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

Statistical Methodology Research and Development Program

Annual Report 2018/2019

Release date: November 15, 2019

Statistics Canada Statistique 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 2018/2019 achievements of the Methodology Research and Development Program (MRDP) sponsored by the Methodology 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 2018/2019

Contents

1. Research Projects

2. Support Activities

1 Research Projects

1.1 Developmental Research – 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. The system and methods are documented in Hidiroglou, Beaumont and Yung (2018). 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: 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 global performance of small area estimates. However, users are often interested only in their specific domain and an indicator of the quality of 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. The goal of this project is to develop and investigate new local diagnostics for the evaluation of small area estimates.

Progress:

First, we derived an interval on the local random effect ensuring that the design MSE of the best predictor is smaller than the design MSE of the direct estimator. Our first diagnostic evaluates the conditional probability that the local random effect is in the interval. If the probability is too low, it may be preferable to choose the direct estimator over the small area estimator. The second diagnostic is the P-value of a design-based hypothesis test that the random local effect is not larger than the limit of the interval. We developed the theory and conducted empirical studies. Preliminary results were presented at the Colloque francophone sur les sondages in October 2018. A paper is currently being written (Lesage, Beaumont and Bocci, 2019).

SUB-PROJECT: LFS unemployment rate estimation using cross-sectional and time series models

The goal of this project is to study the use of time series methods for the estimation of unemployment rates by CMA/CA across Canada as well as to investigate and compare the LFS unemployment rate estimates under the cross-sectional Fay-Herriot (FH) model and time series models.

Progress:

We studied LFS unemployment rate estimation using FH model, spatial FH model, spatio-temporal FH model and cross-sectional/time series model. For spatial FH model, studied the LFS estimates based on the SAR SFH of Petrucci and Salvati (2006) and the CAR SFH of You and Zhou (2011). For cross-sectional/time series model, we studied the model of You, Rao and Gambino (2003) and the spatio-temporal FH of Marhuenda, Molina and Morales (2013). We also studied the R sae package and gave a presentation (You, 2018) comparing the SAE estimates for LFS unemployment rates based on different models using the R and S-Plus functions. Our results indicate that the FH model is very useful, and in the LFS application the FH model is adequate to improve the direct survey estimates by reducing the relative errors and improve the precision. Spatial FH or time series models do not necessarily provide better results than the FH model in the LFS application.

SUB-PROJECT: Unemployment rate estimation with different variance modeling at CMA-NOCS4 level using a hierarchical Bayes (HB) model with R and G-Est system

The goal of this project is to investigate and study the feasibility of producing small area estimates of unemployment rates using HB modeling approach at CMA-NOCS4 level. Use S-plus SAE functions and compare the results obtained by the G-EST for this problem.

Progress:

We obtained the SAE estimates of LFS unemployment rates at the CMA-NOCS4 level for 3220 small domains using the G-EST SAS package and the HB estimates based on HB sampling variance models using R and S-Plus functions. Estimates based on FH model using REML and

ADM have been obtained using SAS G-EST package and R functions. The G-EST SAS package and the R SAE functions perform very well in terms of computation efficiency. By comparing the model-based estimates and the direct estimates, benchmarked EBLUP of You, Rao and Hidiroglou (2013) or HB benchmarking procedure (You, Rao and Dick, 2004) may be needed to reduce possible bias when the small area level is low and the number of areas is very large.

SUB-PROJECT: Evaluation and comparison of HB modeling of totals and rates at CMA/CA level

The unemployment rate can be modelled directly in the Fay-Herriot model. This is the approach that has been taken so far. Alternatively, the numerator and denominator of the rate can be modelled separately and the rate can be derived from these two separate total estimates based on area level models.

Progress:

Area level models have been studied for unemployment rate, unemployment total and inLF (in labour force) total estimates. For unemployment rate and unemployment total, direct sampling variances are smoothed since the sample sizes are small for many CMA/CAs. For unemployment rate, FH model with smoothed sampling variance is used and for unemployment total, unmatched log-linear model with smoothed sampling variance is used. For inLFS total, the sample size is relatively large, so direct sampling variances are used in the unmatched log-linear model with sampling variance modeling. The model-based estimates are compared with the census estimates and the results demonstrate that the proposed models improve the direct LFS estimates substantially in terms of bias reduction and CV reduction. For unemployment rate, we also derived a simulated rate estimator based on the HB simulated unemployment total and inLFS totals from the Gibbs sampling procedure. Our results demonstrate that the direct modeling on the LFS rate performs much better than the derived HB rate based on the simulated HB total samples when the results are compared with the census estimates. Details can be found in a research report (You, 2019).

SUB-PROJECT: Small Area Estimation for Global Affairs Canada (GAC)

The task agreed upon with Global Affairs Canada was to determine the feasibility of producing annual small area estimates of unemployment rate and employment count for specific domains. These domains are defined as census metropolitan areas (CMA) by occupation (NOCS4) and also, CMA by industry (NAICS4). The experiments were conducted on data from the year 2016. If estimates obtained from small area estimation methods appear promising then this exercise could be reproduced and prove useful in non-census years. In census years, the desired rates and counts can be produced directly from the census. For these experiments, we used the Fay-Herriot model.

The goal of small area estimation is to produce a domain estimate of better quality than the direct estimate by combining Labour Force Survey (LFS) data with external information from all domains. The first step in the small area estimation process is actually to smooth direct estimates of variance. This is done by constructing a variance smoothing model which uses the identified auxiliary variables. The next step involves modelling the LFS direct estimates of the characteristics of interest. The predictions of both these models are incorporated to finally produce a small area estimate and its corresponding measure of quality.

Progress:

This project was completed and a final report was submitted to the clients (Bocci and Beaumont, 2018). Up to three auxiliary sources were used to model the survey estimates of unemployment rate and employment count. The production of small area estimates involved mainly the construction of models and the evaluation of their performance. This required the analysis of graphs and other diagnostics to determine the appropriateness of each model. Although both the unemployment rate and the employment count were investigated, only small area estimates of employment count for the domains of interest were released to the client because the diagnostics for unemployment rate were not convincing. Overall, it seemed that the small area estimates for employment count were an improvement over the direct survey estimates. It appeared that the small area estimates were somewhat affected by the quality of the auxiliary data. Furthermore, gains in using the small area estimation techniques were more apparent at the CMA-NAICS4 level than at CMA-NOC4 level.

SUB-PROJECT: Small Area Estimation of annual average and median hourly wage by economic region x occupation

The task was to determine the feasibility of producing small area estimates of annual average and median hourly wage for small domains defined as economic region by occupation. Direct estimates of annual average (or median) hourly wage along with quality measures at the desired domain levels are calculable using the Statistics Canada’s Labour Force Survey (LFS). These survey estimates are typically not of good quality for domains with small sample sizes. Thus, small area estimation techniques were considered for the desired domains. These involved modelling the LFS survey estimates using auxiliary information from Census 2016 representing a similar concept of the average (or median) hourly wage at the domain level. If estimates obtained from small area estimation methods for 2016 appear promising then this exercise could be reproduced and prove useful in non-census years. In census years, the desired averages can be produced directly from the census.

Progress:

We produced small area estimates for the annual median hourly wage for the years 2011, 2016 and 2017 and the annual average hourly wage for the years 2016 and 2017. The results of our preliminary investigation were presented somewhat formally to the clients and survey methodologists in March 2019. In addition, a document describing the principal steps in producing the small area estimates for this project was written.

