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

Statistical Methodology Research and Development Program

Annual Report 2017/2018

Release date: October 22, 2018

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.

This report summarizes the 2017/2018 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 2017/2018

Contents

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 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: Research on the design-based mean square error of small area estimates

In small-area estimation theory, the variability of the estimates is measured using the model-based mean square error (MSE). Statistical agencies are interested in estimating the design MSE of these estimates in line with the traditional design MSE estimates of direct estimators for areas with adequate sample size. These design MSE estimates tend to be unstable when the area sample size is small. In this research project, we propose composite MSE estimators by taking a weighted sum of the design MSE and model MSE estimators. We study the properties of these composite estimators by looking at the design bias, the relative root mean square error and the coverage rate of the confidence intervals.

Progress:

We developed a research paper (Rao, Rubin-Bleuer and Estevao, 2018) on this topic and created programs to carry out simulation studies for both area-level and unit-level models. The results of these simulation studies show that the proposed composite MSE estimators provide a good compromise in estimating the design MSE. This paper was submitted and accepted for publication in the journal Survey Methodology.

Sub-project: Small area estimation for Global Affairs Canada (GAC)

The task agreed upon with Global Affairs Canada is 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 are 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.

The goal of small area estimation is to produce a domain estimate of better quality than the direct estimate by combining available, 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 sources. The next step involves modelling the direct estimates of the statistics of interest. The predictions of both these models are incorporated to finally produce a small area estimate and its corresponding measure of quality.

Progress:

We secured all the data sets necessary for this project. Three auxiliary sources are used: PD7 data, Employment Insurance data and data from the Job Vacancy and Wage Survey. The majority of our efforts was concentrated on the construction of a model and the evaluation of its performance. Underlying model assumptions must be verified. This involved analyzing graphs and other model diagnostics to determine the appropriateness of each model. With regards to the variance smoothing, we constructed and compared six variance smoothing models. The behaviour of the data in these experiments is such that much of this exploratory work is customized. That is, we had to construct more refined smoothing models than we used in our previous experiments. Once we are satisfied with our models, we will further investigate the quality of the resulting small area estimates before releasing them.

Sub-project: Small area estimation using non-probability panel data

Statistics Canada has been investigating alternate sources of data to not only reduce response burden but also to improve standard (direct) estimates coming from its surveys. In particular, data have been acquired from a non-probabilistic survey conducted on cell phones that offers rewards to respondents. This cell phone application asks a handful of questions on health, income and employment status. This project investigated the feasibility of using this auxiliary source in the context of small area estimation.

Progress:

Three small area estimation experiments were conducted. The first experiment was to evaluate if these panel data could be used as auxiliary information for small area estimation of certain health statistics from the Canadian Community Health Survey (CCHS) at the health region level. Similarly, a second experiment, with the same intention as the first, consisted of using the panel data in small area estimation methods for certain employment estimates from the Labour Force Survey (LFS) at the census metropolitan area/census agglomeration level. A third experiment was aimed at reducing response burden and investigated whether small area estimation methods using panel data could be applied on a smaller CCHS sample to achieve estimates of similar quality to the standard direct estimates derived from the full sample.

In the first two experiments, it appeared that there was little gain, if any in certain cases, in using panel data as auxiliary information in the small area estimation process. Small area estimates improved the direct estimates, but often, the improvement did not seem to come from the use of the panel data specifically. In the third experiment, there was some potential for applying small area estimation techniques on a reduced CCHS sample. It was found that small area estimates (SAE) estimates based on a reduced sample size (33% of the full sample)

would achieve similar precision as direct estimates based on the full sample. However, the precision gains of using panel data over a simple common mean SAE model was quite small. Similar precision gains or better gains could be obtained using existing administrative data. Major deficiencies with the panel data, such as measurement errors, coverage and representativity, might explain the lack of correlation between panel data and CCHS data and their ineffectiveness in improving SAE estimates beyond a common mean model. The report did not comment on the general question of whether the gains in using small area estimation methods outweigh the risks of model misspecifications nor did it assess whether the costs of acquiring these panel data in particular are worth the potential (small) gains in using them for small area estimation purposes.

