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

Statistical Methodology Research and Development Program Achievements, 2019/2020

Release date: September 29, 2020

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 STATCAN.infostats-infostats.STATCAN@canada.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

Depository Services Program

- Inquiries line 1-800-635-7943
- Fax line 1-800-565-7757

Standards of service to the public

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

Note of appreciation

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

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

This report summarizes the 2019/2020 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 survey programs; these activities would not otherwise be carried out during the provision of methodology services to those survey 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, contact:

Susie Fortier
(613-220-1948; susie.fortier@canada.ca)

Statistical Methodology Research and Development Program

Annual report 2019/2020

Contents

5. Divisional research and other activities

6. Research papers sponsored by the Methodology Research and Development Program

1. Modeling, data integration and data science

1.1 Small Area Estimation

Standard design-based estimates of population parameters, called direct estimates, are generally reliable provided that the sample sizes in the domains of interest are not too small. Indirect estimates, that borrow strength over areas or over time, often yield substantial gains of efficiency for small domains at the expense of introducing model assumptions. In recent years, there has been a renewed interest at Statistics Canada in investigating indirect model-based estimation methods for small domains and a system has been developed. The system and methods are documented in Hidiroglou, Beaumont and Yung (2019). The ultimate objective is to implement such methods for the production of official statistics, when judged appropriate. The main goals of this project are:

i) to develop new estimation methods for small domains that address issues found in real surveys;
ii) to study properties of existing methods under different scenarios to better understand how and when to use them;
iii) to determine suitable small area estimation methodology for some candidate surveys;
iv) To develop and test prototypes implementing new or existing methods that could be beneficial to statistical programs.

So far, progress has been made in the following sub-projects. They are described below.

SUB-PROJECT: EBLUP and HB modeling for LFS small area estimation with sampling variance smoothing vs modeling

The goal of this project is to study and evaluate the Fay-Herriot (FH) model with different sampling variance smoothing and modeling methods for the estimation labor force characteristics. In particular, we consider the models of You and Chapman (2006) and you (2016) for the estimation of the unemployment rate and compare these model-based estimates with census estimates. We plan to determine an appropriate small area model with sampling variance smoothing or modeling for the estimation of the unemployment rate in the Labor Force Survey (LFS).

Progress:

We studied Empirical Best Linear Unbiased Prediction (EBLUP) and Hierarchical Bayes (HB) approaches for the estimation of the unemployment rate based on the Fay-Herriot model with

sampling variance smoothing and modeling. Model-based estimates have been obtained and compared with the census estimates at the Census Metropolitan Area/Census Area (CMA/CA) level with different sample sizes. The results have shown that the Generalized Variance Function (GVF) smoothing method of You and Hidiroglou (2012) is the most efficient approach for small area estimation using LFS data. HB sampling variance modeling based on log-linear models of You (2016) is also efficient in terms of bias and Coefficient of Variation (CV) reduction compared to the direct survey estimates. Using unsmoothed sampling variance with the EBLUP of Wang and Fuller (2003) performs poorly, particularly when the sample size is small. Our recommendation is that sampling variance smoothing is needed and is an important step for the estimation of the unemployment rate using the Fay-Herriot model, if the EBLUP approach is used. If the HB modeling approach is used, we recommend sampling variance modeling based on a log-linear GVF model. Models and results are summarized in a research report (You, 2020a).

SUB-PROJECT: Model-based estimation of LFS totals

In this project, we study the linear FH model and non-linear unmatched models for the estimation of LFS totals including the employment total. We plan to evaluate benchmarking and non-benchmarking estimates for the totals and compare them with the census estimates. This study will provide models for the total estimation and evaluate the use of log-linear unmatched models and possible benchmarking procedure for the model-based total estimates.

Progress:

Log-linear unmatched Fay-Herriot model (You and Rao, 2002) based on HB approach is applied to the LFS total data to estimate employment total, in-labour-force total and unemployment total at CMA/CA level. For total estimation, the linear Fay-Herriot model performs poorly when total estimates are directly modelled. Transformation to the rate or mean is needed if the Fay-Herriot model is used. For employment total estimation, unmatched log-linear model using HB approach performs very well and efficiently by reducing both the bias and CV. Both inverse gamma prior and log-linear model prior of You (2016) lead to similar results for the total estimation. Unmatched log-linear model performs slightly better than the transformation approach by having slightly smaller CV. Benchmarking does not necessarily reduce the bias, given the fact that the model is adequate for the data. Programs and summary reports have been written and completed (You, 2020b).

SUB-PROJECT: Estimation of the design Mean Square Error in Small Area Estimation

The use of the Fay-Herriot model to produce small area estimates has increased at Statistics Canada in the last years. These estimates are typically accompanied with estimates of their

model Mean Square Error (MSE). However, users are typically accustomed with estimates of the design MSE. The design MSE has the advantage over the model MSE of not integrating out the specificity of a particular domain, and may be more relevant to users as an indicator of the quality of the estimates. Design-based estimates of the design MSE are known to be unstable (e.g., Rao, Rubin-Bleuer and Estevao, 2018). We plan to investigate the use of a conditional approach to obtain a more efficient estimator of the design MSE.

Progress:

We developed an estimator of the design MSE by using a conditional approach. The theory has first been developed for the case of the best predictor, which assumes all the model parameters are known. It is expected to be more efficient than the design-unbiased estimator of the design MSE. The case of the best linear unbiased predictor is currently under investigation. For the case of the empirical best linear unbiased predictor, where all model parameters are estimated, we developed a bootstrap procedure for estimating the design MSE. Our conditional approach will be evaluated in a simulation study. A draft internal report has been written (Beaumont, Lesage and Rao, 2020).

SUB-PROJECT: Assessing the robustness of the Fay-Herriot model for small domains

It is well known that probabilistic surveys make it possible to obtain reliable estimates for population domains for which the sample size is large enough. The demand for estimates for increasingly fine-grained domains has increased in recent years. To meet this demand, without drastically increasing the overall sample size and the collection costs, small area estimation techniques can be used. These techniques are based on determining a model that links the survey data to auxiliary data. The Fay-Herriot model is the one most commonly used in practice. If the model is correctly specified, the approach is valid and can lead to significant increases in accuracy. What happens in a realistic scenario where the model is not perfectly suited? The objective of this project is to assess and quantify, theoretically and/or through simulations, the impacts of poor model specification on the bias and variance of small area estimates, especially for the smallest domains. Various types of poor model specification are assessed.

Progress:

A simulation study under various scenarios was undertaken to test the effects of non-linearity of the model mean. A main conclusion is that small area estimates remain more efficient than direct estimates under model misspecification, particularly for the smallest domains. The report by Buresi (2019) contains greater detail on these findings.

SUB-PROJECT: Local diagnostics for the Fay-Herriot model

Model validation tools such as graphs of residuals are often used to assess the plausibility of the Fay-Herriot model. Model-based Mean Square Error (MSE) estimates are then used to evaluate the efficiency gains of small area estimators over direct estimators. All these techniques are useful to assess the overall performance of small area estimates. However, users are often interested in their specific domain only and a quality indicator for their specific domain estimate is more relevant to them. The model-based MSE achieves partially this goal but integrates out the local random effect (linking model error) that is of interest to users of a specific domain. The design MSE would be more relevant to these users but design-unbiased estimates of the design MSE are known to be very unstable (e.g., Rao, Rubin-Bleuer and Estevao, 2018). The goal of this project is to develop and investigate new local diagnostics for the evaluation of small area estimates.

Progress:

In previous research, we developed two new local diagnostics. They both involve the standardized residual for each domain as well as the ratio of the linking model variance over the sampling model variance. In the last year, we completed writing a paper that was submitted to Survey Methodology (Lesage, Beaumont and Bocci, 2020).

SUB-PROJECT: Small area estimation of health indicators for neighborhoods in Ontario

This project examined the potential of small area estimation techniques for estimating 19 health indicators (proportions) of interest for 147 neighborhoods in Ontario using the Canadian Community Health Survey (CCHS) annual data. The survey is not designed to produce reliable estimates of all these proportions at this level. Survey estimates were modelled as a function of Census variables to produce small area estimates using the Fay-Herriot methodology.

Progress:

Small area estimates for all 19 proportions were produced. Internal documentation was written, which includes the general strategy and a summary of results. The small area estimates had sometimes substantially smaller mean square error estimates than the direct CCHS estimates.

SUB-PROJECT: Small area estimation of labour characteristics by self-contained labour areas

Direct estimates of unemployment rates and employment totals can be calculated from the Labour Force Survey (LFS) at various levels of aggregation for a given time period. Two

requests were made to investigate the feasibility of producing small area estimates at the fine level of Self-contained Labour Areas for which the direct LFS estimates are typically not reliable. We used the Fay-Herriot area-level model to produce annual and monthly estimates of unemployment rates and employment totals by modelling the LFS estimates using Demography estimates and Employment Insurance data as auxiliary information.

Progress:

Small area estimates of annual unemployment rates and employment totals by self-contained labour areas were produced for the years 2011 to 2016 and forwarded to analysts for examination.

Small area estimates of monthly unemployment rates for the months 201904 to 201911 were produced for each self-contained labour area. Two documents, Bocci and Beaumont (2019a, 2019b), describe the methodology used to produce these small area estimates and show diagnostics, quality measures and comparisons. Work on small area estimation of total employment continues.

SUB-PROJECT: Small area estimation course

This project was to construct a course covering basic principles of small area estimation and diagnostics, as well as a software system developed by Statistics Canada to produce small area estimates in practice.

Progress:

The course was completed and presented in English. It was translated into French. After completing the revision of the translation, the French version will be ready to be offered.

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

References

Rao, J.N.K., Rubin-Bleuer, S. and Estevao, V.M. (2018). Measuring uncertainty associated with model-based small area estimators. *Survey Methodology*, 44, 2, 151-166. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2018002/article/54958-eng.pdf.

Wang, J., and Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.

You, Y. (2016). Hierarchical Bayes sampling variance modeling for small area estimation based on area level models with applications. Methodology Branch research paper, ICCSMD-2016-03-E.

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.

You, Y., and Hidiroglou, M. (2012). Sampling variance smoothing methods for small area proportion estimators. Methodology Branch Working Paper, SRID-2012-08E, Statistics Canada, Ottawa, Canada.

You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3-15.

1.2 Real-Time Estimation via Time Series Modeling

Other projects related to time series are described in section 5.6.

SUB-PROJECT: Time Series Modelling and progress towards Real-Time Estimation

Time series modelling (nowcasting) is seen as a potential means to produce advance indicators for the agency's economic indicators. Exploration was made towards models to produce early indicators, and to look at the possibility of producing real-time indicators that could frequently be updated to reflect the most current information available.

Progress:

Literature reviews were conducted to identify models commonly used in nowcasting (ARIMA models and Unobserved Component Models), develop knowledge of the methods and available tools for their application. Further work was also done to investigate alternative models available in R (TBATS, robust ARIMA as well as exponential smoothing variants for forecasting) and those based on machine learning. A draft version of practical guidelines for developing nowcasts was created (Matthews, Patak, Picard and Mischler (2020)). This document includes an inventory of available families of time series models along with advantages and disadvantages in the context of nowcasting, and a discussion of producing nowcasts in the case of a national statistical organisation.