SUB-PROJECT: Small Area Estimation for the Monthly Survey of Manufactures

Back in 2016, ICMIC methodologists were asked to determine the feasibility of using small area estimation techniques for the Monthly Survey of Manufactures (MSM). From those initial investigations, a small area approach was developed to produce monthly small area estimates for total sales of goods manufactured by census metropolitan area/census agglomerations (CMA/CA) x industry. Specifically, 12 CMA/CA were considered in the manufacturing industry resulting in approximately 320 domains per month. Since that time, the survey went through a redesign and a historic revision was planned in early 2019.

Progress:

With a recent redesign of the survey, the method to obtain the direct estimates of the total and the corresponding direct estimates of the variance changed. This led to a re-evaluation of the small area estimation methodology in this fiscal year. As a result, the small area estimation models and the strategy were altered. Under this new strategy, monthly small area estimates were produced for the periods 201212 to 201805. A document describing the latest strategy was written. The most important change in the strategy is that now take-none domains are estimated using the SAE system. These new estimates are synthetic.

Furthermore, we provided specifications and computer code to MSM methodologists so that these estimates could be incorporated into the same business system used to generate the direct estimates of the survey (Bocci and Beaumont, 2019). This work happened from September to December 2018. In May 2019, we re-run the small area estimates again for the periods 201212 to 201805 because the direct estimates underwent the planned historical revision.

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

References

Marhuenda, Y., Molina, I. and Morales, D. (2013). Small area estimation with spatio-temporal Fay-Herriot models. *Computational Statistics and Data Analysis*, 58, 308-325.

Petrucci, A., and Salvati, N. (2006). Small area estimation for spatial correlation in watershed erosion assessment. *Journal of Agricultural, Biological and Environmental Statistics*, 11, 169-182.

You, Y., and Zhou, Q.M. (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology*, 37, 1, 25-37. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011001/article/11445-eng.pdf.

You, Y., Rao, J.N.K. and Dick, P. (2004). Benchmarking hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, 6, 631-640.

You, Y., Rao, J.N.K. and Gambino, J. (2003). Model-based unemployment rate estimation for the Canadian Labour Force Survey: A hierarchical Bayes approach. *Survey Methodology*, 29, 1, 25-32. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2003001/article/6602-eng.pdf.

You, Y., Rao, J.N.K. and Hidiroglou, M. (2013). On the performance of self benchmarked small area estimators under the Fay-Herriot area level model. *Survey Methodology*, 39, 1, 217-229. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2013001/article/11830-eng.pdf.

1.2 Developmental Research – 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 nonunique 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. With encrypted data, the challenge is even greater.

SUB-PROJECT: Automated estimation of linkage errors

The reporting of linkage errors is an important requirement according to the agency record linkage process model (Statistics Canada, 2017) and approved guidelines for the reporting the linkage accuracy (Statistics Canada, 2019). Yet, it is still a major challenge that currently limits the automated production and use of linked data. Previous solutions include clerical reviews (Newcombe, Smith and Howe, 1983; Dasylva, Abeysundera, Akpoué, Haddou and Saidi, 2016; Dasylva, 2018) and the use of specific statistical models. Clerical reviews consist in the visual inspection of record pairs to determine their match status. They are labour intensive and inherently subjective (Newcombe et al., 1983). To avoid these issues, one may use various mixture models from the literature, including log-linear mixtures (Fellegi and Sunter, 1969; Winkler, 1988; Winkler, 1993) with or without the assumption of conditional independence or mixtures involving transformed normals (Belin and Rubin, 1995). Unfortunately, these models have limitations. Indeed, assuming conditional independence (Fellegi and Sunter, 1969) leads to biased estimators because this condition is seldom met in practice. Also, log-linear mixtures with interactions (Winkler, 1988; Winkler, 1993; Thibaudeau, 1993) are not generally known to have the identification property. A model without this property leads to biased estimators regardless of the sample size. Finally, mixtures with transformed normals attempt to estimate linkage errors by modeling the distribution of a pair (probabilistic) linkage weight (Belin and Rubin, 1995). However, it requires a good separation between the matched weight distribution and the unmatched one. Besides, all previous models suffer from their focus on in a single pair. Thus, they cannot yield measures of linkage error at the record level, e.g., whether a record with many adjacent links has one true positive link.

Progress:

A new statistical model is proposed for the automated estimation of linkage errors when linking two complete and duplicate-free registers of a large iid population with N individuals. It is based on the concept of *neighbour* of a given record, which is generally defined as another record such that the resulting record pair meets some criterion, for example the presence of a link by some linkage method regardless of whether it is deterministic, hierarchical, probabilistic, based on machine learning (Supervised machine learning solutions can estimate linkage errors through cross-validation. However, these estimates are often unreliable because of errors in the training data.) or otherwise. Intuitively, the distribution of the number of neighbours can provide much information about the occurrence of linkage errors. Indeed, in the situation that is considered, this connection is summarized in the following table.

Table 1
Number of neighbours and linkage errors

Number of neighbours (n_i)	False negatives	False positives
0	1	0
1	?	?
Many	?	$\geq n_i - 1$

A model is used to predict the numbers of false positives and false negatives when they are not fully determined by the number of neighbours n_i.

The model parameters are related to the targeted error measures (Statistics Canada, 2019), when linking neighbour records without any conflict resolution.

The new model provides many advantages. It accurately estimates the error rates for *any* linkage method, whether probabilistic or nonprobabilistic (deterministic, hierarchical, machine learning, etc.). It can also provide accurate estimates of probabilistic linkage weights, which lead to automated linkage decisions (i.e., a single threshold and no grey zone if using the probabilistic method) that are *optimal* among *all possible decisions by any linkage method.* Additionally, the new model is easy to use because it is oblivious of the interactions among the linkage variables, even if specifying interactions may yield more efficient estimators. Further benefits include the ability to estimate the false negatives due to blocking (define a neighbour solely from the blocking criteria), and record-level estimates of linkage errors.

The model has been successfully applied in simulations and in an empirical study with administrative data. Further details can be found in the working paper (Dasylva, Goussanou, Ajavon and Abousaleh, 2019).

The new methodology is being modified to obtain more precise estimators by taking a larger estimation sample and to accomodate situations that involve two or more files that have some duplicates and some undercoverage, including the resolution of conflicts.

SUB-PROJECT: Automated coding using a record linkage methodology

Coding consists in assigning the right code to a field from a finite code set, according to some input. It is a classification problem and an essential part of the editing task when working with census, survey or administrative files. Some solutions use a nearest neighbour approach by linking each new observation to a subset of coded observations (Wenzowski, 1988; Chu, Yeung and Dasylva, 2018). This subset represents a pool of donors from which one is selected to assign the corresponding code to the new observation. This methodology requires the estimation of the coding error and the setting of the linkage parameters according to the target error rate.

Progress:

A methodology has been described for estimating the rate of coding error and for setting the underlying parameters according to a target error rate (Dasylva, 2019). In simulations, it performed well, judging by the agreement between the target error rate and the achieved rate (Savard, 2019). The methodology was also applied in an empirical study using the 2016 census mother tongue data, where it enhanced the solution by Chu et al. (2018) by setting the linkage parameters according to the target error rate. As in the simulations, the achieved error rate was close to the target.

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

References

Belin, T.R., and Rubin, D.B. (1995). A method for calibrating false-match rates in record linkage. *Journal of the American Statistical Association*, 90, 430, 694-707.

Dasylva, A., Abeysundera, M., Akpoue, B., Haddou, M. and Saidi, A. (2016). Measuring the quality of a probabilistic linkage through clerical reviews. Proceedings: Symposium 2016, Growth in Statistical Information: Challenges and Benefits, Statistics Canada.

Fellegi, I.P., and Sunter, A.B. (1969). A theory of record linkage. *JASA*, 64, 1183-1210.