This project is completed and a document was written which summarized the study and reported results. The project was also presented to the Advisory Committee on Statistical Methods in September 2017 (Beaumont and Bocci, 2017).

Sub-project: Empirical best linear unbiased predictor (EBLUP) and Hierarchical Bayes (HB) small area estimation with sampling variance modelling for totals and/or counts using matched and unmatched area level models with Monthly Survey of Manufactures (MSM) application

Compare EBLUP and HB estimates under different priors on variance components, particularly for the sampling variances when the area-specific sample sizes are small. Investigate the impact of sample size and the sampling variance models on the small area estimates. Propose proper methods and models to handle the sampling variance when the sample size is small with application using MSM data.

Progress:

For the EBLUP approach we considered the Fay-Herriot model when the direct sampling variance estimates are used. For the HB approach, we considered both the Fay-Herriot model and unmatched log-linear model with different priors on sampling and model variances. We compared the EBLUP estimator with HB estimators through an application using the MSM survey data. Both the EBLUP and unmatched HB estimators based on the unmatched log-linear model lead to positive estimates for all areas. The Fay-Herriot Hierarchical Bayes (FH HB) estimator may lead to negative estimates for some areas when the sample sizes are small and the direct coefficients of variation (CVs) or the input sampling variances are large. For the Fay-Herriot model, we suggest to use EBLUP approach instead of the HB approach when the sample sizes are small. For proportions or counts data, unmatched log-linear models may be used. A research report has been finished (You, 2018).

Sub-project: Small area estimation with benchmarking under model misspecification

Study EBLUP and/or HB estimators with the benchmarked EBLUP and/or benchmarked HB estimators, under model misspecification, bias of fixed effects and small area means, by simulation study for the Fay-Herriot model.

Progress:

We studied the self-benchmarked EBLUP estimator of You, Rao and Hidiroglou (2013) under model misspecification through simulation study. In particular we considered skewed normal model misspecification and unequal mean model misspecification. The result shows that the benchmarking procedure does not improve the EBLUP under non-normal distribution of the

random effects and unequal mean model misspecification. A research paper is completed (You, 2017).

Sub-project: SAE using R sae package

Investigate the SAE computing package using R and S-Plus. In addition, the time series options of the R package developed by Isabel Molina will be investigated to see if it can satisfy StatCan needs. The package will be tested using LFS data for the estimation of unemployment rates.

Progress:

The R package of Isabel Molina was tested. Some issues were found and feedback was provided to the authors. The gains by using time series methods were not found to be huge. Further explorations are needed before implementing a method in the SAE system. A report has been finished (You, 2018) and a presentation using R and S-Plus for SAE will be given.

Sub-project: SAE in the Labour Force Survey

In the previous year, we developed small area estimation models for the estimation of employment counts and unemployment rates by census metropolitan area/census agglomeration (CMA/CA) using LFS data. We also compared our SAE estimates for May 2011 with estimates from the National Household Survey. The goal of this project was to re-evaluate our models in light of new collected data and to compare our SAE estimates for May 2016 with those coming from the Census.

Progress:

Analyses of model residuals and graphs indicated that our models are still reasonable. We also confirmed that our SAE estimates were significantly closer to Census direct estimates than LFS direct estimates, especially for the smallest areas. These findings are reported in Hidiroglou, Beaumont and Yung (2018).

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

Reference

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/pub/12-001-x/2013001/article/11830-eng.pdf.

1.2 Developmental Research – Record Linkage

The record linkage research covers three areas including analysis with linked data, improvements to error estimation and new linkage techniques.

Progress:

Automated coding

A new methodology has been proposed to further automate the coding of the mother tongue in the census. The coding of this variable has been automated to a large extent with G-CODE. However, a small portion of the related write-ins are rejected by G-CODE are coded manually through a clerical review process. These manual operations generate large operational costs, which must be reduced drastically. To this end, a supervised method was proposed and implemented in a SAS prototype. It uses nearest-neighbour imputation and training data from the 2016 census. The prototype also incorporates and adapts techniques that have been previously used in the context of record linkage. In detail, the prototype attempts to code ("impute" language code) each G-Code-rejected mother tongue write-in by finding its "most similar" G-Code coded mother tongue write-in. Here, the similarity score of two given write-ins is a composite score based on the n-gram similarities of the respective mother tongue write-ins, the place-of-birth write-ins as well as the ethnicity write-ins. The prototype uses a mother-tongue Soundex as an artificially generated blocking variable to reduce the number of pairs of write-ins that need to be examined. At the composite score threshold of 60% or better, the prototype currently yields a "codability" of 70% and among the codable write-ins, it yields a 90% accuracy rate.