A number of criteria are proposed to determine if advance estimator should be produced, in the context of official statistics. A number of proof of concept exercises were completed to assess the feasibility of these methods on specific projects with varying circumstances. The team evaluated time series modelling and nowcasting methods on monthly Retail Trade data (Matthews and Patak, 2020). This evaluation included the use of predictors from leading indicators in the form of statistical outputs, secondary data, or partial survey responses to quantify trade-off between precision and timeliness for models based on auxiliary information sources. Results from this work were presented to the analysts from different areas within Statistics Canada, leading to applications of the methods in other contexts. For example, modelling was applied for other indicators, nowcasting was applied to mixed frequency data to generate and update nowcasts, results were compared with machine learning techniques, and use of sentiment indicators as an auxiliary variable was investigated.

The team continued to expand the use of SAS High Performance Forecasting and forecasting tools for general use, as well as establishing a stable processing platform and improving understanding of the methods. A number of features of the model selection process were investigated leading to some feedback to SAS to point out minor inconsistencies and clarify

technical issues around the impact of events on degrees of freedom. The software continues to be used to support modelling of critical non-respondents, identification of breaks in series, and other modelling applications.

To allow for efficient and intuitive identification of seasonal patterns for forecasting models progress towards a test for seasonality with flexible periodicity (non-parametric) to identify seasonal patterns in daily, weekly, monthly data was made. Both a re-sampling approach and a multiple comparisons approach were considered. Investigation into a non-parametric seasonality test based on re-sampling to generate an empirical distribution function was conducted and documented (Lapointe and Mischler, 2019).

For further information, please contact:
Steve Matthews (613-854-3174, steve.matthews@canada.ca).

1.3 Data Science – Machine Learning

SUB-PROJECT: Modeling opioid

Opioid related overdoses and deaths are currently at crisis levels in Canada. There is a need to understand the broader circumstances of those experiencing opioid related events to help inform the development of policies to address the upstream factors To address this information need, Statistics Canada, in partnership with several BC-based organizations (i.e., British-Columbia Opioid Steering Committee) has created a comprehensive linked data source that brings together administrative data to better understand the socio-economic determinants of those experiencing adverse opioid events (i.e., deaths and overdoses). The goal of this project is to identify Artificial Intelligence and Machine Learning techniques to support predictive and trajectory analyses using big structured data focussing on adverse opioid events as a case study.

Progress:

A heterogeneity analysis of the study cohort has been performed, via a cluster analysis based on 15 discrete variables and 2 continuous variables. Since the variables are a mixture of discrete and continuous ones, the *k* prototype method by Huang (1998) was used for the cluster analysis. In order to assess the appropriate number of clusters, we performed nine independent rounds of *k* prototype clustering (with 2 to 10 clusters) and diagnostics were used to assess optimality of the number of clusters as well as subject-matter expertise. In order to assess cluster stability (with respect to random initializations), we further performed 10 independent rounds of *k* prototype clustering. For each pair of clusters, we computed two similarity measures. Cluster stability was assessed via the resulting density plot and histogram of the two similarity scores. We verified that cluster stability analysis based on either score would yield the same conclusion as that based on the other, by generating the scatter plot of the two scores clustering pairs and noting that these two scores very closely follow a strictly increasing nonlinear transformation of each other. The resulting clusters exhibit distinct life event profiles that strongly corroborate with what has been reported in the public health literature. A manuscript is in preparation about the findings of the cluster analysis. A subsequent study and a subsequent study will take place to estimate the opioid event conditional risk by cluster, in the attempt to investigate the feasibility of predicting imminent opioid events based on individual recent life event history.

Reference

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

SUB-PROJECT: Predicting crop yield

Statistics Canada publishes annual crop yield (production per unit area of harvested land) estimates at the end of each reference year. In addition, full-year crop yield predictions are published several times during the latter half of the reference year. For the 2019 July predictions, a model-based method – essentially, variable selection via LASSO, followed by robust linear regression – was applied to phase out the crop yield question on the July survey for one province, namely Manitoba. The goal of this project is to examine whether certain machine learning techniques may produce better predictions than the current deployed method for the Manitoba/July 2019 yield prediction using up-to-July longitudinal measurements of parcel-level vegetation levels, local-level weather conditions, and parcel-level information on crop type, etc.

Progress:

In order to evaluate and find a better model for predicting yield than the currently deployed one, we must work under the following operational constraint: the trained model to be deployed in each given year could only have been trained on data from strictly earlier years. Hence, we have developed a forward validation protocol where, for a given year, data from that year were used as validation data, while data from years belonging to a fixed training window strictly prior to the given year were used as training data. This was repeated for multiple validation years. Data dating back to 2000 were available and we used data from 2000 to 2017 for training and validation, and data from 2018 and 2019 for testing. We applied several machine learning techniques including XGBoost, Support Vector Machine (SVM), random forest and glmnet. Grid search was used for hyperparameter tuning. We developed two metrics to compare results with the current approach: average production-weighted relative error and an average production-weighted standard deviation of these relative errors. Based on these measures we were able to show that the best-performing machine learning models indeed exhibit improvements over the current model. We are now in the process of implementing this new approach for production for the reference year 2021, with a parallel run in 2020.

SUB-PROJECT: Machine learning approach for statistical classification

We aim to improve the efficiency and accuracy of statistical classification, where text supplied by a respondent is categorized (or coded) to facilitate tabulation and inference. For example, text responses to questions about job title and main duties are regularly converted to an occupation class (4 digit numeric code) corresponding to the National Occupation Classification. Machine learning algorithms promise to outperform traditional manual coding,

where a person reviews the supplied text and selects a code, yielding comparable accuracy, but faster and at a lower cost.

Progress:

Automatic coding functions, based on machine learning and specifically fastText, were added to the G-code software. A number of trials and evaluations were performed with text inputs from surveys and past Censuses related to various classification systems (industry, occupation, major field of study, etc.)

SUB-PROJECT: Machine Learning projects for the High Level Group for the Modernization of Official Statistics (HLG-MOS)

The High Level Group for the Modernization of Official Statistics (HLG-MOS) Machine Learning project began in the spring of 2019 and consists of three work packages: 1) Proof of concepts; 2) Quality Measure and 3) Integration of Machine Learning (ML) in Organisations. Statistics Canada has been involved in all three work packages to varying extents. Our participation will be discussed in this report.

The goal of the proof of concepts is to share knowledge, experience and code (if applicable) amongst the groups involved in the project. The proof of concepts covers multiple different uses of ML such as classification, imputation, image processing and sentiment analysis. The goal of the quality measure work package is to develop a quality framework which will allow official statisticians to compare statistical algorithms, including machine learning ones. The framework will be based on existing concepts but will be tailored for statistical algorithms. Finally, the goal of the Integration work package is to identify best practices and to guide National Statistical Organisations new to machine learning how best to integrate it into their environment.

Progress:

As part of the proof of concept package, Statistics Canada employees have been involved in a coding and classification proof of concept looking at the integration of the FastText algorithm into our Generalized Coding tool (G-CODE). The new algorithm was used to code industry and occupation for two health surveys at Statistics Canada and results, experiences and code were shared with the HLG-MOS Machine Learning group. Methodology Branch employees were also involved in providing assistance or advice on two proofs of concept undertaken by other countries. The Office for National Statistics (ONS) proof of concept investigated the use of decision trees to develop edit rules for one of their household expense surveys. A similar study was undertaken within Statistics Canada’s Agriculture Taxation Data Program and the

methodologist involved provided some advice to the ONS. A proof of concept from Belgium investigated the use of machine learning to predict Energy Balances in order to produce more timely data. The time series section within the Methodology Branch has been working with the researchers in Belgium to see how machine learning methods compare to time series methods. Results have been obtained and are being shared with the researchers in Belgium.

A Branch employee is leading the work package on quality measures which will produce a quality framework for statistical algorithms. The framework will consist of the following five dimensions: interpretability, accuracy, timeliness, cost effectiveness and reproducibility. A final version of the framework is expected to be available in the fall of 2020.

For the integration package, a small face-to-face sprint was organized at Statistics Canada in October 2019. The participants of the sprint were the three work package leads and the project manager. During the sprint, the participants had the opportunity to meet the Chief Statistician and Assistant Chief Statistician, Stéphane Dufour, who co-chairs the HLG-MOS executive board. The main output of the sprint was a roadmap for work package 3. The Branch employee leading work package 2 was the principle organiser of the sprint and participated as well.

For further information, please contact:
Yanick Beaucage (613-854-2397, yanick.beaucage@canada.ca).

1.4 Data Integration - Record Linkage

Record linkage plays an important role in the production of official statistics. However, it is susceptible to errors because it is often based on quasi-identifiers that are not unique and recorded with variations and typographical errors. This project looks at the production and use of linked data, including the accurate estimation of linkage errors. Other activities related to record linkage are reported in section 4.1.

SUB-PROJECT: Estimation of linkage errors in a privacy-preserving setting

Various methods have been developed for linking encrypted data, including methods using bloom filters (Kroll, Niedermeyer, Schnell and Steinmetzer, 2014). However, a major challenge has been the accurate estimation of linkage errors and the related problem of setting the linkage parameters. Previous solutions have relied on training data that are often unavailable.

Progress:

This project has addressed the problem by demonstrating the accurate estimation of linkage errors in a privacy-preserving setting, without any training data. It has used a new error model based on the number of neighbours of a given record (Dasylva, Goussanou, Ajavon and Abousaleh, 2019). A finite population was created including names, birthdates, addresses, etc. Two related complete and duplicate-free registers were created and linked using bloom filters (Kroll et al., 2014), including blocking criteria based on the first few bits and comparisons based on the Dice index. The linkage parameters were the number of bits used for blocking and the Dice index threshold. The parameters were set to different values and for each setting the error rates were estimated and compared to the actual rates based on the actual match status. For each setting, the estimated rates were close to the actual rates. These results have shown that it is possible to accurately link encrypted data without any training data. The details are provided by He (2019).

SUB-PROJECT: Survival analysis with linked data

Linkage errors are a source of bias in analysis including survival analysis. Adjusting for these errors raises many challenges including the accurate estimation of errors and the adjustment method given the estimated rates.

Progress:

This project has considered the cohort mortality study based on the linkage between the Canadian Mortality Database and the Canadian Community Health Survey. It has experimented with two new methodologies for conducting a survival analysis with linked data,

including a new error estimation methodology (Dasylva et al., 2019) and a new adjustment methodology (Dasylva, 2018). A number of issues have been identified that must be addressed in future work. A first challenge is the run long time of the numerical procedure that is used to compute the adjusted survival parameters. The second challenge is to account for the Canadian Community Health Survey sample design when estimating the linkage errors. The details are provided by Miller (2019).

For further information, please contact:
Abel Dasilva (613-408-4850, abel.dasylva@canada.ca).

References

Dasylva, A. (2018). Pairwise Estimating Equations for the Analysis of Linked Data, PhD thesis, Carleton University, 2018.

Kroll, M., Niedermeyer, F., Schnell, R. and Steinmetzer, S. (2014). Cryptanalysis of basic bloom filters used for privacy preserving record linkage. *Journal of Privacy and Confidentiality*, 6, 59-79.

