external agencies (United Kingdom's Office for National Statistics (ONS) and United Nations Economic Commission for Europe (UNECE)) and internally through courses and project-oriented demos.

Lately, the QARC team has been helping with the integration of Pytorch into G-Code. PyTorch is an open-source machine learning library based on the Torch library, used for applications such as computer vision and natural language processing, primarily developed by Facebook's AI Research lab.

The QARC also worked on implementing a variety of quality controls on the coding processes for the LFS, CCHS, PSIS, BPER, JVWS, SBR and the new Statistical Building Register (SBgR). The QARC team has started working on the "QC (Quality Control) by Score" for machine learning (ML) text coding, which will determine quality control sampling rates in relation to the score provided by machine learning algorithm (i.e., fastText, XgBoost, PyTorch). The QARC team has also investigated a calibration methodology to insure a more coherent relationship between machine learning scores and accuracy of the coded data.

A paper was written for Statistics Canada's Symposium (Evans and Oyarzun, 2021) that explains the integration of a FastText model and a quality control process to the coding activities in the LFS.

For more information, please contact:
Javier Oyarzun (613-302-8454, javier.oyarzun@statcan.gc.ca).

Reference

Evans, J., and Oyarzun, J. (2021). Need for speed: Using fastText (Machine Learning) to code the Labour Force Survey. Proceedings: *Symposium 2021*, *Adopting Data Science in Official Statistics to Meet Society's Emerging Needs,* Statistics Canada, Ottawa.

5.8 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 the Heritage Canada, Public Safety Canada, Telefilm Canada, The City of Edmonton, Public Services and Procurement Canada, Global Affairs Canada and others.

The group also contributed to various corporate consultation initiatives.

For further information, please contact:
Paul Kelly (613-371-1489, paul.kelly@statcan.gc.ca).

5.9 Confidentiality

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.

PROJECT: Differential privacy support

The agency continued to look for ways to innovate and modernize its access solutions. One topic of interest to the agency in the realm of disclosure control is the adoption of differential privacy as a disclosure control framework. This topic warrants further investigation in the coming years.

Progress:

In the fiscal year 2022-2023, a differentially private noise mechanism was applied to count-based data from the Canadian Vital Statistics Death Database (CVSD). Results were then provided to clients from the Canadian Centre of Substance Use and Addiction (CCSA) to support policy research related to substance use in Canada.

PROJECT: De-identification

The confidentiality support group continued to offer its expertise in the understanding and development ideas related to de-identification and anonymization. Most recently Statistics Canada has contributed to the "Privacy Implementation Notice 2023-01: De-identification" developed by the Treasury Board of Canada Secretariat (Privacy Implementation Notice 2023 01: De identification - Canada.ca).

PROJECT: Random Tabular Adjustment

Statistics Canada continued to look for opportunities to apply innovative methods to allow for more information to be available to Canadians. Previous research has led to the development of the Random Tabular Adjustment method (Random Tabular Adjustment is here! (statcan.gc.ca)). The confidentiality support group continued to support the application of these ideas on new programs including the 2021 Census of Agriculture, the 2023 Annual Survey of Research and Development in Canadian Industry (RDCI), and the 2023 Livestock Survey.

PROJECT: External consultation

Statistics Canada has offered its expertise to several groups both domestically and internationally. Internationally, Statistics Canada has helped to develop the UNECE "*Synthetic Data for Official Statistics: A Starter Guide*". We have also been collaborating with the CBS (Statistics Netherlands) on some of the work they have been doing to familiarize their research community with synthetic data. Domestically, we

have given workshops and consulted groups such as the North American Association of Central Cancer Registries (NAACCR), Institut de la statistique du Québec (ISQ), Treasury Board of Canada Secretariat (TBS), the Bank of Canada, Public Health Agency of Canada (PHAC), Elections Canada, and Health Canada.

For more information, please contact:
Steven Thomas (613-882-0851, steven.thomas@statcan.gc.ca).

5.10 Data Science Communities of Practice

PROJECT: Applied Machine Learning Text Analysis Community of Practice

The Machine Learning Text Analysis Community of Practice (CoP) is a centralized inter-departmental place for practitioners of various expertise to discuss practical applications of Natural Language Processing (NLP) within the Government of Canada (GoC). Various practitioners across 25 federal departments come together each month to learn, discuss and adopt ethical applications of NLP. Monthly meetings bring 100-130 attendees to share each other's solutions and problems. About three-quarters of the participants are from federal departments outside Statistics Canada.