In addition to this work, a joint methodological framework was proposed for automated coding and record linkage, with an emphasis on linkage/coding errors.

Machine learning solutions for G-Link

When implementing a probabilistic linkage with G-LINK, the thresholds are often set in a manual and thus labour intensive manner. K-means clustering can be used to determine such thresholds in an unsupervised manner, i.e., without any clerical review. This project is a continuation of previous work by Quadir and Bao (2016). It has used synthetic data to evaluate the performance of different k-means strategies depending on selected features, including the pair linkage weight, a categorical vector of dummy variables and a categorical vector of ordinal variables. In the latter two options, the categorical vector is based on the rules outcomes for the different linkage variables. Two types of synthetic data were considered including binary linkage variables and more realistic synthetic data that included names, birthdates and addresses. These latter data were produced with a modified version of the Febrl data generator, which was prepared by the Systems Engineering Division and refined in the course of this work. The different k-means algorithms were compared to another solution for setting the thresholds without any clerical review. In this solution the thresholds are based on a mixture model under the conditional independence assumption. All the solutions have been compared in terms of the difference between the nominal (i.e., target) error rates and the actual rates. The results have shown that the best k-means performance is achieved when the features are based on the linkage weight, or on a categorical vector comprising of dummy variables. When the linkage weight is used, the performance of k-means is sensitive to the accuracy of the linkage weights. The performance of k-means is also sensitive to the initial cluster centroids. The model-based solution is competitive with these k-means options and provides a slightly better performance, i.e., a smaller difference between the nominal and actual error rates.

Impact of preprocessing on linkage errors

This project has looked at the impact of preprocessing on linkage errors using data from the Social Data Linkage Environment (SDLE), where some preprocessing takes place for names

(including titles) and postal codes. Three levels of preprocessing were applied to the same data set, including no preprocessing, preprocessing limited to titles and full preprocessing (for names and postal codes). For each preprocessing option, the linkage errors were measured using the Linkage Error Estimation Prototype (LEEP) and clerical reviews. The results have shown little benefit from the preprocessing of titles, see Pascal (2017).

Prototype for estimating linkage errors

A SAS prototype has been developed for the estimation of linkage errors with or without clerical reviews. The model-based estimates have been evaluated by linking consecutive personal tax files with the probabilistic method. The accuracy of the error estimates was accurately measured by using the actual match status of the record pairs. The results have shown that the model-based estimates are accurate when the pair linkage weights are accurate.

Estimation of linkage errors without any clerical review

The accurate estimation of errors is an important problem in record linkage. A methodology has been proposed for a deterministic linkage of social data, where the key, which comprises of all the linkage variables, is almost unique. It is based on a model and various assumptions but uses no clerical reviews. The methodology has been applied to the linkage between tax and hospitalization data using the variables sex, birthdate and postal code. For the False Positive Rate (FPR), the resulting estimate is within the 95% confidence interval of the clerical-review estimate.

Regression with linked data

Chipperfield, Bishop and Campbell (2011) have described a methodology for logistic regression with linked data using the linked pairs and clerical reviews. This methodology has been extended in two directions. The first extension is the use of a logistic model to model the probability hat a link is a false positive. The second extension modifies the original estimating equation by Chipperfield et al. (2011) into an optimal estimating equation according to the quasi-likelihood framework, see Tsui, Dasylva and Chu (2017). These two extensions have been evaluated using simulations.

Capture-recapture with linkage errors

A new methodology for capture-recapture with linkage errors has been described, see Dasylva (2018). It addresses the limitations (see Dasylva, 2017) of previous solutions by Ding and Fienberg (1994) and by Di Consiglio and Tuoto (2015).

For further information, please contact:
Abel Dasylva (613 951-7618; abel.dasylva@canada.ca).