1.5 Mixed Methods and qualitative approach

SUB-PROJECT: International Engagement - A Mixed Methods Evaluation

Statistics Canada engages in international activities for various reasons. As a leading statistical agency, bringing leadership and technical expertise to the international community is an ongoing effort. These efforts are carefully monitored through the International Engagement Program.

Program monitoring to this point has focused on quantitative metrics as an evaluation tool; characteristics such as the number of participants, type of roles, event cost, travel cost, and participation by field are reported on quarterly basis. There has been however, growing interest in better understanding the value of international engagement beyond quantitative program statistics. As a result, this study applies mixed methods research to generate further insights. Mixed methods research draws on potential strengths of both qualitative and quantitative methods to analyze topics with complexity.

Progress:

Evaluation of the international engagement program is complete and results outline the strengths and challenges of the program. The project included cognitive interviews of program participants and quantitative data collection; and both were integrated with a mixed methods research approach. A post mortem was also completed in assessing the feasibility of applying mixed methods as a widespread approach at Statistics Canada.

SUB-PROJECT: A mixed-method exploration of measuring psychosocial factors of workplace mental health

The Workplace Mental Health Performance Measurement Project group, which is co-led by Public Services and Procurement Canada and Statistics Canada, in consultation with human resources professionals and organizational psychologists, undertook a concept mapping exercise collectively to map indicators to 13 psychosocial factors in federal organizations using various data sources, including the Public Service Employee Survey (PSES). The main objective of the study was to showcase a rigorous methodology to evaluate the proposed conceptual mapping between the items from the 2017 PSES and the 13 psychosocial factors.

Progress:

A mixed-method approach to social inquiry (Greene, 2007) was used with triangulation and complementarity. First, a concept mapping exercise (qualitative approach) was conducted to

determine how well the PSES items mapped into the 13 psychosocial factors as perceived by a select group of public service employees. This concept mapping exercise was different from the previous one undertaken by the Project group in that participants were asked to complete the exercise independently rather than collectively as a group. Agreement scores between raters were calculated. Second, exploratory and confirmatory factor analyses (quantitative approach) were conducted to examine whether the previously mapped PSES items appropriately measured the 13 psychosocial factors. Different criteria were considered to evaluate model fit. The findings from the concept mapping exercise and factor analyses were triangulated and complemented (mixed-method approach) to evaluate the conceptual mapping proposed by the Project group. In conclusion, further development of items as well as considerations of alternative data sources may be needed to better measure psychosocial factors.

A detailed report (Arim, Bougie, Michaud, Tabuchi, Yung and Kohen, 2019) was prepared for the Workplace Mental Health Performance Measurement Project Group.

For further information, please contact:
Laurie Reedman (613-894-2779, laurie.reedman@canada.ca).

Reference

Greene, J.C. (2007). Mixed Methods in Social Inquiry. San Francisco: Jossey-Bass.

2. Confidentiality

Confidentiality research at Statistics Canada continues 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. The Center for Confidentiality and Access group at Statistics Canada also continues to offer consultation services to internal and external partners as a way to help develop capacity in disclosure risk identification and treatment.

2.1 Random Tabular Adjustment

Random Tabular Adjustment is a disclosure control method that relies on adding random noise to estimates rather than suppressing them. The primary focus is to avoid suppression for continuous variables collected under economic surveys.

Progress:

At the end of 2018-19, Statistics Canada made an important inroad into offering this alternative strategy with the release of the Survey of Innovation in Business Strategy (https://www150.statcan.gc.ca/n1/daily-quotidien/190326/dq190326b-eng.htm). This marked the first occasion that the Random Tabular Adjustment (RTA) methodology was used.

Research for 2019-2020 built upon this success by developing the method to be applied on a wider variety of products at Statistics Canada as the agency moved towards zero suppression for its economic surveys. The main challenge being investigated was adding a correlated noise function. This would have several benefits including: maintaining correlation structures between related variables; minimizing the impact that noise addition has on aggregate cells by adding negative noise to related cells; and preserving trends found in repeated surveys. A paper is in progress that describes the new methodology. A new SAS tool is in development to apply the method. Training material is in development to develop a user base.

2.2 Synthetic Data

Continuing with the focus on offering new alternative access options, Statistics Canada is investing in researching methods for creating synthetic data. Synthetic data can take on a variety of forms and possess a variety of quality characteristics but the main focus is always to offer a microdata access option that poses little or no disclosure risk and therefore, can be released to the general public.

Progress:

The focus for 2019-2020 was to develop clear fundamentals on the terminology of data synthesis with the focus on creating synthetic data of high analytical value. The first challenge has been describing and differentiating this from synthetic "dummy" files. An update to a guidance document is in development.

Sample surveys pose special challenges such as dealing with sampling weights and ensuring proper analysis takes place. The challenges with creating synthetic data of high analytical value in this context were discussed at Meeting 68 of Statistics Canada's Advisory Committee on Statistical Methods held in April 2019.

On November 2, 2019, the Canadian Partnership against Cancer at The Hack4Cancer Hackathon, was held one day ahead of the Canadian Cancer Research Conference in Ottawa, Ontario. Statistics Canada developed the synthetic data for this session which was the second time data of this type was publicly released by the agency. Unique challenges including utility constraints were brought up and will be the focus for the future (https://cc-arcc.ca/hack4cancer_hackathon/).

The success of the work at Statistics Canada in this area of research was highlighted by Kenza Sallier winning the International Association for Official Statistics 2020 IAOS Prize for Young Statisticians for her paper "Toward More User-Centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis" (Sallier, 2020).

2.3 User Support and Consultation

Statistics Canada continues to support its internal and external partners in the area of confidentiality and access.

Progress:

The Research Data Center program allows trusted researchers to access detailed microdata in a secured environment. The research program supports this through user support and the development of vetting rules for analytical outputs and support for confidentiality concerns for specific projects.

The Generalized System G-Confid is Statistics Canada's generalized solution for confidentiality risk assessment and treatment for continuous variables usually found in the context of economic surveys. The research program supports implementation user requests for the system.

Statistics Canada ensures the anonymity of any microdata file through the support of its Microdata Release Committee. The committee is supported by the methodology branch and its team of confidentiality experts who ensure that the proper risk identification and mitigation strategies are in place for any public release of data. In 2019-20, approximately 20 different public data files have been released through this process.

During the period, a variety of consultations have taken place with domestic and international partners. International consultations include Denmark, United Kingdom, Australia, and the Caribbean. Statistics Canada also plays an active role with the United Nations Economic Commission for Europe and the Modernisation of Official Statistics Group. A paper highlighting the access and confidentiality challenges at Statistics Canada was presented at the 2019 Work Session on Statistical Data Confidentiality in October 2019.

Domestic consultations have included the Canadian Council of Cancer Registries, Public Health Agency of Canada, Health Canada, The International Association of Privacy Professionals (IAPP) Canada, Canadian Mortgage and Housing Corporation, CANON, BC Stats, Canada Council and the Tri-agency Institutional Programs Secretariat among others. Strategies to improve access were presented at the 2019 Joint Statistical Meetings in 2019.

For further information, please contact:
Steven Thomas (613-882-0825, steven.thomas@canada.ca).

3. Theory and framework

3.1 Theory and framework - Data Integration

Data integration: Combining data from probability and non-probability samples

The use of data collection or acquisition methods that do not rely on a probability sampling design has recently increased. For instance, there has been a growing use of crowdsourcing since the beginning of the COVID-19 pandemic. A crowdsourced sample can be defined as any sample of volunteers, typically obtained through the Internet. These samples are sometimes quite large but may yield estimates with significant selection biases. How to obtain meaningful estimates and make valid inferences from those large non-probability samples is an important question that still requires research and experimentations. Methods that attempt to address this question often combine data from those non-probability samples with data from a probability sample. This topic is called statistical data integration.

This project has three main objectives:

- to assess the possibility of using current data integration methods to obtain estimates with reduced bias obtained from non-probability samples;
- to develop or adapt new methods to solve practical issues;
- To develop and test prototypes that implement the most promising methods.

SUB-PROJECT: Using classification trees to weight a non-probability sample

One possible data integration method that may achieve bias reduction consists of modelling the probability of participating in the non-probability sample and weighting each participant by the inverse of its estimated probability. Chen, Li and Wu (2019) used a logistic function to model the participation probability assuming that the explanatory variables are given. A special case of this model is the homogeneous group's model, where the participation probability is assumed uniform within each group. We build on this work to extend the idea to classification trees (Breiman, Friedman, Stone and Olshen, 1984). The objective of classification trees in this context is to build a set of exhaustive and mutually exclusive groups that are homogeneous with respect to the participation probability and then use the method of Chen, Li and Wu (2019) to estimate these participation probabilities. Classification trees can be viewed as a nonparametric method of finding relevant explanatory variables and their interactions.

Progress:

In the previous year, we developed an R program that implements an algorithm to build a fully-grown tree. We evaluated the algorithm through a simulation study. Preliminary results indicate that the algorithm is promising for reducing the bias of estimates obtained through non-probability samples. The results were presented at the 2019 meeting of the Statistical Society of Canada (Chu and Beaumont, 2019).

Although our results indicated that classification trees could reduce selection bias, they also showed that the resulting estimates were somewhat inefficient. This may be explained by overfitting; i.e., the creation of too many groups. Pruning is usually recommended to avoid overfitting. It is usually done in two steps. First, a sequence of subtrees of decreasing size is determined starting from the fully-grown tree. Then, the best of these subtrees is chosen, typically using cross-validation. Cross-validation does not seem straightforward to implement in this data integration context. We spent some time developing an alternative method for subtree selection based on the Akaike Information Criterion. This was described in a paper to be presented at the Advisory Committee on Statistical Methods in June 2020 (Beaumont and Chu, 2020).

SUB-PROJECT: A Bayesian approach to survey estimation using probability and non-probability samples

A Bayesian method of data integration was recently published in the Journal of Official Statistics (Sakshaug, Wisniowski, Ruiz and Blom, 2019). This paper deals with the estimation of model parameters when the dependent variable *y* and the vector of explanatory variables *x* are observed in both a probability and non-probability sample. They use the non-probability sample as a means of obtaining a prior distribution under the assumption of a simple random sampling design. The goal of this project is to extend the method to the estimation of finite population parameters, with complex designs, and see if we can obtain model-based estimates more efficient than the standard survey-weighted estimates.

Progress:

We have read and studied the paper of Sakshaug et al. (2019). A summary report has been written (You, 2019). We plan to extend the work of Sakshaug et al. (2019) to develop a Bayesian inference approach based on a super population model to estimate population means and totals under simple random sampling as well as more complex designs, and borrow information from non-probability sample data as prior information under the Bayesian framework.

SUB-PROJECT: Statistical data integration using a prediction approach

We consider the problem where a non-probability sample is available that contains a vector of auxiliary variables, x, for each sample unit. We assume that this non-probability sample covers a significant portion of the population. A probability sample is also available that contains x as well as the variable of interest y for each sample unit. The indicator of participation in the non-probability sample is available in the probability sample. This scenario is relevant to a survey on postal traffic conducted by La Poste in France. Alain Dessertaine proposed a predictor in that scenario. The goal of this project is to study the properties of that predictor.