Newcombe, H., Smith, M. and Howe, G. (1983). Reliability of computerized versus manual death searches in a study of the health of Eldorado uranium workers. *Computers in Biology and Medecine*, 13, 157-169.

Statistics Canada (2017). *Record Linkage Project Process Model,* Catalogue No. 12-605-X, Statistics Canada.

Statistics Canada (2019). *Guidelines for Reporting the Record Linkage Accuracy*, Statistics Canada.

Thibaudeau, Y. (1993). The discrimination power of dependency structures in record linkage. *Survey Methodology*, 19, 1, 31-38. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1993001/article/14477-eng.pdf.

Wenzowski, M.J. (1988). ACTR A generalised automated coding system. *Survey Methodology*, 14, 2, 299-307. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/1988002/article/14586-eng.pdf.

Winkler, W.E. (1988). Using the EM algorithm for Weight Computation in the Fellegi-Sunter Model of Record Linkage. *Proceedings of the Section on Survey Research Methods*, American Statistical Association, 667-671.

Winkler, W.E. (1993). Improved decision rules in the Fellegi-Sunter Model of Record Linkage. JSM 1993: In Proceedings of the 1993 Joint Statistical Meetings, Survey Research Methods Section, August 8–12 1993, San Francisco, CA. Alexandria VA: ASA; 274-279.

1.3 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

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, updated and delivered training presentation in various forums and met with international delegates to discuss current and future development of the generalized systems. The group met with delegations from Italy, Japan, Singapore and Ireland.

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

Progress:

The Banff team led the organization of a hackathon to encourage exploration of new imputation methods. The event consisted of a dedicated day where participants explored a dataset and were encouraged to find and apply imputation methods to treat non-response. The hackathon was titled “Can you beat BANFF” and ultimately the proposed methods were compared with those available from the BANFF procedures. In the end, a number of methods were identified and the most promising method was missing forest, available through the R-package *missForest*.

Subsequently, in order to further increase the functionality of the Banff Processor, the Banff development team demonstrated how external programs can be integrated into the Banff framework. Given the availability and popularity of external, open-source data editing programs, these may be appealing options for current and future users. Specifically, a workflow containing both standard Banff procedures and an imputation step using the R-package *missForest* was developed and successfully applied in a testing environment.

The use of R in general as a research and production tool was also explored. The results of this exploration were documented and discussed both with Statistics Canada’s Advisory

Committee on Statistical Methods and at an international conference on the use of R for Official Statistics (Fortier and Thomas, 2018).

SUB-PROJECT: Generalized System development – Statistical Data Editing Framework

Progress:

The Banff edit and imputation system consists of nine custom SAS procedures performing various statistical data editing (edit and imputation) tasks. Built into these procedures is a data editing framework that allows information outputs from one procedure to act as inputs into another, allowing for complex data editing workflows, facilitated by the Banff Processor. This work was presented at the United Nations Economic Commission for Europe (UNECE) Workshop on Statistical Data Editing (Neuchatel, Switzerland, September 2018).

Additionally, the Banff research and development team contributed to version 2.0 of the Generic Statistical Data Editing Model (GSDEM) released June 2019 (https://statswiki.unece.org/plugins/servlet/mobile?contentId=117771706#content/view/117771706), intended as a reference for all official statisticians whose activities include data editing.

SUB-PROJECT: Generalized System development – release of G-Sam 1.02.001

Progress:

A new version of G-Sam was developed, tested and released. Version 1.02.001 includes a partial restructuring and recoding of all three modules (Stratification, Allocation and Selection) to improve efficiency, and to facilitate technical and methodological support. Other improvements include the simplification of user inputs and improvements to the SAS log messages generated by G-Sam, to better serve user needs.

SUB-PROJECT: Generalized System development – Release of G-EST 2.02

Progress:

A new version of G-EST was developed, tested and released. Major changes for version 2.02 included improved performance for estimating sampling variance using Taylor linearization, the management of questionnaire skip patterns, a bug fix related to the estimation of sampling variance in presence of nonresponse and some methodological changes to the calibration exclusions. A second release (version 2.02.008) reintroduced parallel processing for sampling variance and a new version (3.2.1) of SEVANI.

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

1.4 Prospective Research – Data Integration

The advent of the World Wide Web in the 1990s opened the door to new modes of information collection for surveys, namely large opt-in Web panels and Big Data. *Opt-in Web panels* are composed of individuals who use the Web regularly and who are asked questions on various topics. *Big Data* is a generic term for data sets so large or complex that the capabilities of traditional data processing applications are inadequate. Web panels as well as Big Data often do not use probability-based sampling designs.

This project has three main objectives:

i) To assess the possibility of using sample matching or other data integration techniques for some of the programs of Statistics Canada in order to reduce respondent burden and/or data collection costs.
ii) To develop or adapt new methods to solve practical issues.
iii) To develop and test prototypes that implement most promising methods.

SUB-PROJECT: Review of data integration methods

Statistics Canada has recently embarked into a modernization phase. One key component of the different modernization initiatives is to make greater use of different data sources by combining them. This topic is known as data integration in the statistical literature. The goal of this project was to make an extensive literature review on data integration statistical methods.

Progress:

The literature review was conducted in the summer 2018. 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 and inverse propensity score weighting. This review was presented at the Colloque francophone sur les sondages in Lyon and a paper was written and submitted for publication (Beaumont, 2019).

SUB-PROJECT: Using regression trees to weight a nonprobability sample

Data from nonprobability sources, such as Web panel data, are known to suffer from selection bias. One possible approach to deal with selection bias is to model the selection probability in the nonprobability sample and weight each sample unit by the inverse of this probability. This method is known as inverse propensity score weighting. It is particularly useful when the sample contains many variables of interest as a single weighting strategy can be applied to all these variables.

If auxiliary variables were known for the whole population then the problem would be basically identical to weighting for nonresponse in a probability survey. However, most of the time, such information is not available. To tackle this issue, Chen, Li and Wu (2019) proposed a fully parametric method to combine the nonprobability sample with a probability sample that contains the values of all the auxiliary variables. If the set of useful auxiliary variables is given with the appropriate interactions then their method can be applied directly. The real challenge is to choose these auxiliary variables and interactions.

Regression trees are now becoming popular as a nonparametric method that can automatically choose the auxiliary variables and handle interactions in standard regression problems. The goal of this project is to extend regression trees to the context of combining a probability and nonprobability sample and to write an R program that implements the method.

Progress:

We developed an extension of regression trees to the context of combining nonprobability and probability samples. We also developed and R program that implements the method and evaluated it through a small simulation study. Preliminary results look promising. The results of this project were presented at the 2019 SSC meeting (Chu and Beaumont, 2019)

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

SUB-PROJECT: Analysis of categorical data obtained by probabilistic linkage

Description:

Under the assumption that the only source of linkage error is false positives the method of Chipperfield, Bishop and Campbell (2011) of making inference with binary data in using logistic regression was applied and a simulation study examined the effectiveness of the proposed method. Results showed that the estimators are unbiased and exhibit smaller variance than those obtained from clerically reviewed data. A set of SAS macros were developed to analyzing binary data by logistic regression.

This approach will be generalized to categorical data, contingency tables data and count data and represent a valuable solution when integrating by record linkage multiple sources of data.

The objective of this research is to develop a SAS prototype for analysing linked categorical data in using logistic model, log-linear model or Poisson model.

Progress:

A SAS macro was developed and applied to analyzing binary data by logistic regression in the solely presence of unlinked records (false negatives). A pseudo-likelihood approach which assigns weights to linked records is applied. “Non-link” weighting adjustment is effective if the weighting is related to the variables that influence the probability of being an unlinked record and to the variable being studied.

Results were presented at the FCSM conference and the CANSSI conference held in Montreal by Saïdi, Chu, Dasylva and Labrecque-Synnott (2018).