References

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/pub/12-001-x/2011001/article/11444-eng.pdf.

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

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

Pascal, L. (2017). Optimisation de la préparation des données pour le couplage d'enregistrements. International Cooperation and Corporate Statistical Methods Division (ICCSMD), internship report.

Quadir, T., and Bao, A. (2016). Application of Machine Learning Algorithms in G-Link. Household Survey Methods Division (HSMD).

1.3 Developmental Research - Generalized Systems

The Generalized Systems Section (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;
- G-Confid: The generalized disclosure control system;
- Economic Disclosure Control and Dissemination System (EDCDS).

Aside from providing support and training related to generalized systems, the team also take on development research related to disclosure control, data visualization, variance estimation and other survey methods on which the systems are built upon.

Sub-project: Mixed variable error localization

In BANFF, a valid record is defined as a record whose numerical entries satisfy a set of linear inequalities, labelled "edits" within the system. Linear inequalities represent a very narrow subset of possible edits, and BANFF users (both internal and external) have expressed desire to see a broader set of edits addressed in BANFF, including "or" statements, and the integration of categorical variables. Many edits of this type can be expressed mathematically as "mixed variable" problems – mathematical functions acting on a combination of discrete and continuous variables. The objective of this project is to explore the feasibility of treating mixed variable problems in one of BANFF's most complex procedures: error localization.

Progress:

A literature review produced some interesting leads to follow. This included alternative definitions of the error localization problem, which were ultimately deemed too ambitious to research in the short term but could be considered in the future. A paper by DeWalt from Statistics Netherlands outlines a framework for mixed variable error localization that aligns well with the existing error localization methodology in BANFF. Original research focused on the integration of an "or" statement into BANFF's set of edits, and how it could be incorporated into BANFF. Future work includes developing a prototype within SAS, exploring the more general mixed variable problem (instead of just the "or" problem), and investigating the feasibility of incorporating these edits into other BANFF procedures (such as donor imputation).

Sub-project: Disclosure control - Random Tabular Adjustment

The Random Tabular Adjustment (RTA) framework is an approach of perturbative tabular adjustment that employs a Bayesian analysis to reduce disclosure risk. This approach provides key benefits beyond the proposed deterministic methods in the literature, all the while meeting the ultimate goal of publishing complete tables of results with no suppression.

Progress:

A working RTA prototype has been developed and tested on real survey data. The methodology has been applied to the Monthly Retail Trade Survey Retail trade - Sales by the North American Industry Classification System (CANSIM 080-0020) and the Monthly Coal Supply and Disposition Survey - monthly production of Exports of Coal (CANSIM 135-0002). Testing using real data has illustrated some of the challenges for client divisions because of the alternative impacts on quality. Discussions with other survey areas along with considerations for time series have taken place.

An RTA steering committee is now in place and giving guidance on the direction of the development of the method. Overview presentations explaining the method and its potential to analysts and users were prepared and given at the Federal Committee on Statistical Methods, the Statistical Society of Canada the Economic Statistics Federal-Provincial-Territorial meeting, Methods and Standards Committee and the Economic Statistics Forum. In general, the methodology and approach are supported and encouraged. The challenge is to be able to find an appropriate application for the method.

Work will continue on this project in the next year. The method will continue to be applied in different situations. Some generalization of the approach will be developed in order to allow it to be easily applied in different situations. There will also be research into introducing correlated noise in order to solve some of the application issues for some surveys.

Sub-project: Generalized system continuous development and support

The Generalized Systems Section 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 continuous support to users, updated and delivered training presentation in various forums and met with international delegates to discuss ongoing and future development of the generalized systems. Using alternative training media including WebEx was investigated and successfully applied to G-Confid training for a group from the Canadian Revenue Agency. The group met with delegations from Italy, Japan, Singapore and Ireland.

The investment project for G-Confid improvements neared completion with minor bug fixes remaining. New functions include the use of survey weights, the ability to analyze negative values and the support of waivers in the risk assessment. The new functionality is being integrated into the Economic Disclosure Control and Dissemination System (EDCDS), a metadata-driven processing system to implement estimation and confidentiality for use by the Centre for Special Business Projects (CSBP). The new ideas have been presented to several internal groups including CSBP and Field 5 management with special consideration for

agriculture surveys. The new functionality was also presented at the United Nations Economic Commission for Europe (UNECE) work session on Confidentiality in Skopje, Macedonia.