Progress:

We developed variance estimators, including a bootstrap variance estimator, for evaluating the quality of the proposed predictor. The details are given in an internal draft report. We plan to continue the collaboration with La Poste, Toulouse School of Economics and the University of Besançon and transform this draft into a joint paper for possible publication in a peer-reviewed journal.

SUB-PROJECT: Review of data integration methods

The goal of this project was to conduct an extensive literature review on statistical data integration methods. Design-based methods such as multiple frame methods and design-based calibration were reviewed as well as model-based methods like statistical matching, model-dependent calibration, inverse propensity score weighting and small area estimation.

Progress:

A paper was earlier written and submitted to Survey Methodology. We received reviewers' comments this year and revised the paper accordingly. The paper will be published in the June 2020 issue of Survey Methodology (Beaumont, 2020a). The model-based portion of the paper was also presented at the Italian Conference on Survey Methodology in June 2019 in Florence, Italy.

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

References

Breiman, L., Friedman, J.H., Stone, C.J. and Olshen, R.A. (1984). Classification and RegressionTrees. CRC Press.

Chen, Y., Li, P. and Wu, C. (2019). Doubly robust inference with non-probability survey samples. *Journal of the American Statistical Association* (published online).

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

3.2 Theory and framework - Quality

The Quality Secretariat, whose support activities are described in section 5.4, is also involved in development activities.

SUB-PROJECT: Quality measurement and reporting for statistical programs using integrated data

Statistics Canada, like many other national statistical agencies, is facing a paradigm shift in the production of its official statistics. This change is resulting in a transition from a traditional model with statistics produced using methods developed for probability sample survey designs and direct collection from respondents to a new model where the data are extracted from various secondary sources, such as administrative data or satellite data. These data are then combined or integrated to produce the desired statistics. In this new context, being able to properly measure and report data quality becomes a major challenge, given that the current methods and the terms used are closely associated with sampling theory.

Progress:

An overview and a work plan were developed based on the Rancourt (2018) article and were presented to the Scientific Review Committee and at the International Total Survey Error Workshop (Beaulieu, Chepita, Fortier, 2019). A working group was created to review the existing quality indicators and identify gaps in the various stages of producing official statistics. The work to develop a format for nutrition labels has begun and been discussed with some programs.

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

Reference

Rancourt, E. (2018). Admin-First as a Statistical Paradigm for Canadian Official Statistics: Meaning, Challenges and Opportunities. Presented at the 2018 International Symposium on Methodology Issues.

3.3 Framework – Responsible Machine Learning

Statistics Canada is using machine learning techniques to solve large scale data problems. Some of the research projects in this area is described in section 4.7. In parallel, National Statistical Offices are facing unprecedented pressure to demonstrate to citizens, businesses and users that they are trustworthy and transparent institutions. Therefore, a responsible approach is to develop a framework tailored to the use of machine learning techniques, including guidelines for building and implementing ethically and methodologically sound processes, and metrics for assessing and communicating the quality of machine learning processes and their products.

Progress:

The development of the framework and guidelines was done by a multi-disciplinary team and interim versions have been vetted by various internal communities (Data Science, Methodology, Senior Management (Quality, Methods and Standards Committee)) and external partners (such as Canada Post and the Advisory Committee on Statistical Methods). The framework is aligned with the Directive on Automated Decision-Making and its Algorithmic Impact Assessment tool, developed by the Treasury Board Secretariat (2020). It is based on 4 themes: Respect for people, Respect for data, Sound Application and Sound Methods and a number of attributes for each theme. The guidelines have not been officially adopted yet, but are being tested by managers of data science applications. A checklist accompanying the guidelines is also being tested and is used to help assess the responsible machine learning processes. Final adoption of the guidelines and checklist, development of a dashboard for monitoring and implementation of a formal review process are still to be done.

For further information, please contact:
Yanick Beaucage (613-854-2397, yanick.beaucage@canada.ca).

Reference

Treasury Board Secretariat (2020). Directive on Automated Decision-Making, Government of Canada website https://www.tbs-sct.gc.ca/pol/doc-eng.aspx?id=32592.

3.4 Framework – Necessity and Proportionality

In this information era, data sources and the information needs are constantly increasing and evolving. To address these mounting demands and, the equally important need to keep respondents' data safe, Statistics Canada is working in consultation with statistical and privacy experts from around the world to develop a new methodology framework based on the principles of necessity and proportionality.

Progress:

In October 2019, Statistics Canada adopted a new Necessity and Proportionality Framework to jointly maximize the production of information and privacy protection when developing data gathering approaches. This framework provides both a justification and a guide for designing strategies to gather sensitive data using surveys, administrative sources obtained from the public or private sector, or any other method. The approach is the result of consultations that were carried out with domestic and international experts and practitioners in the domain of official statistics, professional statistical associations, privacy experts, ethics experts, and the Office of the Privacy Commissioner. The Necessity and Proportionality Framework is an adaptation of the scientific approach (Rancourt, 2019) to the context of both statistical methodology and privacy protection. It rests on a solid description of why a given data source is needed and a thorough ethical assessment. To support this work, a secretariat operating under a Chief Ethics and Scientific Integrity Officer was set up. The secretariat's work supports the activities of an internal Data Ethics Committee. Statistics Canada integrated the Necessity and Proportionality Framework into its data acquisition process such that the ingestion of any new data source has to follow the framework. More information can be found at Statistics Canada's Trust Centre (https://www.statcan.gc.ca/eng/trust).

For further information, please contact:
Eric Rancourt (613-298-9403, eric.rancourt@canada.ca).

4. Support (Ressource Centre)

4.1 Record Linkage Ressource Centre

The objectives of the Record Linkage Resource Centre (RLRC) are to provide consulting services to both internal and external users of record linkage methods, including recommendations on software and methodology and collaborative work on record linkage applications, to evaluate alternative record linkage methods and develop improved methods. We evaluate software packages for record linkage and, where necessary, develop prototype versions of software incorporating methods not available in existing packages and assist in the dissemination of information concerning record linkage methods, software and applications to interested persons both within and outside Statistics Canada.

Progress:

The support team helped the development team and tracked user inputs to help identify ideas for potential improvements. The RLRC also provided internal and external G-Link users with support when help/comments/suggestions regarding G-Link were sought at G-Link info through JIRA tickets.

During the year, much of methodology's work revolved around the development of a new version of G-Link (Version 3.5) which included the addition of profile-based linkage (similar to deterministic linkage), identification and treatment of orphan records and integrated pseudo keys. Additionally, work has been done to integrate a clerical review tool (quality assessment) and correct and improve some threshold estimators.

RLRC also worked on a variety of other record linkage-related projects during the year, including holding two more instances of the Record Linkage Forum. Our record linkages helped us document performance and issues pertaining to management and developers and was used as an opportunity to field test new G-Link3.5 features and develop more systematic and theoretically coherent approaches of defining and adjusting record linkages under servers and SAS Grid. RLRC updated the tutorial of G-Link 3.5.

For further information, please contact:
Abdelnasser Saïdi (613-863-7863, abdelnasser.saidi@canada.ca).

4.2 Generalized Systems

Developmental Research – Generalized Systems

The Generalized Systems unit (GenSys) is responsible for research, development and support of the following systems:

- G-Est: The generalized estimation system;
- G-Sam: The generalized sampling system;
- Banff: The generalized edit and imputation system.

Aside from providing support and training related to generalized systems, the team also take on development research related to data visualization, variance estimation and other survey methods related to survey processes.

SUB-PROJECT: Generalized system ongoing support and development

The Generalized Systems unit facilitates the use of the systems for new and existing surveys as well as statistical programs undergoing redesign.

Progress:

The Generalized Systems support team provided ongoing support to users and prospective users (both within Statistics Canada and in other organizations), updated and delivered training presentation in various forums and met with international delegates to discuss current and future development of the generalized systems. The G-Est team began development on a new version of SEVANI to address challenges posed to the system by large datasets with complex imputation strategies. This development was done in consultation with clients and stakeholders. This work is now at the prototype stage, with user testing underway. As part of the strategic planning for Generalized Systems, a periodic review of options needs to be conducted. The focus of this review is not on technical problems requiring development of methodology, but is on issues related to processing platforms, and efficient use of existing tools such as open-source languages. As part of this review, a scan of alternative tools that are available in other software was undertaken to identify and compare the functionality with that offered by the Generalized Systems. In addition, a template was developed to allow for consideration of potential enhancements to generalized systems through a business-case approach.

SUB-PROJECT: Generalized System development – Exploration of additional methods and validation techniques

Progress:

The generalized systems team continued the development of an assessment tool based on Monte Carlo non-response simulations, and visualization tools were explored to complement the empirical estimates currently produced. This work was done jointly with Keren Li, a graduate student from the University of Ottawa who prepared a complete work term report (Li, 2019) detailing the visualizations and the process to produce them in the Impact tool. This work was also presented at the 2019 Joint Statistical Meetings conference of the American Statistical Association (Gray, 2019). In addition, the tool has been used to evaluate the imputation strategy for a current survey as a proof of concept. Early research was also conducted into expanding the scope of Banff's nearest-neighbour donor imputation methods, and included a literature review and testing of nearest-neighbour distance metrics.

SUB-PROJECT: Generalized System development – Statistical Data Editing Framework and support for United Nations Economic Commission for Europe initiatives

Progress:

The team continued participation in United Nations Economic Commission for Europe working group to develop editing framework, including contributions to finalize the official report and documentation on the Generic Statistical Data Editing Model (GSDEM) released in June 2019. This report was finalized during the period, by the working group, with significant contributions from the team (UNECE (2019)). A member of the Banff team participated in the organizing committee for the United Nations Economic Commission for Europe Workshop on Statistical Data Editing, scheduled for 2020. Work so far included organizing topics, communication, and reviewing abstracts.

SUB-PROJECT: Generalized System development and version updates

Progress:

A new version of G-Sam was developed, tested and released. G-Sam version 1.03.001 was released in March 2020 (Statistics Canada, 2020). It included a number of non-critical bug fixes, and an update to the allocation module to eliminate redundancies and simplify user inputs. Specifically, a change was made to allow auxiliary values of zero in the allocation optimization problem, which required the addition of new diagnostic tools. A new version of G-EST was developed, tested and released to respond to an error that was found in the

creation of bootstrap replicate weights under specific conditions for variance estimation. G-Est version 2.03.001 was released in November 2019 (Statistics Canada, 2019a). It addressed an error that effected the creation of bootstrap replicate weights. The G-Est team worked swiftly to identify and fix the error, communicating with clients and stakeholders throughout the process.

For further information, please contact:
Steve Matthews (613-854-3174, steve.matthews@canada.ca).

4.3 Questionnaire Design Ressource Centre

The Questionnaire Design Resource Centre (QDRC), in the Methodology Branch, 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 contributed to a corporate initiative related to examining Canadian citizen's perceptions of the secondary use of administrative data as a source of statistical information. Focus groups and consultations were conducted.

The QDRC also started the development of a plan to collect information that may contribute to a sensitivity scale related to survey topics and sources of data.

The group also contributed to various corporate consultation initiatives.