SUB-PROJECT: E-M estimation of probabilistic linkage parameters with a dependence assumption between record fields

Description:

It is well known that failure to comply with the conditional independence assumption of the Fellegi-Sunter (F-S) model generally produces biased m (linked probabilities) and u (unlinked probabilities) distribution estimates. This limitation impacts the classification process (determining of thresholds and optimality of the F-S decision rule) and error levels for false positives and false negatives. The E-M algorithm with interactions between a record’s fields estimates linkage parameters and produces substantive improvements in linkage effectiveness.

The objective of this study is to implement the E-M algorithm in G-Link without a conditional independence assumption to calculate probabilistic linkage parameters. We will be working to improve the prototype developed in 2015 by Dasylva et al.

Progress:

SAS macros were developed in 2015 by Abel Dasylva et al. to compute E-M weight using log-linear models in PROC CATMOD with interaction terms and optional clerical data under the assumption of missing outcomes omitted (under a missing at random model). Authors tested using data generated from theoretical models.

The team mapped and documented Input/output/intermediate files from both algorithms and futher investigated the technical requirements for implementation in G-Link. The research goals of this project is to transform G-Link datasets for new macros and test macros with real data in comparing with “true” linkage status using known unique identifiers. We also will study the case where clerical data are available, in specifying interactions for new model, bounds on the parameters and output model diagnostics.

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

References

Chen, Y., Li, P. and Wu, C. (2019). Doubly robust inference with non-probability survey samples. Unpublished manuscript.

Chipperfield, J.O., Bishop, G.R. and Campbell, P. (2011). Maximum likelihood estimation for contingency tables and logistic regression with incorrectly linked data. *Survey Methodology*, 37, 1, 13-24. Paper available at https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2011001/article/11444-eng.pdf.

1.5 Prospective Research – Non-probabilistic approaches

SUB-PROJECT: Experiments with nonprobability data sources

We have acquired data from a nonprobability panel of volunteers which contains variables similar to a few variables in the Canadian Community Health Survey (CCHS). Panel members are asked to respond to a short questionnaire via an application on the cell phone. In return, they get rewards from their preferred program. It is well known that estimates derived directly from panel data are subject to selection bias. These nonprobability data offer an opportunity for investigating the effectiveness of data integration techniques at reducing selection bias.

Progress:

We evaluated two techniques: statistical matching, also known as sample matching (Rivers, 2007) and model-dependent calibration. The success of both methods depends on the strength of auxiliary variables available in both sources. In our experiment, we used the following variables: age, sex, education, marital status and health region. The estimates obtained using statistical matching and calibration were then compared with CCHS estimates. We observed that both methods reduced the substantial bias observed with the direct panel estimates. Statistical matching seemed to be slightly more effective than calibration. This might be due to its nonparametric nature unlike calibration which relies on a linear model with missing interactions. However, a nonnegligible bias persisted. Two reasons might explain its presence: i) auxiliary variables used were not strong enough predictors of the health variables of interest and ii) measurement errors were likely present in the panel data. Our results were presented at the SSC and CANSSI conferences in the summer 2018. A report describing our findings was written and published in the SSC proceedings (Chatrchi, Beaumont, Gambino and Haziza, 2018).

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

Reference

Rivers, D. (2007). Sampling for Web Surveys. *Proceeding of the Joint Statistical Meeting*, Salt Lake City, Utah.

1.6 Prospective Research – Data Science

As part of the 2018 International Methodology Symposium, Methodology Branch organised a panel on “Data science and machine learning for official statistics – successes, opportunities and challenges”.

In 2018, the Methodology and Analytical Study Branches launched a Data Science Centre of Excellence (DSCoE) on behalf of the sector. The DSCoE is hosted in Methodology and operates in close collaboration with the various other data science agents and groups in the organisation.

The main role of the Data Science Centre of Excellence is to offer the scientific backbone needed for the data science activities. The team working in the DSCoE is involved in:

- Providing consultation and guidance on the application of advanced data science methods in the context of data creation, processing and analysis for official statistics;
- Directly contribute to data science pilot or exploratory projects, in particular those involving advanced techniques related to data processing and analysis (i.e., predictive analytics);
- Leading and conducting research related to support and advance the applications of DS for the production of official statistics;
- Evaluating or providing training activities to advance capacity in DS skills across the Agency;
- Providing support and advice on the ethics component of AI related work.

One of the joint projects included an in-depth review of machine learning teachniques for automatic classification, with 4 sub-topics: unbalanced training data sets; transfer learning, hyperparameter tuning and quality assessment. The discussions and results were presented at the Advisory Committee on Statistical Methods (Yeung, Chu, Laroche and Fortier, 2018).

Further use of data science tools and techniques are reported elsewhere in this report, within the relevant subtopics.

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

1.7 Divisional Research

SUB-PROJECT: Re-evaluation of the regression composite estimator for the Labour Force Survey

The Labour Force Survey started using the regression composite estimator nearly 20 years ago after extensive analysis. While it has been very effective at reducing the variance of month-to-month and level estimates, recent preliminary analysis shows that it may be resulting in an unexpectedly large bias in some estimates. The source of this bias is not immediately obvious, but it may be connected to the drift problem discussed in Fuller-Rao (2001).

The goal of the project is to investigate possible adjustments to the regression composite estimator, including re-evaluating the combining factor used in combining two different composite estimators. This factor was set based on empirical studies done by Gambino, Kennedy and Singh (2001) and repeating the study with more recent data may be informative. As well we would like to investigate adjustments to the composite control total categories, as some of the categories contain very few respondents and should be collapsed.

Progress:

The key papers on the regression composite estimator by Fuller and Rao (2001), Gambino et al. (2001) and related references in those papers were extensively reviewed. A review of a more recent paper by Preston (2015) was also done. Monthly estimates covering the period of January 1997 to December 2018 using the generalized regression method were compared with estimates from the regression composite estimation method for a large number of LFS series. Preliminary results indicate that there are some differences between the estimates of these two methods. The differences are more pronounced in some series than others. At this point it is not clear what could be the cause of these differences. It is believed that these differences may be explained by the theoretical properties of the two methods and/or impact of nonresponse and/or the choice of the number of calibration control variables (and totals). Work is under way to formally identify and explain the potential cause(s) of these differences and then at a future time investigate potential solution(s) where feasible.

For further information, please contact:
Emmanuel Benhin (613-862-7638, emmanuel.benhin@canada.ca).

SUB-PROJECT: Reporting Quality using Confidence Intervals

In 2017, the Methods and Standards Committee (MSC) approved our recommendation to adopt as a best practice the use of confidence intervals for measuring and reporting the quality of estimates. The purpose of the project is to conduct research to support the use of confidence intervals for reporting quality.

Progress:

During the period, a set of rules was developed for deciding whether an estimate and its confidence interval could be released or whether they should be suppressed for quality reasons. The proposed rules for social surveys were presented to the Social Statistics Methods Division (SSMD) Technical Committee (Neusy, Boulet, Duggan, Mach, Mantel and Reedman, 2018).

Standard text for user guides was written for surveys that decide to use confidence intervals for reporting quality. The text was used for the Study on International Money Transfers.

A presentation (Neusy, 2019a) was given to the Special Surveys Division managers on the problems with using CVs and the advantages of using confidence intervals to report quality.

An internal note documenting the derivation of the Wilson interval for weighted counts was written (Neusy, 2019b).

For further information, please contact:
Elisabeth Neusy (613-863-3513, elisabeth.neusy@canada.ca).