Sub-project: Generalized System continuous development and support

A research project was undertaken to study new functionalities as potential candidates to add to the stratification and allocation methodologies available in G-Sam. The potential candidates were the Sethi Algorithm (a generalization of the Lavallée-Hidiroglou algorithm), the Kozak algorithm and the genetic algorithm.

Progress:

Simulation studies demonstrated that, under certain conditions, both the Kozak algorithm and the genetic algorithm provided lower sample sizes for a fixed coefficient of variation, compared to three classical approaches: the cumulative square root f method, the geometric method and a generalized version of the Lavallée-Hidiroglou method specified by Baillargeon and Rivest (and called the Sethi algorithm in the study). For univariate stratification the Kozak algorithm always outperformed the classical approaches and took the least runtime. The genetic algorithm sometimes found as good a solution as the Kozak algorithm although taking significantly longer runtime. The main advantage of the genetic algorithm is that it alone offers multivariate stratification and a multi-domain feature.

An additional aspect of this research project was to consider using developed R packages rather than developing in-house SAS-based solutions. The R-package *SamplingStrata* was used for the Genetic Algorithm while the package *stratification* was used to implement the others. In order to adapt the functionality for use at Statistics Canada, a SAS program was developed that permits SAS users to make use of SAS data sets as inputs. The call to R within SAS is done transparently using PROC IML, and the outputs translated into SAS, so that users only need to know SAS in order to run the R program.

An internal report and a slide deck document the simulation studies, while another slide deck documents the call to R within SAS.

Sub-project: Confidentiality of regression-based output

This project serves the need of the Canadian Centre for Data Development and Economic Research (CDER) to improve upon the assessment of the output of regression modelling for confidentiality.

Progress:

The investigation included a review of the relevant literature and discussions with CDER representatives. Special considerations were needed in the context of CDER where researchers often re-run their analysis using nearly identical subsets of microdata, which only serve to create slivers of a few business enterprises, the values of which may be prone to disclosure when two sets of regression outputs are compared. A working draft of the set of guidelines was written. Some of the guidelines require further elaboration in the next fiscal year.

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

1.4 Developmental Research - Collection

The data collection research portfolio has for objective to support collection and operations research activities related to the corporate priorities. The 2017/2018 collection research activities cover projects that are related to the declining response rates for household surveys, to active collection management initiatives and to help prepare the future multimode environment. One project is a new initiative while the other three are continuing projects that have been started in the previous years.

Sub-project: Development of a case prioritization system to optimize collection for CATI surveys (Gildas Kleim, Kenza Sallier)

The objective of this project is to develop a procedure for calculating a prioritization score to identify which case will be interviewed when a CATI interviewer becomes available, all in the Integrated Collection and Operations System (ICOS). The algorithm has three components. The first takes into account previously entered information about arrangements made with the respondent for appointments, the second uses auxiliary information to predict the best time to make the initial call, and the third takes into account the progress of collection currently underway, in particular the results of previous attempts made in each case, to predict the best time to call the respondent. A more specific aspect of this project is developing a method that gives a good estimate of the probability that a household can be contacted within a given time slice.

Progress:

We have continued to assess the machine learning methods presented in the article “Bayesian Hierarchical Rule Modelling for Predicting Medical Conditions” (The Annals of Applied Statistics, 2012) by applying them to real survey data, in particular from the 2015 National Apprenticeship Survey. We contacted one of the researchers who wrote the article, and he sent us his programs (written in R) which are now used as a basis for work.

For further information, please contact:
Gildas Kleim (613-853-9553; gildas.keim@canada.ca).