For further information, please contact:
Paul Kelly (613-371-1489, paul.kelly2@canada.ca).

4.4 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.

SUB-PROJECT: Updating the Quality Guidelines

This initiative, the purpose of which is to update the Quality Guidelines, has three objectives: a) to provide a relevant reference document to all the other data producers in the Canadian statistical system, b) to adapt to the new administrative data reality by dealing with the main statistical production processes and c) to play a role in ensuring compliance with current quality assurance methods.

Progress:

Some revisions and additions were made to the draft prepared in 2018/2019. Specifically, guiding principles and best practices involving alternative data, as well as greater emphasis on ethical practices, privacy and proportionality, were added. The new edition of the Quality Guidelines was published in December 2019 (Statistics Canada, 2019b).

SUB-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 others) 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 measures with many partners during the reporting period. Internally, training workshops were held through various courses offered to staff and through more targeted training for teams working on a specific program. With national partners, formal presentations on quality management practices were given to six organizations, in addition to holding many discussions as part of the Data Governance Standardization Collaborative, as well as the Data Quality Working Group. The latter group, co-chaired by Statistics Canada, aims to define a data quality framework that applies to all Government of Canada organizations, as part of implementing the data strategy. Involvement in a data quality panel (Beaulieu, 2020) also led to consultations with other partners. The

quality of a statistical process carried out by another federal organization was also validated. Internationally, presentations were given and discussions held in the context of formal visits by delegations from other countries. Consultation was also provided to a country drafting its own quality assurance framework, and involvement as a United Nations Panel of Experts on national quality assurance frameworks continued, with the aim being to finish writing the United Nations National Quality Assurance Framework Manual for Official Statistics (United Nations, 2019).

For further information, please contact:
Martin Beaulieu (613-854-2406, martin-j.beaulieu@canada.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/.

4.5 Data Analysis Resource Centre

The main goal of Data Analysis Resource Centre (DARC) is to give 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 or other departments, as well as to analysts and researchers from academia or the Research Data Centres (RDCs).

Progress:

Consultations

As part of the DARC mandate, consultation services were provided as requested by various clients. The most requests came from Statistics Canada's analysts and methodologists; these consultations covered topics on survey bootstrap, constructing confidence intervals, testing hypotheses, analysis using linked data, estimating hierarchical linear models with survey data, etc. We also helped our clients with implementation of methods in SUDAAN, SAS, STATA and R software. External consultations were delivered to a variety of clients from other federal and provincial departments and agencies as well as academics from universities. The requests included various topics in analysis of survey data; for example, testing differences between medians, fitting logistic regression, reweighting for nonresponse, degrees of freedom for variance estimation, etc. We also provided a methodological review of a study comparing promotion rates between groups. Expert advice was also given to the analysts and researchers from the Research Data Centres (RDCs). The topics included combining cycles of surveys, using bootstrap weights, multi-level models, quantile regression, analysis of linked datasets, etc.

Provision of Training and Training Material

The team continued presenting topics in statistical data analysis as part of the Data Interpretation Workshop.

At the Annual Conference of the RDC Analysts, we gave a presentation on analysis with linked data.

We drafted Data Visualization, Best practices, intended to be a tool for Statistics Canada analysts and dissemination teams.

We completed two fact sheets answering frequently asked questions:

- Why Fay's adjustment is necessary when using mean bootstrap weights?
- Using survey bootstrap weights versus bootstrapping.

Collaboration with analysts

Two articles were published in the September 2019 issue of Health Reports. (https://www150.statcan.gc.ca/n1/pub/82-003-x/82-003-x2019009-eng.htm).

For further information, please contact:
Harold Mantel (613-863-9135, harold.mantel@canada.ca).

4.6 Time Series Research and Analysis Centre

The objective of the time series research is to maintain high-level expertise and offer needed consultation in the area, to develop and maintain tools to apply solutions to real-life time series problems as well as to explore current problems without known or acceptable solutions. Some of the research activities related to time series are reported under real-time estimation in section 1.2.

SUB-PROJECT: Seasonal Adjustment and Trend-Cycle Estimation

This sub-project consists of evaluation of methods to apply seasonal adjustment and trend-cycle estimation. These methods are intended to support analysis of time series data on official statistics, and the research allows the Time Series Research and Analysis Center to maintain state-of-the-art expertise in the area.

Progress:

A presentation and proceedings article were prepared and delivered to compare X-12-ARIMA and SEATS seasonal adjustment methods. This was conducted by finding analagous state-space models for comparison, and represents progress towards developping state-space-based seasonal adjustment (Matthews and Dochitoiu, 2019). Further to this work, collaboration with the University of Ottawa has been ongoing to develop methodology to apply seasonal adjustment with desirable properties through these models.

Development continued on variance estimation for seasonally adjusted estimates. In particular a presentation and proceedings paper were prepared and delivered for the 2019 Joint Statistical Meetings (Verret and Dochitoiu, 2019). The approach used was the application of a coordinated bootstrap method to estimate the design variance for a business survey. While several issues remain for individual components of the variance, the work represents significant progress towards establishing methods to estimate variance of seasonal adjustment for surveys with a typical business survey design.

The Seasonal Adjustment Dashboard was implemented for the monthly economic surveys to support analysis and interpretation of seasonally adjusted results by methodology and subject matter experts. This automated tool was programmed in R-Shiny and produces an interactive summary of seasonally adjusted results for an individual series according to a template. The tool was presented at the Seasonal Adjustment Practitioners Workshop hosted by the United States Census Bureau (Matthews, Ferland, Verret and Habli, 2019). The tool has now been tested on several surveys over a number of months and has proven to be effective to analyse and understand seasonally adjusted results. Further work on the dashboard will not be

conducted under research budgets as the tool should only require small customizations to be adapted to a given project.

SUB-PROJECT: Support and Enhancement of Time Series Tools

This sub-project consists of the support and development of tools to apply time series methods, notably seasonal adjustment. Development was continued for the Time Series Processing System, currently used in production applications throughout the agency and on demand support was offered to each program.

Progress:

Notably, Version 3.08 of the TSPS (Ferland, 2019) was released to further enhance the offering of techniques available within the system. The new features include custom outputs for the production of the Seasonal Adjustment Dashboard, enhanced benchmarking capabilities and options to generate forecasts or nowcasts. The new benchmarking functionality consists of a new interpolation module added to replace embedded missing data points with either cubic spline interpolations (smooth line) or linear interpolations (straight line segments) along with several projection options for missing data points at the start (before the first value) and end (after the last value) of the series. This module will be particularly useful in benchmarking of stock variables such as inventories and balances. The forecasting options that were introduced allow for production of forecasts and prediction intervals from regARIMA models, which is an important first step in preparing for basic nowcasting methods.

SUB-PROJECT: General Consultation and training in Time Series

As part of the Time Series Research and Analysis Centre (TSRAC) mandate, consultation is offered as requested by various clients within Statistics Canada. Topics most frequently covered are related to the identification of break in series, application of seasonal adjustment and time series modelling in various situations and specific applications of benchmarking and reconciliation.

Progress:

In addition, formal and informal exchanges are held as needed with other statistical organisations (United States Census Bureau, Bureau of Labor Statistics, Eurostat, Statistics Norway, Australian Bureau of Statistics etc.) and academic organizations (University of Waterloo, University of Ottawa) to collaborate and provide input on current topics.

Several journal papers were reviewed on topics related to application of seasonal adjustment, as well as a paper on an applied benchmarking approach to preserve movements in derived series when benchmarking is done in several steps.

Existing courses were updated and delivered as needed to participants from within and outside of Statistics Canada. The newly updated course on "Time Series Modeling and Forecasting" was offered for a second time, and an active learning component was added to the course "Seasonal Adjustment with X-12-ARIMA" to allow for participants to bring data to seasonally adjust in order to gain experience applying the method in a practical setting. In addition, courses on benchmarking, raking and time series components were also delivered.

For further information, please contact:
Steve Matthews (613-854-3174, steve.matthews@canada.ca).

4.7 Data Science Community of Practice

The Statistics Canada Machine Learning Community of Practice has the goal of facilitating collaboration and knowledge transfer as well as improving our machine learning operations at Statistics Canada.

SUB-PROJECT: Machine learning capacity development

Through various activities pertaining to machine learning bringing together 30 to 50 people, such as lunch-and-learns, presentations, reading groups, viewing groups and information-sharing on a site developed and updated by the members, the Community is, through its active presence, still collaborating in the development of the machine learning capacities of Statistics Canada's employees.

Progress:

The Community organized a number of presentations covering multiple areas of machine learning, such as an introduction to computer architecture, automated merchandise classification, the use of an R package for creating synthetic data, an introduction to GitHub and its local version GitLab, content modelling by subject, even a presentation on a new Framework for Responsible Use of Machine Learning Processes. It held two reading groups on articles about machine learning for national statistical organizations, even the use of machine learning for factoring in non-response. The Community also created a viewing group to look at free online machine learning courses, enabling the participants to discuss the topics covered after the viewing. The Community developed a list of existing methodology projects that explore or use machine learning and participated in an exercise for centralizing the presentations of all the communities involved with data science at Statistics Canada.

For further information, please contact:
Yanick Beaucage (613-854-2397, yannick.beaucage@canada.ca).

5. Divisional research and other activities

5.1 Economic Statistics Methods Division

SUB-PROJECT: Evaluation of model-assisted survey regression estimators using Lasso and regression trees

In this project, we evaluate through simulation studies the properties of different survey regression estimators of finite population totals. We include the LASSO method (McConville, Breidt, Lee and Moisen, 2017) and regression trees (McConville and Toth, 2019). The LASSO method is used to select appropriate explanatory variables for the regression estimator. Regression trees select a set of appropriate groups and the resulting regression estimator reduces to a post-stratified estimator. First, we evaluate the methods in the context of a probability sample and then we consider the case of non-probability samples.

Progress:

An extensive simulation study was performed. The results are discussed and summarized in Lundy and Rao, (2019). One main conclusion is that regression trees yields the most efficient regression estimator when the sample size is small and the number of auxiliary variables is large.

For further information, please contact:
Wesley Yung (613-951-4699, wesley.yung@canada.ca).

References

McConville, K.S., Breidt, F.J., Lee, T.C.M. and Moisen, G.G. (2017). Model-assisted survey regression estimation with the LASSO. *Journal of Survey Statistics and Methodology*, 5, 131-158.

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

5.2 Social Statistics Methods Division

SUB-PROJECT: Neural networks for the imputation of income tax

The Administrative Personal Income Masterfile (APIM) is an administrative database of personal income built from T1s and other tax slips. On this file, the partial income information available for non-tax filers is completed through imputation. The imputation of taxes is a particularly complicated step. Federal and provincial income taxes are imputed through four consecutive rounds of donor imputation. First, taxable income is imputed, and this is used in deriving tax credits and imputing federal and provincial taxes in subsequent rounds. The imputation process uses a range of auxiliary variables such as various types of income and demographics to find donors, as well as formulas for the derivation of net tax and tax credits using inputted tax parameters. Changes in tax rules from year to year must be monitored and taken into account in the imputation process. This process is onerous and has not always been updated properly in the past.