Progress:

Throughout 2022-2023, eleven practitioners from various Statistics Canada's fields and seven teams from departments, including the Public Service Commission of Canada, Canada Mortgage and Housing Corporation, Global Affairs Canada, Employment and Social Development Canada, Justice Canada, Social Sciences and Humanities Research Council of Canada, and Canada Revenue Agency presented their high-quality NLP solutions. Presenters illustrated and demonstrated their modern methodologies to quickly process their data source whether that be survey data, administrative data and public reports. Discussions after the presentation broke down the complex concepts for attendees to comprehend. Around 200 members were from Statistics Canada and 280 members were from other federal departments.

For more information, please contact:
Joanne Yoon (343-542-5625, joanne.yoon@statcan.gc.ca).

6 Other activities

6.1 Survey Methodology Journal

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

Progress:

The June and December 2022 issues (48-1 and 48-2) were released. The June 2022 issue contains ten papers. Six papers were published in the December 2022 issue, which featured the 2022 Waksberg paper by Roderick Little entitled "Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference", as well as a special invited paper by Changbao Wu entitled "Statistical inference with non-probability survey samples". The latter includes five discussions by eminent survey statisticians and a rejoinder.

In 2022, 55 papers were submitted to the journal. The average number of days from submission to initial decision was 52. All submitted papers were reviewed within 120 days, except for one paper that was exceptionally reviewed in 170 days and another one in 144 days, and 75% of them were reviewed within 90 days. Among those 55 papers, 28 were rejected, 16 were accepted and 11 had not received a final decision (including papers that were not revised by the authors before the deadline) at the end of May 2023. From April 2022 to March 2023, the *Survey Methodology* pages were viewed 57,383 times (compared with 45,071 views for the previous year).

We are currently planning special discussion papers and special issues in forthcoming releases. For instance, in the June 2023 issue, a special paper by Natalie Shlomo on statistical disclosure control and privacy will be published to honour the memory of Chris Skinner, a giant in survey statistics, who passed away in 2020. Shlomo's paper will be accompanied with testimonials from Danny Pfeffermann, J.N.K. Rao and Jae-Kwang Kim. Another special paper, by Pascal Ardilly, David Haziza, Pierre Lavallée and Yves Tillé, is being planned for the December 2023 issue to honour the memory of another giant in survey statistics, Jean-Claude Deville, who passed away in 2021. The paper will review the most important of his contributions to the field, which include among others, calibration and cube sampling. It will be followed by discussions/testimonials from colleagues and friends. The December 2023 issue will also feature the 2023 Waksberg paper by Raymond Chambers, and a special section with a few selected papers presented at the 2021 Colloque francophone sur les sondages. The Guest Editor for this special section is Alina Matei. In 2024, a special issue is planned for three papers that were presented at the 2022 Morris Hansen Lecture event on the use of non-probability samples. All three papers will be discussed by international experts in the field. An introduction by Partha Lahiri, the Guest Editor for this special issue, will precede the papers. A special discussion paper by Carl-Erik Särndal, entitled "Progress in survey science: yesterday – today – tomorrow", is also currently being planned for publication in a future issue in 2024 or 2025 along with discussions from eminent survey statisticians. Finally, the June 2025 issue will be dedicated to celebrate the 50th anniversary of *Survey Methodology*.

For more information, please contact:
Jean-François Beaumont (613-863-9024, jean-francois.beaumont@statcan.gc.ca).

6.2 Knowledge Transfer – Statistical Training

The Working Group on Statistical Talent Development, whose primary mandate remains statistical training within the Branch and the organization, has had another busy and productive year. Our course program is fully adapted to training in a virtual setting. Several courses were offered this year, including those related to time series, questionnaire design, sampling, the CANCEIS edit and imputation system, imputation, weighting, modeling and data ethics.

In terms of new activities, the group continues to design and prioritize learning activities that can be developed in a timely manner and focus on active learning. This year, a new Rao-Wu bootstrap course was offered for the first time. This method is used to estimate the sampling variance for several Statistics Canada social surveys. The course was offered in both official languages and more sessions are planned for next year. We also worked on revamping courses on analysis with survey data. The goal is to offer both courses on this topic in the coming year. A course on the use of macros with the SAS software has also been developed and will be offered in the coming year. As data science remains one of the priority areas within the organization, three courses on the subject have been developed.