Sub-project: Segmentation of the Canadian population to derive efficient communication and operation strategy to increase response rate for social surveys

Response rates to surveys are decreasing and statistical agencies must find innovative ways to improve collection strategies while maintaining high-quality data. This project focuses on household surveys at Statistics Canada and applies a segmentation approach to create clusters and more specifically, to identify geographic areas and their associated characteristics where household non-response tends to be higher. A prior study was conducted at Statistics Canada using data from the 2011 Canadian Census of Population, whose goal was to understand the statistical profile of people who responded to the census right away versus those who required follow-up. This previous study provided motivation to apply the same framework in a household survey context with a goal of better understanding the statistical profile of households who do not respond to surveys. This information can then be used to tailor advertising campaigns, communication tools and collection strategies for specific clusters that are known to have higher non-response. There are three main steps to this segmentation

project: linking households that were selected for surveys to the census data, modelling the non-response to identify significant predictor variables and performing the cluster analysis.

Progress:

A feasibility study to test the proof of concept of a segmentation analysis for household surveys using the 2011 National Household Survey (NHS) data and survey data from the Canadian Community Health Survey (CCHS), the Survey of Household Spending (SHS) and the Survey of Financial Security (SFS) was completed (Halladay, 2018). The conclusion reached was that the cluster analysis proved to be useful in identifying statistical profiles of survey respondents and non-respondents. A number of recommendations and improvements were recommended for further research.

Once the 2016 Census long-form data became available, the segmentation project was updated to take this into account. Instead of only using CAPI surveys, all voluntary household surveys that were in the field around the same time as the 2016 census were considered. The non-response was modelled using stepwise logistic regression and a number of census variables were deemed significant. These variables (i.e., household characteristics associated with non-response) were then standardized and used in the cluster analysis.

Some improvements from the previous study include

- increased linkage rates from household surveys to census data;
- investigated various geographical levels. Aggregate dissemination areas (ADAs) proved to be the most useful for our purposes;
- importance factors were applied to the variables in the clustering to give a higher weight to variables that are thought to have a larger impact on non-response;
- some studies into selecting the best initial seeds for the k-means clustering algorithm were conducted (they were not randomly selected like previously).

The clusters were finalized and presented to the communications team. Documentation was completed.

For further information, please contact:
Amanda Halladay (613-854-1937; amanda.halladay@canada.ca).

Sub-project: CAPI collection using a non-clustered sample

The Survey of Financial Security (SFS) sample design was changed between 2012 and 2016. Despite the fact that collection for SFS is done in person, in 2016 a non-clustered design was used in most areas. Only in areas that were deemed too rural or for which the frame did not have high enough quality address information was a clustered design used.

Progress:

Since clustered designs are typically used to simplify collection for in person surveys, after the 2016 SFS, collection paradata was examined to determine the effect of increasing the use of a non-clustered sample on collection operations. The main observations of this project suggest that increasing the use of a non-clustered design did not negatively impact collection. Overall the response rate was higher in 2016 than in 2012. In 2016, the average number of attempts was higher than in 2012 for both respondents and non-respondents. For cases where there

were two or more attempts made on a sample unit, the average number of days between attempts and the average number of days from the first attempt to the last attempt were higher for 2016. This is consistent with practices of spreading out attempts over the entire collection period. Importantly, the non-clustered sample was not substantively different from the clustered sample in terms of collection paradata patterns.

SFS has a long collection period (three months) and a relatively long interview meaning that few interviews are done within one day. It may be that these characteristics of SFS make it a good candidate for a non-clustered design. In this context, the use of a non-clustered design, which can significantly improve the efficiency of the design as it did for SFS, can be endorsed for other surveys with similar characteristics as SFS.

For further information, please contact:
Cilanne Boulet (613-851-0031; cilanne.boulet@canada.ca).

1.5 Prospective Research - Non-probabilistic approaches

Sub-project: Non-traditional statistical research

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.

Rivers (2007) examined a technique called sample matching, which is related with statistical matching or data fusion. In this method, a probability sample is drawn from a sampling frame, and each sample unit is matched to a panel member based on given characteristics available on both data sources. Note that exact matches are not required. Values of the panel members are then given to the sample unit. This technique can be viewed as donor imputation with the recipients being all sample units and donors being panel members.

The purpose of this research is two-fold. The first goal is to assess the possibility of using sample matching or other data integration techniques for some of the programs of Statistics Canada. This approach might be useful, for example, for programs where response rates obtained through traditional collection methods are relatively low. It could also be used to reduce collection costs and response burden. The second goal is related to the exploration of alternative data sources.