SUB-PROJECT: Validation of an extension to the Rao-Wu bootstrap variance estimator proposed by Pérez-Duarte, applicable to multi-stage sampling, with probability proportional to size sampling at the first stage using a Monte Carlo simulation

The variance estimator from Rao-Wu's bootstrap, also called "rescaled bootstrap", is used by several Statistics Canada surveys. This estimator has been proven on samples from one-stage stratified sampling designs, where units are sampled using a simple random draw. For multi-stage sampling, the Rao-Wu estimator is effective if the first stage is sampled using a simple random draw at all stages and if the sampling fraction is small. For surveys that use a multi-stage sample or where the first stage is sampled using a draw proportional to size with replacement, an extension of the Rao-Wu estimator exists.

Osiewicz and Pérez-Duarte (2012) propose a modification to the extension of the Rao-Wu estimator that would estimate the sampling variances of estimators from sampling designs proportional to size sampling without replacement and two-stage sampling designs, where the first-stage draw is without replacement and is proportional to size sampling using the Rao-Hartley-Cochran (RHC) method and the second stage is a simple random selection without replacement.

The purpose of this research is to validate this extension using a Monte Carlo simulation.

Progress:

For the purposes of this study, eight sampling designs were tested. All designs use the same population and two-stage sampling, where first-stage sampling is proportional to size without replacement of RHC and second-stage sampling is simple random sampling. Only the first- and second-stage survey fractions vary. The Monte Carlo expectation of the CV of the estimator for a total of a two-stage design using RHC PPS sampling (Rao, Hartley and Cochran, 1962) as well as the Monte Carlo expectation of the CV of the variance estimator proposed by Osiewicz and Pérez-Duarte (2012) were produced. To date, the CV of the variance estimator has skyrocketed. The reasons for this increase have not yet been explored. An internal article describing the research and the results to date has been written (Devin, 2018). SAS programs are also available.

For further information, please contact:
Nancy Devin (613-618-1027, nancy.devin@canada.ca).

SUB-PROJECT: Electricity meter data for Manitoba

This is a project concerning monthly data from electrical meters in Manitoba. It is a big data project which will allow for an unprecedented understanding of electricity usage within Manitoba. An objective of this project is to produce quarterly total estimates of residential electrical energy consumption for Manitoba's dissemination areas.

Progress:

In November 2018, a presentation was given at the Statistics Canada Methodology Symposium (Duddek, 2018), and the corresponding paper was submitted for the proceedings. This was a summary of the work done to aggregate the hydro data to 2016 Census geography and then evaluate coverage, amongst other objectives.

Smart meter data were also studied as part of the evaluation of an imputation strategy for the Household and Environment Survey (HES). As part of this study, meter data were used jointly with heating degree day data from the Environment Canada website. This enabled the development of a meter level regression model tying quarterly electrical consumption to the weather. We were then able to implement outlier detection and treatment to deal especially with some unusual under-coverage of measured kWh for the first quarter of 2015.

This work was presented to the Technical Committee on Household Surveys on March 1, 2019 and entitled "Using Manitoba Hydro data to evaluate imputation of electricity consumption in the Households and the Environment Survey" (Duddek, 2019).

For further information, please contact:
Christopher Duddek (613-862-9234, christopher.duddek@canada.ca).

SUB-PROJECT: Exploring the use of Neural Networks to Automatically Classify Logos on SHS Shopping Receipts

This research project plans to assess the feasibility of using coded receipts from past cycles of the Survey of Household Spending (SHS) to train a neural network algorithm to automatically classify the logos appearing on these receipts by store names. Currently, relevant information on SHS receipts such as store name, items purchased, date, purchase total, etc., is all captured manually. Alternatively, methods of automated capture could potentially save time, resources and budget. This project focuses on automatically extracting the store name based on its logo. Two strategies will be tested using neural networks, which are a type of machine learning algorithms. First, we will attempt using transfer learning techniques by taking advantage of pre-trained neural networks algorithms. Then, a custom convolutional neural network algorithm will be developed. The training set created for this project will be derived from the SHS databases and the scanned SHS receipts. A cropping technique will be developed to extract store logos from the receipts and a label will be assigned to them based on its coded store name found in the SHS database. Finally, after training the algorithms the results of the two strategies will be assessed using a test set and metrics such as accuracy, sensibility and specificity. A conclusion on the feasibility of applying such techniques to automate the receipt capture process in the future will be given.

Progress:

- Coded receipts from SHS 2015, 2016 and 2017 were used to build the training, validation and test set. Only receipts from the Top 20 stores were kept. The Top 20 stores were chosen based on frequency of receipts from the store and the fact that they had an image based logo. In total, we had over 40,000 labelled logos which we assigned in a 60%, 20%, 20% split into a training, a validation and a test set respectively.

- A logo cropping algorithm was developed to automatically crop the logo at the top of the receipt.
- A technique using a pre-trained neural network was attempted but was overfitting, meaning the algorithm was very good at predicting the store name for the test set but was very bad at predicting the store names for new receipts. This technique was quickly dropped since we did not believe it would return results of good quality.
- A custom convolutional neural network (CNN) was built and trained. This algorithm correctly predicted the store name from the Top 20 receipts in our test set with an accuracy of over 95%. The errors made by the algorithm were minimal and were mostly due to misclassification during the manual coding or receipts of poor quality.
- The use of a custom CNN for automatically classifying shopping receipts logos was deemed successful and demonstrates promising first results towards automating the capture of relevant information on SHS receipt.
- The project was documented and presented to various groups (Lee, 2018a), (Lee, 2018b), (Mayer, 2019a) (Mayer, 2019b).

For further information, please contact:
Émilie Mayer (613-220-1138, emilie.mayer@canada.ca).

SUB-PROJECT: Develop quality indicators to measure price index accuracy

In price indexes, it is difficult to measure the accuracy of estimates for various reasons. Often, index estimates are produced in a context somewhat removed from the traditional survey methodology either because of the survey design used or alternative data sources. In this context, it is not always possible to directly calculate traditional indicators such as the CV. However, there is a demand for this type of information either to inform users or to support internal planning. For producer price indexes and the Consumer Price Index, we want to develop quality indicators based on the estimation context. We already produce CVs and confidence intervals specific to the design of certain indexes. We want to adapt this approach to other indexes and develop new indicators in cases where the current method does not apply. Our objectives are to conduct a more extensive literature review on what already exists and to further explore and reflect on the new indicators to develop.

Progress:

In fiscal year 2018/2019, we completed a literature review that comprised an overview of different quality frameworks used to define the quality of price indexes. The review covered the frameworks used by Statistics Canada and by various statistical agencies, such as the Bureau of Labor Statistics and Eurostat.

Particular attention was given to indicators related to index accuracy. In this context, we documented different cases in which bootstrap is used to estimate sampling variance for price indexes using a probability-proportional-to-size design. Variance was calculated using the approach proposed by Beaumont and Patak (2012). This probabilistic variance estimation method was also applied to a price index based on a sample drawn from a non-probabilistic design. For the latter, we assumed different designs and documented the required assumptions.

We also delved into variance estimation of the model. This variance has a different interpretation than sampling variance and is of interest in a context in which price samples come from a non-probabilistic design. We documented the process of model variance estimation using the approach proposed by Zhang (2010). The proposed approach was applied to the Retail Services Price Index. A series of indicators was also developed to evaluate index accuracy for a series of domains. For response rate use, we conducted empirical simulations to evaluate the risk of bias and establish acceptability thresholds.

For further information, please contact:
Justin Francis (613-863-0276, justin.francis@canada.ca)
Jean-Sébastien Provençal (613-513-9441, jean-sebastien.provencal@canada.ca).