- A course on machine learning modeling for classification was developed earlier this year. This course introduces a number of approaches, such as XGBoost, FastText, and transformer (self-attentive) models. Participants are invited to view a series of videos, which are followed by an in-depth discussion with the group and a facilitator, and finally a demonstration of the application of the method in an existing Statistics Canada program. The course was first given in June 2022.
- A course on introduction to machine learning was also developed over the past year. This training presents the main methods of supervised learning and applications using the Python programming language. A pilot version of the course was offered in June 2023.
- A course on the responsible use of machine learning methods has been designed. This course currently consists of two modules, one on equity and one on the introduction to explainable artificial intelligence. The idea of this training is to make employees aware of the risks associated with the use of machine learning. Pilot versions of the two modules were offered in June 2023. We plan to create a third module on bias in data for the next year.

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

For more information, please contact:
Keven Bosa (613-863-8964, keven.bosa@statcan.gc.ca).

6.3 Statistics Canada's International Methodology Symposium

Statistics Canada's 2022 International Methodology Symposium "Data Disaggregation: Building a more-representative data portrait of society" was the second full Symposium presented virtually. Previously held in the National Capital Region of Canada, Symposium 2022 was accessible online November 3rd and 4th, 2022. The Symposium was again provided free of charge to participants, and included plenary, parallel and poster sessions that covered a wide variety of topics and was preceded by three concurrent workshops on Wednesday, November 2nd.

Progress:

Following on the success of last year's first-ever virtual Symposium, this year's Organizing Committee coordinated the significant efforts of the program committee with the timely implementation of their vision by the Logistics Committee in partnership with Statistics Canada's Conference Services. Following a general call for papers in July, the program committee selected the contributions for presentations and for posters, found organizers for the invited sessions, organized workshops, then filled and finalized the

program. The Logistics Committee arranged to have the 2022 version of the website set up, and all our 900 registrants made use of the updated online abstract submission form.

The Symposium's keynote speaker was Grace Sanico Steffan from the United Nations Office of the High Commissioner for Human Rights, who shared her views on Breaking the Cycle of Invisibility in Data. The 19th Waksberg Award winner, Roderick J. Little from the University of Michigan, presented how Bayes, buttressed by design-based ideas, is the best overarching paradigm for sample survey inference. Numerous other presentations rounded out the event, which included three parallel sessions (one invited and two contributed) when not in plenary.

The translation of papers for the proceedings had to be delayed until the next reporting period, but these will eventually be reviewed and prepared for electronic publication – and are expected after the 2021 Symposium proceedings are released in the summer of 2023.

For more information, please visit:
https://www.statcan.gc.ca/en/conferences/symposium2022/index.

7 Research papers sponsored by the Methodology Research and Development Program

Abado, M., and Stefan, G. (2022). Reconstruction attack risk using Statistics Canada census data. Proceedings: *Symposium 2022*, *Data Disaggregation: Building a more-representative data portrait of society*, Statistics Canada, Ottawa (to appear).

Baillargeon, J. (2022). Bidirectional Pro-rating. Internal document, Statistics Canada.

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

Beaumont, J.-F., Bosa, K., Brennan, A., Charlebois, J. and Chu, K. (2023). Handling non-probability samples through inverse probability weighting with an application to Statistics Canada's crowdsourcing data. *Survey Methodology* (accepted in 2023 and expected to appear in 2024).

Bosa, K., and Beaumont, J.-F. (2023). How to allocate the sample to maximize benefits from small area estimation techniques? Presentation at the Statistical Society of Canada Annual Meeting, May 2023.

Charlebois, J. (2023). Nielsen Homescan Spending Data: Weighting by inverse propensity modelling of "frequent participation" in the Panel. Internal report, Social Statistics Methods Division, Statistics Canada.

Chen, W. (2022). Optimal feature extraction for probabilistic record linkage with model-based trees. Internal report, Statistics Canada.

Dasylva, A., Beaumont, J.-F., Bosa, K. and Maranda, G. (2023). Measuring the accuracy of a prediction for a finite population total. Presentation at the Annual Conference of the Statistical Society of Canada, May 2023.

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

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

Dasylva, A., and Goussanou, A. (2022b). A new model for the automated identification of duplicate records. In *Proceedings of the Survey Methods section*, Statistical Society of Canada. Available at https://ssc.ca/sites/default/files/imce/dasylva_ssc2022.pdf.