Progress:

We have recently acquired data from a non-probability 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. To reduce the selection bias, we considered two techniques: statistical matching and calibration. The success of both methods depends on the strength of auxiliary variables available in both sources. In our experiment, we used demographic variables such as 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 (further described with a divisional research sub-project) seemed to be slightly more effective than calibration. This might be due to its nonparametric nature unlike calibration which relies on a linear model. However, a non-negligible 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 will be presented at the Statistical Society of Canada (SSC) and Canadian Statistical Sciences Institution (CANSSI) conferences.

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, 2007.

1.6 Prospective Research - Data Science

This section covers important research and exploration activities related to data science, machine learning and artificial intelligence that are not reported elsewhere. These activities were mainly carried out by two groups of methodologists involved in the active learning pilot project. As part of these activities, they explored topics in depth to increase their own capacity and organized several outreach events to share knowledge and facilitate the integration of new techniques into regular tasks.

In particular, the groups participated in the Innovation Fair and organized three learning events that included 12 seminars. These events attracted an average of 150 people each time. Newsletters with useful information on machine learning (such as a glossary and a list of related courses available online) were prepared and shared on a monthly basis. The groups also dissected more advanced scientific articles and exchanged on these topics. These include exploration work on the effective use of computer tools - including but not limited to SAS - in the context of large data, on the application of survey methods such as sampling and outlier detection in the context of large data and on various specific machine learning algorithms.

In order to facilitate the transfer of knowledge, the groups also worked on creating prototypes in R and/or python to illustrate the use of these methods. In doing so, they also greatly contributed to the initial development of a strategy to bring these tools into our IT environment. The tools and methods have also been used in more concrete projects, for example as an initial exploration for the automatic classification of comments received on the census and an automatic identification of the cause of death in health-related projects.

Finally, the group also documented the current or potential applications of machine learning techniques in the various steps of the Generic Statistical Business Process Model.

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

1.7 Divisional Research - Business Survey Methods Division (BSMD)

This project is used to finance research and development work related to divisional priorities. In the Business Survey Method Division, the some of the budget was set aside to address upcoming issues, if any. The focus and priorities were given to projects which could lead to publications.

Sub-project: Study on calibration in the Integrated Business Statistics Program (IBSP)

Calibration is an estimation approach implemented into the IBSP estimation system. This method has proven quite effective for the first surveys integrated, which mainly measure revenue and expenditure variables and use revenue as a calibration variable. However, with the integration of more and more and different types of surveys (such as measurements related to rare characteristics and/or events), we question the effectiveness of calibration as regards the accuracy and level of estimates (large variations between two occurrences). To better understand the issues and provide possible solutions, a calibration working group was established. Following some meetings and discussions, it was suggested to use a data set from a survey already integrated into the IBSP in order to better understand the IBSP methodology, to determine in which circumstances calibration is more or less effective, and whether methods and/or parameters should be suggested to improve the efficiency of calibration. The ultimate goal of the working group is to provide tools to help methodologists assess and make informed and effective decisions on whether or not to use calibration for the surveys for which they are responsible (whether or not they have been integrated into the IBSP).

Progress:

A study was initiated using data from the 2015 Industrial Consumption of Energy Survey, a survey that had already been integrated into the IBSP. Results on the progress of that study were presented to the working group on a regular basis and in response to discussions and suggestions, new studies were conducted. These include the comparison of estimates and standard errors before and after calibration using the current methodology, the comparison of estimates using various approaches and parameters to improve calibration (such as excluding units in calibration, modifying constraints, etc.). In the end, two calibration methods were selected to perform simulations in order to measure the extent of bias as well as the mean square error before and after calibration (using the two methods). The results were presented to the BSMD Technical Committee in the fall of 2017. The Committee supported the proposal to develop a model to study calibration using an existing survey as well as implementing diagnostics at the IBSP estimation stage to help analysts and Methodology identify potential issues and react quickly. Following this meeting, a calibration guide was drafted and a study model will be added to the guide using the Industrial Water Survey. The study will begin in June 2018. The work on diagnostics in the IBSP will be done at a later date.

For further information, please contact:
Marie-Claude Duval (613-854-1929; Marie-Claude.Duval@canada.ca).