SUB-PROJECT: Development of a Robust Estimation Prototype

In traditional estimation procedures, the domain estimates may be negatively affected by the presence of influential units in the sample. Various authors have developed robust estimation methods to attenuate the instability of the estimates. In this project, we look to develop a SAS program to implement methods based on minimizing the maximum estimated influence a unit can have on the robust estimate (Beaumont, Haziza and Ruiz-Gazen, 2013). The influence is measured through the conditional bias. This program will allow us to study the effectiveness of the methods by applying them to some of our surveys.

Progress:

We have created a set of SAS macros to handle a variety of functions associated with domain robust estimation of totals under stratified simple random sampling without replacement. It includes the following functions that can be run individually or from top to bottom for this design: (1) calculation of design weights; (2) calculation of calibration weights with auxiliary information for one or two partitions of the population; (3) production of domain robust estimates of totals; (4) creation of coherent domain estimates to account for constraints on the variables; (5) calculation of recalibrated weights to ensure that standard weighted estimates are identical to robust estimates; (6) non-robust estimation of domain totals with associated variance estimates.

Some functions such as the calibration weights function and the coherence estimation function, have many options and features to meet a variety of requirements.

Every function can be run on its own by providing the necessary inputs. To make it easier to run the programs, we have created validation structures to check the inputs to each program and provide notes, warnings and errors.

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

References

Beaumont, J.-F., and Patak, Z. (2012). On the Generalized Bootstrap for Sample Surveys with Special Attention to Poisson Sampling. *International Statistical Review*, 80(1), 127-148.

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

Fuller, W., and Rao, J.N.K. (2001). A Regression Composite Estimator with Application to the Canadian Labour Force Survey.

Gambino, J., Kennedy, B. and Singh, M. (2001). Regression Composite Estimation for the Canadian Labour Force Survey: Evaluation and Implementation.

Osiewicz, M., and Pérez-Duarte, S. (2012). Flexible variance estimation in complex sample surveys: Rescaled bootstrap in multistage, pps surveys – DRAFT.

Preston, J. (2015). Modified regression estimator for repeated business surveys with changing survey frames. *Survey Methodology*, 41, 1, 79-97. Paper available at https://www150.statcan.gc.ca/n1/pub/12-001-x/2015001/article/14160-eng.pdf.

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of Unequal Probability Sampling without replacement.

Zhang, L.-C. (2010). A model-based approach to variance estimation for fixed weights and chained price indices. *Official Statistics in Honor of Daniel Thorburn*, 149-166.

2 Support Activities

2.1 Confidentiality and Disclosure control

The team provided support and advice on the application of disclosure control techniques both internally and to answer specific external requests.

During this fiscal year Random Tabular Adjustment (RTA) was used for the first time in published Statistics Canada tables. Statistics Canada released a complete set of output tables for the Survey of Innovation and Business Strategy (SIBS) using RTA to protect the confidentiality of contributed values. To accompany this release, a blog titled "Random Tabular Adjustment is here!" and other communication material were prepared to herald this new approach to confidentiality. The blog included a link to the technical presentation of the method (Stinner, 2017). The method was also tested with several different surveys both as a proof-of-concept to subject matter areas and also to identify issues associated with the practical implementation of RTA.

The team also explored the use of data synthesis with practical implementation for the creation of a dataset to be used in a data analysis open competition. The work and the method are documented in Sallier and Girard (2018). Finally, the team started to explore the differential privacy framework (Girard, 2019).

For further information, please contact:
Steven Thomas (613-882-0851; steven.thomas@canada.ca).

Reference

Stinner, M. (2017). Disclosure Control and Random Tabular Adjustment. Proceeding of the 2017 Annual Meeting of the Statistical Society of Canada, Winnipeg.

2.2 Record Linkage Resource Centre (RLRC)

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

Progress:

We continued to support the development team of G-Link and worked jointly to potential sources of current/past fixes/bugs improvements for G-Link. The RLRC also provided support to internal and external G-Link users who sought help or provided comments or suggestions.

In order to improve the Fellegi and Sunter classification and to reduce the burden of manuel review in the version 3.4 of G-Link, much of methodology's work revolved around the development and documentation of automatic definition of thresholds with a graphical approach, improved import, export and batch functionalities, a complete and user-friendly interface and extensive performance improvements. Additionally, specifications and programs were delivered to incorporate newer and faster MixMatch comparison rules as well as specifications for new tools for users. RLRC also worked on a variety of other record linkage-related projects during the year, including the tutorial, and the user guide for the new version 3.4 of G-Link.

The inventory of record linkages done by the methodology Branch was updated in 2018 and the results were presented.

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

2.3 Time Series Research and Analysis Centre (TSRAC)

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.

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

- Consultation and training in Time Series (including course development and delivery);
- Support and Enhancement of Time Series Processing System;
- Seasonal Adjustment and Trend-Cycle Estimation;
- Support to G-Series for Benchmarking and Reconciliation;
- Modelling and Forecasting.

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 in various situations (System of National Accounts, local level estimates, new Labor Statistics Division (LSD) surveys, etc.) and specific contexts for benchmarking and reconciliation. In addition, formal and informal exchanges occur with other statistical organisations (Bureau of Economic Analysis, United States Census Bureau, Bureau of Labor Statistics, Eurostat, etc.) academic organisations (University of Waterloo, University of Ottawa) to collaborate and provide input on current topics.

Progress:

The Guidelines on Time Series Continuity were finalised and approved by the Methods and Standards Committee (Statistics Canada, 2019a, 2019b).

Experts from the center provided consultation services as needed, including on volatility measures, seasonally adjustment of flow series, and calibration of import and export series. Exisiting courses were delivered as needed. The newly updated course on Statistical Modeling and Forecasting was offered for a second time.

The impact on seasonal adjustment of disclosure control methods based on perturbations was investigated and suggestions have been made to include an autocorrelation component in the random tabular adjustment method.

Recent research and exploration work have been communicated with peers (Verret, 2018 and Matthews, 2018).

Support and Enhancement of Time Series Processing System

The theme includes continued development and support for the time series processing system, currently used in production applications throughout the agency.

Progress:

A prototype of Seasonal Adjustment Dashboard was completed in R-shiny and will soon be deployed in pilot mode to a few users. This dashboard facilitates the interpretation of seasonally adjusted data and quickly provides key statistics related to this process.

A module has been developed and documented for stock calibration. A strategy for detecting file corruption prior to the execution of trend-cycle retrieval modules has been developed.

Seasonal Adjustment and Trend-Cycle Estimation

This theme covers analysis and evaluation of new methods and techniques for seasonal adjustment. This includes comparisons between model-based methods and X-12-ARIMA. A longer-term goal of this research is to develop methods suitable for adjustment of high-frequency (daily, hourly, etc.) series. Also included is exploration of Machine Learning to select options to specify in seasonal adjustment, intended to streamline the support of production processes, and increase capacity to quickly produce high volume series.

Progress:

A comparison of the X-12-ARIMA and SEATS methods has been documented (Matthews and Dochitoiu, 2018). Preliminary work on the evaluation of seasonality tests, and the potential development of a non-parametric test are underway Mischler, Lapointe and Patak (2019). Other research projects are underway, including the development of a strategy to produce longitudinal bootstrap samples in order to estimate seasonally adjusted variance by replicates (Verret, 2019); on GARCH models to look at seasonal heteroscedasticity and derive measures of irregularity (in collaboration with an academic researcher) and a strategy to develop preliminary short-run seasonal adjustment.

As part of the project to explore the use of machine learning methods for the selection of seasonal adjustment options, a set of training data was collected and prepared. An objective approach to assess the impact on quality has been determined. The first tests were done on the filter length parameter and the early results are promising.

Support to G-Series for benchmarking and reconciliation

Description:

Since the release of G-SERIES 2.0, the focus of this work turns to support, and future development. Support is offered for users for issues and questions related to the methodology.