Dasylva, A., Goussanou, A. and Nambeu, C.-O. (2023). Measuring the coverage of two data sources through capture-recapture with linkage errors. Internal report, Statistics Canada.

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

Istrate, A., and Mashhadi, S. (2023). Anonymization of Training Text Data and its Effect on the Performance of NLP Models. Internal report, Statistics Canada, Ottawa.

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

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

Mandava, C., and Widhani, N. (2023). Reducing data gaps for training machine learning algorithms using a generalized crowdsourcing application. Statistics Canada. https://www.statcan.gc.ca/en/data-science/network/reducing-data-gaps.

Matthews, S. (2022a). Estimating the Trend-Cycle in Topsy-Turvy Times. Seasonal Adjustment Practitioners Workshop, United States Census Bureau.

Matthews, S. (2022b). Estimating the Trend-Cycle in Topsy-Turvy Times. 2nd Workshop on Time Series Methods for Official Statistics, Eurostat.

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

McGrouther, S., and Baillargeon, J. (2022). Stratification via Hierarchical Clustering. Internal presentation, Statistics Canada.

Medous, E., Goga, C., Ruiz-Gazen, A., Beaumont, J.-F., Dessertaine, A. and Puech, P. (2023). QR prediction for statistical data integration. *Survey Methodology*, 49 (to appear).

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

Pearce, A., Sallier, K. and Laperrière, C. (2023). Guiding Principles: Using the 2021 Census of Populations Data to Produce Statistics on DDAP Groups of Interest. Internal Document, May 2023, Statistics Canada.

Rao, J.N.K., Estevao, V., Beaumont, J.-F. and Bosa, K. (2023). Relative efficiency of area-level and unit-level small-area estimators when unit level auxiliary data is available. Presentation at the Statistical Society of Canada Annual Meeting, May 2023, Ottawa, Canada.

Sivathayalan, A., Chu, K. and Le Moullec, J. (2023). A new method to choose the number of clusters of a mixed dataset under Kproto clustering, Part I: Introduction. Working paper, Statistics Canada.

Sivathayalan, A., and Le Moullec, J. (2023). A new method to choose number of clusters of a mixed dataset under Kproto clustering, Part II: Evaluation. Working paper, Statistics Canada.

Sohrabi, M., and Rao, J.N.K. (2023). Estimation of the poverty measures for small areas under a two-fold nested error linear regression model: Comparison of two methods. Draft manuscript submitted to a peer-reviewed statistical journal. arXiv:2306.04907 [stat].

Statistics Canada (2022a). Workshops, training and references. https://www.statcan.gc.ca/eng/wtc/training.

Statistics Canada (2022b). G-Est 2.03.004 Release Notes. Internal document, Statistics Canada.

Statistics Canada (2023). Generalized Systems Report on Achievements and Objectives. Internal document, Statistics Canada.

Stinner, M. (2023). G-Sam 1.04 Notes on Allocation. Internal document, Statistics Canada.

Toupin, M.-H., and Martin, V. (2023). Degrés de liberté liés à l'estimation du questionnaire détaillé du recensement. Internal report (will be submitted to a scientific journal), Statistics Canada.

United Nations (2023). Synthetic Data for Official Statistics: A Starter Guide | UNECE. Geneva: United Nations. Available at https://unece.org/statistics/publications/synthetic-data-official-statistics-starter-guide.

United Nations Economic Commission for Europe (2023). *UNECE Project on Input Privacy Preservation: Final Report*. https://statswiki.unece.org/x/mQCQFw.

Wang-Lin, A. (2023). Literature Review of Fairness in Machine Learning. Unpublished internal Statistics Canada document.

You, Y. (2023). An empirical study of hierarchical Bayes small area estimators using different priors for model variances. *Statistics in Transition* new series, to appear.

You, Y., DaSylva, A. and Beaumont, J.-F. (2023). An approximate Bayesian approach to estimation of population means by integrating data from probability and non-probability samples. Draft manuscript to be submitted to a peer-reviewed statistical journal.

You, Y., and Hidiroglou, M. (2022). Application of sampling variance smoothing methods for small area proportion estimation. Proceedings: *Symposium 2022*, *Data Disaggregation: Building a more-representative data portrait of society*, Statistics Canada, Ottawa, Canada (to appear).

You, Y., and Hidiroglou, M. (2023). Application of sampling variance smoothing methods for small area proportion estimation. *Journal of Official Statistics*, to appear in the December issue 2023.