Given the multi-level and non-linear relationships between the input data and tax variables imputed, neural networks have the potential to be an appropriate class of model and provide a suitable alternative method to donor imputation. They could also reduce the need to update the imputation programs manually to take into account changes to tax credit calculations. In addition, APIM is a large file with 30 million records giving much data to train on. Thus, the project objective is to evaluate the use of neural networks for the imputation of income taxes in the APIM as an alternative to donor imputation.

Progress:

With advice from the Data Science division, a plan was put forward to explore XGBoost before neural networks. Both methods were further researched to learn how they could be used in the context of this project. An initial model using XGBoost was developed, trained and tested using a subset of the data (300,000 records which is ~10% of the full data set). An initial attempt at parameter tuning was also started but could not be completed due to the work stoppage in March.

In our future work, we plan to compare the imputation results using four different methods: a standard regression model, XGBoost, neural networks and the currently implemented donor imputation. For a subset of the records being imputed, the true values provided on tax forms are available, which would be used as the benchmark to see which method performs the best.

SUB-PROJECT: Area-level model in small area estimation

The demand for small area estimates by users of Statistics Canada’s data has been steadily increasing in the recent years. The development of an SAE production system was started in

the early 2000s. The production system, which is now a part of the G-EST software, handles the basic area-level model by Fay and Herriot (1979) with multiple options such as different methods to estimate the variance components and estimation methods. The basic area-level may not provide satisfactory results, if a single linear model does not provide an adequate explanation on the relationship between the variable of interest and the covariates. The piecewise area-level model is a modification of the basic area-level model in which the area specific auxiliary variable is partitioned into intervals and a separate line segment is fit to each interval. Moreover, the choice of the covariates has a major impact on the quality of estimates and the accuracy of the model. In general, the auxiliary data must come from a source that is independent of the direct estimates, and it must be available at the appropriate levels of geography. Sometimes in practice, the model is fitted on covariates, which are not strongly correlated to the variable of interest or have coverage issues.

This research project has two main objectives: 1) to compare the piecewise area-level model with the basic area-level model and evaluate the impact of the piecewise approach on model parameters and residuals, 2) to examine the model with covariates that are not fully correlated to the dependent variable or have coverage issues (e.g., aggregated data on spending). The models is tested on data from Visitor Travel Survey (VTS), and payment data is used as the auxiliary information in the model.

Progress:

The impact of a piecewise modification to the basic area-level model was explored. The results of the analysis on the VTS data with payment data as the auxiliary information suggested that the piecewise modification improves the model fit. In particular, the piecewise area-level model displayed residuals with a slightly better behaviour in terms of normality, homoscedasticity and linearity. It also resulted in more reliable estimates, with smaller CV's on average than the basic area-level model. In addition to the payment data, the piecewise area-level model was also tested with aggregated data on spending as the auxiliary information. The current iteration of the aggregated data on spending has major coverage issues and quality concerns resulting in significantly fewer domains available for modelling and weaker correlation with VTS spending. The findings were documented in an internal document. Depending on the priorities and resources, the area-level models with both payment data and aggregated data on spending, either as separate covariates or as one combined auxiliary variable, will be tested and evaluated.

SUB-PROJECT: Using the National Road Network to calculate distance

The National Road Network (NRN) is a series of datasets under the responsibility of Statistics Canada, available on the government's Open Data portal (open.canada.ca). These datasets

contain the longitude and latitude coordinates of points along the Canadian road network, as well as information regarding which points are joined by a road. Thus, the datasets can be used to calculate the road distance (within the country) between any two geographical points.

The Canadian Health Measures Survey (CHMS) uses the distance between sampled addresses and the physical medical clinic stationed in each sampled site in models of non-response; indeed, this distance has been used in modeling non-response to the clinic in 3 of the first 5 cycles of CHMS, and the closely related variable of postal code has been used in the other 2. However, CHMS has only been able to use straight-line distance (using PROC GEODIST), since the address data is statistically sensitive. The CHMS team has begun to develop a solution using the NRN files. However, this solution has not been widely tested in order to determine its quality and is not currently generalizable to other projects that may wish to make use of the NRN files as well. The project objectives are to evaluate whether distance by road offers an improvement over distance as-the-crow-flies in non-response modeling for CHMS, and to document the status of related work being done in Methodology and/or elsewhere at Statistics Canada.

Progress:

Non-response adjustment models using cycle 5 CHMS data were compared, using as-the-crow-flies distances and road distances. The results of these comparisons were presented to a meeting of the Spatial Analysis Community of Practice (Emond and Mather, 2019). The SAS code was generalized so the macros could be used by other projects. The researchers have connected with and participated in the Spatial Analysis Community of Practice, where projects and ideas regarding geospatial analyses are shared. The members of this community and record of the meetings (available on GCdocs) serve as a great synopsis of the related work happening at the agency. The results of the research showed that, for CHMS, there is not a lot of value added by using road distance over as-the-crow-flies distance in non-response modeling. Thus, although the survey will use road distances moving forward, ongoing future work in this area is not anticipated.

References

Fay, R.E., and Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*.

Emond, N., and Mather, A. (2019). Distance by Road – CHMS, Presentation to the Spatial Analysis Community of Practice.

SUB-PROJECT: Optical character recognition as part of modernizing the Household Expenditure Program

The Survey of Household Spending collects information about Canadian household spending using two main methods, namely a computer-assisted personal retrospective interview and a diary of expenses (paper format). For the diary component, the respondents are asked to write down their expenses and/or place the receipts for their purchases in a folder for a period of one week. Those receipts are scanned at head office, and the information is currently entered manually from the receipt images. For each collection cycle, the data from approximately 30,000 receipts are entered manually, which is a significant burden for the agency. The methodology team began a research and development project for automating the entry of the information on the receipts. In the first phase, a machine learning algorithm (neural network) was trained to recognize store logos, under the assumption that the layout on the receipts of the items to be captured is different for each header. The experiment proved to be very positive, with the algorithm successfully classifying logos for over 95% of the receipts in the test sample. In order to optimize those results and prepare for capturing the items and amounts on receipts, the team would like to do machine learning research for character recognition from the receipt images. This might make it possible to extract the names of stores whose logos were not recognized by our algorithm and capture the rest of the information of interest. The context lends itself really well to this type of research because we have a very large bank of receipt images and information entered over time (~10 years), which is ideal for training algorithms. Although software is already used internally for optical character recognition (with forms already provided for this purpose), consultation with various stakeholders has shown that there is no approach for capturing unstructured information.

Progress:

An international collaboration was established with Lan Benedikt, a data scientist from the Data Science Campus, who is working on a very similar project for the Living Cost and Food Survey by the Office for National Statistics (United Kingdom). Lan had previously developed expertise on implementing techniques for extracting information from receipts (Benedikt, 2019) and, here at Statistics Canada, many years of entered and coded receipts are available. So, an agreement was established, and Lan came to Canada to work with the team for two weeks (August 26 to September 6, 2019). Extraction methods were identified, programs were written or modified, and some testing was done. A brief presentation was given to the management of the two teams, at the Data Science Accelerator, which will have a very similar project in the future, and to internal clients who are apparently involved in implementing such a production strategy (Benedikt and Mayer, 2019). In concrete terms, this project made it possible to acquire knowledge and programs in image processing, optical character recognition, and information extraction techniques (parsing) using either a word dictionary or

regex methods. Given the complexity and variability of receipts, machine learning has currently been deemed non-relevant for this component of the project, but is still very relevant for the component involving auto-coding of items on receipts. The preliminary results show real potential for the use of character recognition techniques for capturing information on receipts. The project was presented to Statistics Canada's Research and Development Board (Malo and Mayer, 2019), which agreed to fund the next phase of the project, which will run from March 15 to July 31, 2020.

References

Benedikt, L. (2019). Human-in-the-Loop AI in Government: A Case Study. Presentation given for a Statistics Canada Methodology seminar, September 2019.

Benedikt, L., and Mayer, E. (2019). STC and ONS-DSC collaboration: Automatic Capture and Coding of Shopping Receipts, Presentation to management, OID, ISD and DScD, September 2019.

Malo, D., and Mayer, E. (2019). Automation of the Capture of Shopping Receipts, Presentation given to Statistics Canada Research and Development Board, December 2019.

SUB-PROJECT: Constructing confidence intervals for differences of proportions

In 2017, the Methods and Standards Committee (MSC) approved Methodology's recommendation to adopt as a best practice the use of confidence intervals for measuring and reporting the quality of estimates. Following this recommendation, research is required to provide the framework to support the use of confidence intervals. The project objective is to research confidence interval methods and provide recommendations for a variety of situations in which analysts will use them.

Progress:

A presentation was given to the Scientific Review Committee on Social Statistics on release guidelines when confidence intervals are used to report quality (Neusy and Baribeau, 2019). The goal of the presentation was to make progress towards establishing a recommended and approved set of guidelines. The Committee approved a general framework for release guidelines. Since the committee did not approve specific thresholds, interim rules were developed and documented for the Centre for Social Data Integration and Development surveys (Neusy, 2019a). A presentation was given to the Working Group on Quality

Indicators – Estimation (Neusy, 2019b) in order to coordinate the overlap between the two projects. A short article was written for The Survey Statistician on reporting the quality of estimates through confidence intervals (Neusy, 2020a). Simulations were undertaken to evaluate the performance of modified Wilson confidence intervals for domain estimates under stratified random sampling. Simulation results showed that modified Wilson (and Clopper-Pearson) intervals perform well for domain estimates of proportions. The results were documented (Neusy, 2020b). Simulations were also undertaken to evaluate the performance of confidence intervals for the difference of two proportions. Simulations were run with variables that had various levels of correlation, from independent to highly correlated. As well, variables with differences ranging from 0% to 20% in the underlying population were simulated. The performance of confidence intervals was mixed: good for some scenarios and poor for others. Further research is needed to fully analyze and understand the results of the simulations for the differences of proportions.

For further information, please contact:
François Brisebois (613-222-8310, françois.brisebois@canada.ca).

5.3 Statistical Integration Methods Division

SUB-PROJECT: Estimation of linkage errors accounting for grouping and mapping

The general problem is the estimation of linkage error (particularly false positives) based on the linkage weights. This is an extension of the work of Labrecque-Synnott (2019), in which weight-based estimation was explored without taking into consideration the grouping and mapping constraints. The current objective was to take these into account, as almost all linkages conducted in the Social Data Linkage Environment (SDLE) assume a one-to-one mapping between the linked files.

Progress:

Groups of potential links created in a probabilistic linkage project can be divided into simple groups (a single individual from file a potentially linking to multiple individuals on file B, or vice-versa) and complex groups (involving multiple individuals from file A, and multiple individuals from file B). A theoretical solution has been developed for both of these cases. The problem of error estimates taking grouping and mapping into account can be viewed as a conditioning problem within the Fellegi-Sunter (1969) theory of record linkage, for which formulas can be derived. Based on the structure of events involved in record linkage, simplifications can be made to make these workable in practice. These were programmed in SAS and tested on one of the SDLE linkage projects. For complex groups involving a large number of links, the computations can take a very long time even with the simplifications mentioned above. Work has begun to find a suitable approximation for those cases. This work has been documented in Labrecque-Synnott (2020).