New applications of the new functionality are built and tested as identified. Research and plans for the next release will be documented.

Progress:

The team reviewed conditional benchmarking approach proposed by an advanced user and provided comments on working paper. This approach will be evaluated through application to real data, and considered for implementation in a future release of G_SERIES. Improvements have been made to the balancing macro to counter an identified limitation of SAS 9.4.

Modeling and forecasting

Description:

This theme includes continued and expanded use of SAS HPF as a forecasting tool for general use, for identification of breaks in series, and for other modelling applications, as well as investigation into other methods and tools available in SAS and other third-party software packages.

Progress:

- The team conducted modelling of Tourism Data as a proof of concept to produce advance economic indicators and presented findings at Statistics Canada's Methodology Symposium (Patak and Armenski, 2018). This work assessed several classes of time series models, utility of predictive components, automation of model selection and accuracy of forecasts at different horizons.
- The team also undertook work to model export data to mitigate lack of data from an external source (US government shutdown in December of 2018). Month-ahead forecasts were generated for temporary internal use in absence of data from the administrative source affected by the shutdown. Forecasts and quality indicators were generated, and consultation on their use and interpretation was given (Leung and Dochitoiu, 2018).
- The team conducted a literature review to identify models commonly used in now casting models from various organizations and available tools for their application. Development of practical guidelines document is underway in the context of near-real time estimation exploratory project at Statistics Canada.
- Development of state space models was advanced by the team in collaboration with a professor from the University of Ottawa. The work consisted of evaluation and modification of the basic structural model for seasonal adjustment, and a presentation was prepared for International Symposium on Forecasting (to be presented June 2019).
- Additional development of state space modelling applications was made in the context of small area estimation (Matthews and Dochitoiu, 2018) to include time series

components and estimation of sample rotation effects for panel designs (Dochitoiu, 2018) were developed to be further evaluated.

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

2.4 Quality Secretariat

SUB-PROJECT: Update of Quality Guidelines document

Initiative to update the Quality Guidelines with the following objective: a) to provide all other data producers in the Canadian National Statistical System with a relevant reference document; b) to adapt to the new reality of administrative data by covering the main statistical business processes; and c) to facilitate compliance with current quality assurance methods.

Progress:

A redaction plan was presented to and approved by Statistics Canada's Methods and Standards Committee. A reading committee made comments and suggestions on the first draft followed by targeted consultations conducted among some divisions in Statistics Canada in order to gather feedback on specific fields of expertise. Comments have been integrated to the draft of the new edition, which will also include items related to privacy, transparency and ethics. The new editon will be released in 2019-2020.

SUB-PROJECT: Research work and capacity building with international partners

The Quality Secretariat provided advice and capacity building for international partners, with a focus on providing a high-level overview of Statistics Canada's quality management practices and official documents related to quality (i.e., the Quality Assurance Framework and Quality Guidelines).

Progress:

In 2018-19, we gave presentations to delegations from Indonesia and Cameroon and led a one-week workshop in Santo Domingo, Dominican Republic in our meetings with international visitors, we gave presentations centered on their topics of interest; namely, quality management and the Generic Statistical Business Process Model. In Santo Domingo, our workshop assisted participants from over a dozen Latin American countries in the drafting of a roadmap to build a Quality Assurance Framework. During each of these collaborations, we also participated in discussions about modern challenges facing data quality at National Statistical Offices, including the issue of reporting quality for statistical data that are not collected from sample surveys.

Research on the development of new measures or quality indicators in a mixed data context has also begun Reedman (2018) and Reedman and Windross (2018).

SUB-PROJECT: Participation in the United Nations Expert Group on National Quality Assurance Frameworks

The Quality Secretariat served as the co-chair of the United Nations Expert Group on National Quality Assurance Frameworks, and contributed to the drafting of an updated National Quality Assurance Framework by sharing content from our internal Quality Assurance Framework.

Progress:

A draft of the updated United Nations National Quality Assurance Framework (NQAF) was completed and released for peer review in November 2018. Recommendations were subsequently implemented, and this version was adopted at the UN's Statistical Commission in New York in March 2019. The Quality Secretariat served as a co-chair in the commission's side event to discuss the NQAF.

For further information, please contact:
Ryan Chepita (613-851-5340, ryan.chepita@canada.ca).

2.5 Data Analysis Resource Centre

The Data Analysis Resource Centre (DARC) is a team of statistical consultants and researchers within the Methodology Branch. The main goals of DARC are 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 was offered as requested by various clients. Numerous consultation services were provided to Statistics Canada's analysts from about a dozen different divisions. These consultations covered topics on constructing confidence intervals, testing hypotheses about differences between subpopulations, analysis using combined survey cycles, survival analysis, etc. There were several requests concerning comparisons of quantiles or their functions – a topic very rarely discussed in the literature. For a comparison of growth rates of median wealth, we proposed a few possible approaches, one of which has been used. We also helped Statistics Canada's analysts with implementation of methods in R, SUDAAN, SAS SURVEY and STATA software.

The group provided services to other methodologists as well. These consultations included questions on machine learning methods, auto-coding, text classification, degrees of freedom for variance estimation, building regression models, etc. We also created STATA examples for a workshop on analysis using GSS data.

External consultations were delivered to a variety of clients from other federal and provincial governments and agencies. The requests included analysis with non-probability samples, evaluating stability of principal component analysis based on survey data, methodological review of a study comparing promotions between groups, etc.

Finally, expert advice was given to the analysts and researchers from the Research Data Centres (RDCs). The topics included combining cycles of surveys, estimating standard errors for 1991-2006 Census estimates, using bootstrap weights for analysis of survey data, etc.

Provision of Training

The team presented seminars on data analysis from a complex sample design survey and offered various courses related to their areas of expertise, including a course on survival data analysis. Special presentations for training purposes have been developed and delivered including presentations on heat maps, machine learning and the proper use of p-value.

Collaboration with analysts

The article entitled “Reallocating time between sleep, sedentary and active behaviours: Associations with obesity and health in Canadian adults”, coauthored by Rachel Colley, Isabelle Michaud, and Didier Garriguet, was published in the April issue of *Health Reports.* (Health Reports, Vol. 29, no. 4, April 2018, Statistics Canada, Catalogue no. 82-003-X).

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

2.6 Questionnaire Design Resource Centre (QDRC)

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:

As part of Statistics Canada modernisation initiatives, the group explored various way to use technology and modern methods to optimise some of its operational work. The results were summerized in a talk titled *Modernizing questionnaire testing - Using technology to find efficiency (while maintaining quality)* and presented at the QUEST Workshop, an international conference of experts (Solomon, 2018). While the study highlighted mainly the continued advantages on real in-person testing, it helped identify targeted cases or side tasks that could be done remotely or with local support.

The group also contributed to various corporate consultation initiatives.

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

2.7 Knowledge Transfer – Statistical Training

New strategies to achieve our goal of capacity building and talent development have gradually been implemented. In general, the curriculum is now divided into thematic blocks under the responsibility of the appropriate resource centres (and the corresponding activities are reported in their respective sections). For basic survey methods, new versions of several of the courses have been developed. A new version of the sampling course including a presentation to be given by the participants was offered. A new course on robust estimation was offered. An introductory pilot course in data science has been added to the curriculum and taught by a recognized university professor. A course on small area estimation is in developpement. The new training strategies were summarized and discussed during a panel at the Joint Statistical Meetings (Fortier, 2018).

For further information, please contact:
Susie Fortier (613-863-9135, susie.fortier@canada.ca).