Reference

Fellegi, I., and Sunter, A. (1969). A theory for record linkage. *Journal of the American Statistical Association*, 64(328), 1183-1210.

SUB-PROJECT: Optimization of Mapping one-to-one in record linkage

The greedy method based on the higher total weight is presently applied in G-Link in order to do the mapping one to one. With this strategy it is not guaranteed that we reach an optimal solution because we perform local choices. Once the linkage total weight is assigned to each pair, the identification of one to one links can be solved as a linear programming problem

where the objective function to maximise is the sum of weights of the linked pairs under the constraints that each unit of the table A must be linked with only on unit of the table B.

The objectives of this project are:

- to document the algorithm used and apply it to real data;
- to improve the one-to-one mapping methodology, specifically factoring in the number of pairs to be considered;
- to implement the resolution algorithm in G-Link.

Progress:

Progress included a review of existing literature (Bertsekas, 1992; Chipperfield, Hansen and Rossiter, 2018; Hungarian Algorithm, 2013; Jaro, 1989; Jin, 2016; Lee, Xiong, Yu and Li, 2018; Sadinle, 2016; Sahu and Rudrajit, 2007) as well as available algorithms applied in software such as Febrl and Relais.

References

Bertsekas, D.P. (1992). Auction Algorithms for Network Flow Problems: A Tutorial Introduction.

Chipperfield, J., Hansen, N. and Rossiter, P. (2018). Estimating precision and recall for deterministic and probabilistic record linkage. Docklands: *International Statistical Review*.

Hungarian Algorithm (2013). Website. 16 08 2019. <www.hungarianalgorithm.com/index.php>.

Jaro, M.A. (1989). Advances in record-linkage methodology as applied to matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84, 406, 414-420. 14 05 2019. <https://www.jstor.org/stable/2289924>.

Jin, X. (2016). *Parallel Auction Algorithm for Linear Assignment Problem*. Project. Stanford.

Lee, M., Xiong, Y., Yu, G. and Li, G.Y. (2018). Deep Neural Networks for Linear Sum Assignment Problems. *IEEE Wireless Communications Letters*, 7.6, 962-965.

Sadinle, M. (2016). *Bayesian Estimation of Bipartite Matchings for Record Linkage*.

Sahu, A., and Rudrajit, T. (2007). Solving the Assignment Problem Using Genetic Algorithm and Simulated Annealing.

SUB-PROJECT: Spatial Analysis

This research project started this year with the goal of working on spatial analysis in the context of the Business Register (BR). For this year, two objectives were explored (1) Implement a road network distance calculation algorithm; (2) Creation of Walkability/Proximity Score for businesses.

Road Network Distance vs Straight-Line Distance: Statistics Canada's Business Register Experience

Lately, many surveys have started measuring or quantifying the distance between businesses to produce proximity statistics. Many surveys use distance as a simple measure of accessibility, risk, or disparity. Up to now, most surveys have used the straight-line (Euclidean) distance because of its ease of calculation. However, road network distance provides is more accurate for the actual distance between businesses, although this alternative is more computational-intensive and may also be more expensive.

Progress:

In 2019-20, a program was developed using the Dijkstra's Shortest Path algorithm, SAS Proc OPNET, Canadian National Road Network (NRN) and the Business Register to calculate 906,780,919 road network distances and compare them to their straight-line distance counterpart. Many surveys have been using server capacity in high levels to calculate the road network distance, this has created a shortage of computer resources which led to the idea of applying a model to the straight-line in order to obtain a good estimate of the road network distance. Work has also begun on an article/report and a presentation on this subject.

Business Register's Proximity Score

The growth of geocoded data has created an influx of new applications related to geo-distances. The early 2010s saw the emergence of the Walk Score (walkability index), heat maps and other spatial analysis applications. The Walk Score is a type of automated efficiency model focused on location efficiency. Since 2016, the Business Register produces a geocoded location for all businesses. The geocoded location provides user with a latitude-longitude representation of block-face centroids. The latitude-longitude coordinates have previous been used for record linkage, distance calculation for transport surveys and other applications. In 2019, the possibility of using the latitude-longitude coordinates to create a proximity score by industries (North American Industry Classification System (NAICS)) was explored.

Progress:

In 2019-20, a methodology for a Business Register proximity score was developed. The methodology consists of extracting active establishments with a valid address from the Business Register and creating a cluster of all units within a predetermined distance (1 km), using the latitude-longitude coordinates. This provides the number of units in the predetermined radius proximity zone for all establishments. The score is available for users.

SUB-PROJECT: ECONOMIC INDICATORS: Assessing their use in the model

The dwelling occupancy model is used in the Census of Population to predict the occupancy status of private dwellings in Canada as either occupied, unoccupied or void (invalid). The purpose of this model is to allow identifications of as many unoccupied or void dwellings as possible so that their status is verified during the dwelling verification operation, and so that they are subsequently removed from the list of dwellings for which a non-response follow-up is necessary.

Following the 2019 dwelling verification operation, it was hypothesized that economic indicators associated with dwellings' regions would make it possible to better identify the unoccupied or cancelled dwellings within them, and thus using them in the model would help improve it. Therefore, the purpose of this research project was to assess whether employment-related economic indicators could be obtained and used in the model to increase its predictive power. This would help improve the dwelling occupancy status predictions and therefore reduce the cost of the non-response follow-up operation.

Progress:

The project focused on the concrete example of plant closures by assessing how the information in the Business Register, specifically the operational status of each business, the date that the status changed, and the number of employees impacted by that change, could be used to improve the model. Although the results did not prove conclusive compared to those of the initial model, recommendations were made regarding how to make better use of the information in the Business Register if this avenue were to be explored again, and regarding the potential contribution of using a multilevel logistic regression model in predicting dwelling status (Legault, 2020).

For further information, please contact:
Michelle Simard (613-293-3192, michelle.simard@canada.ca).

5.4 International Cooperation and Methodology Innovation Centre

SUB-PROJECT: Development of a prototype system for robust estimation

In many economic and a few social surveys, variables with skewed distributions are collected for the sample units, which may result in the presence of outliers and influential units. Traditional estimation methods may produce estimators that are highly inefficient when samples may contain influential units. The idea of robust estimation is to diminish the effect of influential sample units on the estimates of interest. The concept of the conditional bias is used as a measure of influence as it represents the contribution of each sample unit on the sampling error of an estimator. The traditional sample estimate is decreased by a function of the conditional bias of the sample units. The conditional bias was first proposed by Moreno-Rebollo, Munoz-Reyez and Munoz-Pichardo (1999) and later used to develop a robust estimator by Beaumont, Haziza and Ruiz-Gazen (2013). This work is relevant to the many economic and social surveys at Statistics Canada.

Progress:

A prototype has been developed in SAS to include many of the specifications for robust estimation described in Beaumont (2017). It consists of a series of macros for the various functions related to the production of domain robust estimates. It includes a macro to calculate the traditional domain estimates as well as the domain robust estimates. The domain robust estimates generally do not have the additivity property of the domain estimates, so a macro exists to meet this requirement by creating domain coherent estimates from the domain robust estimates through minimal change of their values. There is also another macro to produce recalibrated weights for the sample units to ensure that they reproduce the domain coherent estimates and any known totals of auxiliary variables. The macros are available for use through the support of the International Cooperation and Methodology Innovation Centre.

SUB-PROJECT: Bootstrap estimation of the conditional bias for measuring influence in complex surveys

In sample surveys that collect information on skewed variables, it is often desirable to assess the influence of sample units on the sampling error of weighted estimators of finite population parameters. The conditional bias is an attractive measure of influence that accounts for the sampling design and the estimation method. It is defined as the design expectation of the sampling error conditional on a given unit being selected in the sample. The estimation of the conditional bias is relatively straightforward for simple sampling designs and estimators. However, for complex designs or complex estimators, it may be tedious to derive an explicit

expression for the conditional bias. In those complex surveys (e.g., the Survey on Household Spendings), variance estimation is often achieved through replication methods such as the bootstrap. Bootstrap methods are typically implemented by producing a set of bootstrap weights that is made available to users along with the survey data. We study how to use these available bootstrap weights to obtain an estimator of the conditional bias. This estimator could then be used to construct robust estimators of finite population parameters that are less negatively affected by influential units than standard weighted estimators. We plan to evaluate our bootstrap estimator of the conditional bias in a simulation study.

Progress:

We developed a bootstrap estimator of the conditional bias using the bootstrap weights and wrote a draft paper (Beaumont and Bocci, 2020b). We have started the simulation study and obtained preliminary results.

SUB-PROJECT: Bootstrap variance estimation for multistage sampling with application to nonresponse

The bootstrap is often used for variance estimation in surveys with a stratified multistage sampling design. It is typically implemented by producing a set of bootstrap weights that is made available to users and that accounts for the complexity of the sampling design. The method of Rao, Wu and Yue (1992) is often used to produce the required bootstrap weights. It is valid under stratified with-replacement sampling at the first stage or when the first-stage sampling fractions are negligible. Some surveys do not satisfy these conditions. The goal of this project is to propose a bootstrap methodology for multistage designs that would be applicable when the conditions for the validity of the Rao, Wu and Yue (1992) bootstrap are not satisfied.

Progress:

We developed a simple bootstrap method that is valid even with non-negligible sampling fractions. It is applicable to any multistage sampling design as long as a valid bootstrap method is available for each individual stage of sampling. Our method is also applicable to two-phase sampling designs with Poisson sampling at the second phase. We use this design to derive bootstrap weights that account for nonresponse weighting. A draft paper is currently being written (Beaumont, 2020c).

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

References

Beaumont, J.-F. (2017). Robust Estimation Prototype, Methodology Specifications. Internal report, Statistics Canada.

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2013). A unified approach to robust estimation in finite population sampling. *Biometrika*, 100, 555-569.

Moreno-Rebollo, J.L., Munoz-Reyez, A.M. and Munoz-Pichardo, J.M. (1999). Influence diagnostics in survey sampling: Conditional bias. *Biometrika*, 86, 923-928.

Rao, J.N.K., Wu, C.F.J. and Yue, K. (1992). Some recent work on resampling methods for complex surveys. *Survey Methodology*, 18, 2, 209-217. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1992002/article/14486-eng.pdf.

5.5 Development program

The Statistical Talent Development Working Group was created and took the lead in continuing on the path towards a modernized statistical program. A number of new and/or redesigned courses in the curriculum were offered for the first time, including a small area estimation course. The data science course piloted last year and taught by a distinguished university professor was a success and became part of the regular curriculum. Furthermore, a new R course was also developed in-house and piloted. One of the objectives moving forward for statistical training is to offer various and efficient delivery of training, with an emphasis on participatory activities where active learning is favoured when possible. Areas that have emerged as priority include data science and machine learning as well as data integration and statistical modeling. With those priorities in mind, a new series of learning activities on statistical modeling are in development.

Other learning opportunities included a large number of recommended online courses, webinars and the data science community of practice. Ad-Lib, a new activity being developed, is an interactive training that is somewhere between a reading group and working group where a theme is given which serves as a starting point for discussion that evolves according to the interests and interventions of the participants and a facilitator over few hours.

Finally, this was the year that the micromission program was launched in the Branch. Micromissions are informal short-term corporate assignments, a form of active learning, that allow Branch employees to be temporarily assigned to projects outside the Branch. The main objective is to provide opportunities for staff to learn and become familiar with priorities and work conducted outside the Branch and, in some cases, outside Statistics Canada. The first year was a success with many employees participating.

For further information, please contact:
Pierre Caron (613-612-6910, pierre.caron@canada.ca).

5.6 Publication - *Survey Methodology*

Survey Methodology is an international journal available at www.statcan.gc.ca/surveymethodology that publishes articles in both official languages on various aspects of statistical development relevant to a statistical agency. Its editorial board includes world-renowned leaders in survey methods from the government, academic and private sectors. The journal is released in fully accessible HTML format and in PDF.

The work related to the editorial and production processes include: correspondence with authors, referees, associate editors, and subscribers; review of referees' comments and author revisions; re-formatting manuscripts; copy editing of manuscripts; liaison with translation and dissemination; and maintenance of a data base of submitted papers. It is part of the knowledge transfer activities.

Progress:

The June and December 2019 issues (45-1 and 45-2) were released in PDF and HTML versions. The June 2019 issue includes 10 papers, including the Special Waksberg Invited Paper (Rubin, 2019). The December 2019 issue included 8 papers and one short note. The journal also published a special issue in May 2019 as part of a special collaboration with the International Statistical Review in honour of Prof. J.N.K. Rao's contributions. That special issue included 8 papers and a short piece by Prof. J.N.K. Rao.

From April 2018 to March 2019, the *Survey Methodology* pages were viewed 27,000 times and nearly 6,000 copies of papers were downloaded using an improved web metrics methodology. Aside from the invited papers for the special issues, 31 papers were submitted for publication.

In 2019, the publication of 3 issues of the journal is planned. In addition to the two regular issues, a special issue showcasing some papers presented at a conference titled "Contemporary Theory and Practice in Survey Sampling: A Celebration of Research Contributions by J.N.K. Rao" will be published in collaboration with the International Journal of Statistics.

For further information, please contact:
Susie Fortier (613-220-1948, susie.fortier@canada.ca).

Reference

Rubin, D.B. (2019). Conditional calibration and the sage statistician. *Survey Methodology*, 45, 2, 187-198. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019002/article/00010-eng.pdf.

6. Research papers sponsored by the Methodology Research and Development Program

Arim, R., Bougie, E., Michaud, I., Tabuchi, T., Yung, W. and Kohen, D. (2019). A mixed-method exploration of the Public Service Employee Survey (PSES) items as potential measures of psychosocial factors in the workplace. A report prepared for the Workplace Mental Health Performance Measurement Project Group.

Beaulieu, M. (2020). Présentation au panel « Qualité des données dans l'ensemble du gouvernement : qu'est-ce que cela veut vraiment dire ? ». Canadian Government data conference.

Beaulieu, M., Chepita, R. and Fortier, S. (2019). Building a quality indicators framework in a multi-source environment. Presented at *International Total Survey Error Workshop.*

Beaumont, J.-F. (2020a). Are probability surveys bound to disappear for the production of Official Statistics? *Survey Methodology*, 46, 1, 1-28. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2020001/article/00001-eng.pdf.

Beaumont, J.-F., and Bocci, C. (2020b). Bootstrap estimation of the conditional bias for measuring influence in complex surveys. Draft internal report, Statistics Canada.

Beaumont, J.-F. (2020c). Bootstrap variance estimation for multistage sampling with application to nonresponse. Draft internal report, Statistics Canada.

Beaumont, J.-F., and Chu, K. (2020). Statistical data integration through classification trees. Paper to be presented at the Advisory Committee on Statistical Methods, June 2020, Statistics Canada.

Beaumont, J.-F., Lesage, É. and Rao, J.N.K. (2020). Estimation of the design mean square error in small area estimation. Draft internal report, Statistics Canada.

Bocci, C., and Beaumont, J.-F. (2019a). Small area estimation methodology of the unemployment rate in special labour areas. Internal report, Statistics Canada.

Bocci, C., and Beaumont, J.-F. (2019b). Small area estimation of unemployment rate at the special labour area level. Internal report, Statistics Canada.

Buresi, G. (2019). Évaluation de la robustesse du modèle de Fay-Herriot pour les petits domaines. Internal document, Statistics Canada.

Chu, K., and Beaumont, J.-F. (2019). The use of classification trees to reduce selection bias for a non-probability sample with help from a probability sample. *Proceedings of the Survey Methods Section*, Statistical Society of Canada, May 2019.

Colley, R.C., Christidis, T., Michaud, I., Tjepkema, M. and Ross, N.A. (2019). An examination of the associations between walkable neighbourhoods and obesity and self-rated health in Canadians. *Health Reports*, 30(9), 14-24.

Colley, R.C., Christidis, T., Michaud, I., Tjepkema, M. and Ross, N.A. (2019). The association between walkable neighbourhoods and physical activity across the lifespan. *Health reports*, 30(9), 3-14.

Dasylva, A., Goussanou, A., Ajavon, A. and Abousaleh, H. (2019). Revisiting the probabilistic method of record linkage. Internal report, Statistics Canada.

Do, Q. (2020). *Mixed Methods Research on International Engagement*.

Ferland, M. (2019). Time Series Processing System - What's new in V3.08. Internal document, Statistics Canada.

Gray, D. (2019). A Generalized Framework to Evaluate Imputation Strategies: Recent Developments. Presented at the Joint Statistical Meetings of the American Statistical Association.

He, Y. (2019). Estimation of errors in privacy-preserving record linkage. Internal report, Statistics Canada.

Hidiroglou, M.A., Beaumont, J.-F. and Yung, W. (2019). Development of a small area estimation system at Statistics Canada. *Survey Methodology*, 45, 1, 101-126. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2019001/article/00009-eng.pdf.

Labrecque-Synnott, F. (2019). Estimation et propagation d'erreurs de couplage. Internal document, Statistics Canada.

Labrecque-Synnott, F. (2020). Weight-based linkage error estimation accounting for grouping and mapping. Internal document, Statistics Canada.

Lapointe, M.-A., and Mischler, L. (2019). Détecter la saisonnalité : Un enjeu actuel de l'analyse des séries chronologiques. Internal document, Statistics Canada.

Legault, J. (2020). Utilisation d'indicateurs économiques dans le modèle d'occupation des logements (DOM). Internal document, Statistics Canada.

Lesage, É., Beaumont, J.-F. and Bocci, C. (2020). Deux diagnostics locaux pour évaluer l'efficacité du meilleur prédicteur empirique issu de modèle de Fay-Herriot. *Survey Methodology* (under revision).

Li, K. (2019). Data Visualization for an Imputation Strategy Evaluation Tool. Work Term Report, submitted to the University of Ottawa.

Lundy, E., and Rao, J.N.K. (2019). Simulation study of model-assisted survey regression estimation. Internal report, Statistics Canada.

Matthews, S., and Dochitoiu, C. (2019). Comparison of seasonal adjustment approaches through state space representation. Proceedings: Symposium 2019, Forecasting, Statistics Canada.

Matthews, S., and Patak, Z. (2020). Towards real-time estimation through time series modelling. Internal document, Statistics Canada.

Matthews, S., Patak, Z., Picard, F. and Mischler, L. (2020). Technical documentation of development of Nowcasting options in official statistics. Internal document, Statistics Canada.

Matthews, S., Ferland, M., Verret, F. and Habli, N. (2019). De-mystifying Seasonal Adjustment: A visual tool to understand the process. 3rd Seasonal Adjustment Workshop, Washington, D.C.

Matthews, S., Patak, Z., Picard, F. and Mischler, L. (2020). Technical documentation of development of nowcasting options in official statistics. Internal document, Statistics Canada.

Miller, J. (2019). Survival analysis of the Canadian Mortality Database linked to the Canadian Community Health Survey. Internal report, Statistics Canada.

Mischler, L. (2019). Reference Week Adjustment of Employment Insurance Statistics. 3rd Seasonal Adjustment Workshop, Washington, D.C.

Neusy, E. (2019a). Interim Release Guidelines for CSDID Surveys. SSMD Internal Document, Statistics Canada.

Neusy, E. (2019b). Reporting Quality using Confidence Intervals. Presentation to the Working Group on Quality Indicators –Estimation.

Neusy, E. (2020a). Reporting the quality of estimates through confidence intervals. *The Survey Statistician*, Country Report, No. 81, January 2020.

Neusy, E. (2020b). Wilson Confidence Interval Simulations. Internal Document, Statistics Canada.

Neusy, E., and Baribeau, B. (2019). Quality-based Release Criteria for Social Statistics Part 2, Presentation to the Scientific Review Committee on Social Statistics, September 2019.

Oyarzun, J., and Zhang, S. (2019). Business Register's Proximity Score. Presented at Statistics Canada's Business Register Analysis Meeting, June 7, 2019.

Oyarzun, J., and Zhang, S. (2020). Road network distance vs straight-line distance: Statistics Canada's business register experience. Internal document, Statistics Canada.

Rancourt, E. (2019). The scientific approach as a transparency enabler throughout the data life-cycle. *Statistical Journal of the IAOS*, 35, 549-558.

Reedman, L. (2019). A Framework for Responsible Machine Learning Processes at Statistics Canada. Presented to Statistics Canada's Advisory Committee on Statistical Methods. Internal paper, October 2019.

Sallier, K. (2020). Toward More User-Centric Data Access Solutions: Producing Synthetic Data of High Analytical Value by Data Synthesis, submitted for publication (awarded first prize in the 2020 IAOS Prize for Young Statisticians).

Statistics Canada (2019a). G-EST 2.03.001 Release Notes/G-Est 2.03.001 Upgrade notice, https://www150.statcan.gc.ca/n1/en/catalogue/10H0035.

Statistics Canada (2019b). Statistics Canada Quality Guidelines. Sixth Edition. Statistics Canada, Catalogue No. 12-539-X. Ottawa, Ontario. https://www150.statcan.gc.ca/n1/pub/12-539-x/12-539-x2019001-eng.htm.

Statistics Canada (2020). G-Sam 1.03.001 Release Notes/G-Sam 1.03.001 Upgrade notice, https://www150.statcan.gc.ca/n1/en/catalogue/10H0031.

UNECE (2019). Generic Statistical Data Editing Model (GSDEM) - Version 2.0, Modern Stats by HLG-MOS https://statswiki.unece.org/plugins/servlet/mobile?contentId=117771706#content/view/117771706.

Verret, F., and Dochitoiu, C. (2019). Estimating the variance of seasonally-adjusted series of monthly Statistics Canada surveys. *Proceedings of the Joint Statistical Meetings of the American Statistical Association*.

You, Y. (2019). A summary report on "Supplementing small probability samples with nonprobability samples: A Baysian approach". ICMIC internal report, Statistics Canada.

You, Y. (2020a). EBLUP and Hierarchical Bayes small area estimation of LFS rates using area level models with sampling variances smoothing vs modeling. ICMIC internal research report, Statistics Canada.

You, Y. (2020b). Report on hierarchical Bayes small area estimation of LFS totals using area level models. ICMIC internal research report, Statistics Canada.