2.8 Knowledge Transfer – *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 2018 issues (44-1 and 44-2) were released in PDF and HTML versions. The June 2018 issue includes 7 regular papers. The December 2018 issue was led by two guest editors Jean-François Beaumont and David Haziza. It includes 10 invited papers selected from all the communications presented at the 9th *Colloque francophone sur les sondages*, which took place in Gatineau from October 11 to 14, 2016.

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 (613-863-9135, susie.fortier@canada.ca).

3 Research papers sponsored by the Methodology Research and Development Program

Beaumont, J.-F. (2019). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology* (submitted).

Bocci, C., and Beaumont, J.-F. (2018). Report on Small Area Estimation for the Global Affairs Canada Contract. Internal document, Methodology Branch, Statistics Canada.

Bocci, C., and Beaumont, J.-F. (2019). Small area estimation methodology applied to the Monthly Survey of Manufacturing-UPDATE. Internal document, Methodology Branch, Statistics Canada.

Chatrchi, G., Beaumont, J.-F., Gambino, J. and Haziza, D. (2018). An investigation into the use of sample matching for combining data from probability and non-probability samples. Proceedings of the Survey Methods Section, Statistical Society of Canada.

Chu, K., and Beaumont, J.-F. (2019). Formation of Homogeneous Self-Selection Propensity Classes for Non-Probability Samples via Probability Samples, SSC.

Chu, K., Yeung, A. and Dasylva, A. (2018). *Census Secondary Mother Tongue Write-in Autocoding: Prototype 2*. PowerPoint presentation.

Dasylva, A. (2018). Design-based estimation with record-linked administrative files and a clerical review sample. *Journal of Official Statistics*, 34, 41-54.

Dasylva, A. (2019). Autocoding as Record Linkage. Working paper.

Dasylva, A., Goussanou, A., Ajavon, A. and Abousaleh, H. (2019). Revisiting the Probabilistic Method of Record Linkage. Working paper.

Devin, N. (2018). Validation de l’extension au Bootstrap de Rao-Wu proposée par Pérez-Duarte applicable aux échantillons tirés selon des plans proportionnels à la taille à un ou deux degrés à l’aide d’une simulation Monte Carlo. Internal document.

Dochitoiu, C. (2018). Modelling Sample Rotation Effects through unobserved components. Internal document.

Duddek, C. (2018). Combining Census and Manitoba Hydro data to understand residential electricity usage. Proceedings: Symposium 2018, Combine to Conquer: Innovations in the use of Multiple Sources of Data, Statistics Canada.

Duddek, C. (2019). Using Manitoba Hydro data to evaluate imputation of electricity consumption in the Households and the Environment Survey, Presentation at the Social Statistics Methods Division (SSMD) Technical Committee.

Fortier, S. (2018). Modern Statistical Training Program for a National Statistical Office. Panel presentation at the 2018 Joint Statistical Meetings, Vancouver.

Fortier, S., and Thomas, S. (2018). (R)evolution of generalized systems and statistical tools at Statistics Canada, presented at the *6th Conference on the Use of R in Official Statistics,* The Hague.

Girard, C. (2019). Making Head or Tail of Differential Privacy. Internal seminar.

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.

Lee, B. (2018a). Receipt Logo Recognition Using Neural Network Algorithm, Simon Fraser University Co-op Term Work Report.

Lee, B. (2018b). Receipt Logo Recognition Using Neural Network Algorithm, Social Statistics Methods Division (SSMD) internal presentation.

Lesage, É., Beaumont, J.-F. and Bocci, C. (2019). Deux diagnostiques locaux pour évaluer l’efficacité de l’estimateur composite issu du modèle de Fay-Herriot. Draft, Statistics Canada.

Leung, J., and Dochitoiu, C. (2018). Use of time series forecasts as a contingency in the absence of expected data – Application to international trade aggregates. Internal document.

Matthews, S. (2018). Current Challenges with Quality Assurance of Seasonal Adjustment. Presented at the Second Seasonal Adjustment Practitioners Workshop, Washington, D.C.

Matthews, S., and Dochitoiu, C. (2018). Comparison of Seasonal Adjustment Approaches through State Space Representation. Internal Document.

Matthews, S., and Dochitoiu, C. (2018). State Space Modelling for Small Area Estimation with Time Series Components. Internal document.

Mayer, É. (2019a). Using Convolutional Neural Networks to Automatically Classify Logos on Shopping Receipts (MLCoP), Presentation at the Statistics Canada Machine Learning Community of Practice.

Mayer, É. (2019b). Using Convolutional Neural Networks to Automatically Classify Logos on Shopping Receipts, Presentation at the Symposium for Data Science and Statistics.

Mischler, L., Lapointe, M.A. and Patak, Z. (2019). A non-parametric test for detecting seasonality. Co-op term report.

Neusy, E. (2019a). Reporting Quality Using Confidence Intervals. Internal presentation at the Special Surveys Division Managers Meeting.

Neusy, E. (2019b). Wilson Confidence Intervals for Weighted Counts. Internal document.

Neusy, E., Boulet, C., Duggan, J., Mach, L., Mantel, H. and Reedman, L. (2018). Quality-based Release Criteria for Social Statistics. Internal presentation at the Social Statistics Methodology Division (SSMD) Technical Committee.

Patak, Z., and Armenski, T. (2018). Should I stay or should I go: A cross-border traveller's tale, Statistics Canada Symposium on Statistical Methodology, to appear in conference proceedings.

Reedman, L. (2018). A modest attempt at communicating about Quality. Presented at 2018 European Conference on Quality in Official Statistics, Krakow, Poland.

Reedman, L., and Windross, M. (2018). *The NSO, the NSS and Beyond!* Presented at 2018 European Conference on Quality in Official Statistics, Krakow, Poland.

Saïdi, A., Chu, K., Dasylva, A. and Labrecque-Synnott, F. (2018). Analysis of Binary Data Obtained by a Probabilistic Record Linkage, CANSSI Conference, Montréal.

Saïdi, A., Chu, K., Dasylva, A. and Labrecque-Synnott, F. (2018). Logistic regression with Linked Data, FCSM Conference, Washington.

Sallier, K., and Girard, C. (2018). Towards a successful implementation of Synthesis in a National Statistical Agency: A model for cooperation. Presented at the 2018 Privacy in Statistical Databases conference, organised by the Unesco Chair in Data Privacy, Valencia, Spain.

Savard, S.-A. (2019). Une approche de codage automatique avec données d'apprentissage basée sur la méthodologie du couplage d'enregistrements. Working paper.

Solomon, J. (2018). Modernizing questionnaire testing - Using technology to find efficiency (while maintaining quality). Presented at the QUEST Workshop, Germany.

Statistics Canada (2019a). Directive on Maintaining Time Series Continuity in Economic, Social and Environmental Statistics Programs. Internal document.

Statistics Canada (2019b). Guidelines on Maintaining Time Series Continuity in Economic, Social and Environmental Statistics. Internal document.

Verret, F. (2018). Some discussions on calendar effects in X12-ARIMA. Presented at the Second Seasonal Adjustment Practitioners Workshop, Washington, D.C.

Verret, F. (2019). Variance Estimation for Seasonally Adjusted Estimates. Presented to Economic Statistical Methods Division technical committee.

Yeung, A., Chu, K., Laroche, R. et Fortier, S. (2018). Exploring modern coding methods. Internal document. Presented to Statistics Canada's Advisory Committee on Statistical Methods.

You, Y. (2018). Area level modeling approaches to small area estimation using R and S-Plus with applications. International Cooperation and Methodology Innovation Centre (ICMIC divisional presentation.

You, Y. (2019). Hierarchical Bayes small area estimation of LFS status using linear and non-linear area level models. International Cooperation and Methodology Innovation Centre (ICMIC) research report.