Sub-project: Detection and treatment of outliers in business surveys

Outliers are currently declared as such in several business surveys using predetermined thresholds. The units that are declared as outliers have their sampling weights changed to

one, and the weights of the remaining units are adjusted so that the sum of all the weights are equal to the population size. The problem with this procedure is that it is ad hoc, and will result in biased estimates (on account of the weight changes). This is documented in Hidiroglou and Srinath (1981).

Progress:

It is known that methods based on robust procedures are better than the ones that are ad hoc, and that change the sampling weight arbitrarily. Recently, Beaumont, Haziza and Ruiz-Gazen (2013) used the conditional bias associated with a sample unit to define the measure of influence in finite population sampling. This procedure which is robust uses the conditional bias to obtain estimators by down weighting the most influential sample units. The method was adapted to SEPH, and specifications were written (Hidiroglou, 2017).

For further information, please contact:
Mike Hidiroglou (michel.hidiroglou@canada.ca).

Sub-project: Stratification of asymmetric population

Business survey population are known to be highly asymmetric. The goal of this project is to compare three procedures to determine the stratum boundary in the case of asymmetric population: cumrootf (Dalenius and Hodges, 1959), geometric (Gunning and Horgan, 2004) and Lavallée-Hidiroglou (1988), and to see the impact of extended algorithms to aim for optimal solutions.

Progress:

A number of data sets were used to reflect the superiority of optimization-based methods over the approximate methods. Furthermore the equivalence between sample size and coefficient driven stratification boundaries was demonstrated.

A paper entitled “Stratification of Skewed Populations: A Comparison of Optimization-Based Versus Approximate Methods” was submitted to the International Statistical Review journal. The paper was accepted and appeared in the journal in its March 2018 issue.

For further information, please contact:
Mike Hidiroglou (michel.hidiroglou@canada.ca).

Sub-project: Modifying the Hidiroglou-Berthelot (HB) method

The Hidiroglou-Berthelot (HB) ratio method is commonly used in surveys for the detection of outliers. The procedure puts more emphasis on small changes associated with large reported values than with large changes associated with the smaller units. The method is based on a two-stage transformation of the data. The transformed data declare units as outliers if they are outside an interval. This interval is based on the interquartile distance obtained from the resulting distribution.

A problem recently noticed with the use of the interquartile distance is that the resulting boundaries may not be reasonable. This results in the declaration of too many false outliers.

This problem takes place when a large proportion (> 25%) of the observations has the same ratio: the ratio being between the current and previous observation of a given unit, or the ratio between two variables within a given unit. Moreover, when the proportion of such cases is greater than 50%, all other observations whose ratios are different are declared as outliers.

Progress:

The HB method was adjusted to avoid this problem. The suggested adjustment involves using interpercentiles based on the 10^{th} and 90^{th} percentile instead of the first and third quartiles. This results in bounds that are reasonable and yields acceptable outliers. This work is documented in Hidiroglou and Emond (2018).

For further information, please contact:
Mike Hidiroglou (michel.hidiroglou@canada.ca)

Reference

Estevao, V., Hidiroglou, M.A. and You, Y. (2015). *Methodology Software Library Small-Area Estimation Unit Level Model with EBLUP and Pseudo EBLUP Estimation Methodology Specifications*.

Sub-project: Impact of the use auxiliary data on sample design

Calibration is the default estimation method in the Integrated Business Statistics Program's estimation system. One problem with calibration is that the resulting calibration weights that are the product of the original design weight times a *g*-factor (that reflects the auxiliary data) may be negative. Although the Generalized Estimation System can modify the calibration weights to be positive, there can be instances when it is not possible. An alternative solution is to minimize the occurrence of negative weights at the sample design stage. Another way to minimize (eliminate) the existence of negative weights is to use rejective sampling.

Progress:

A number of solutions were investigated to reduce (eliminate) the presence of negative weights. One solution involves increasing the number of minimum units within a stratum, and to declare units that are outliers in the auxiliary data as take-all (Hidiroglou, 2018). Another solution is to use rejective sampling. This method was originally pioneered by Hájek. It was later used by Fuller (2009) and Chauvet, Haziza and Lesage (2017) in the context of auxiliary data. Stephan and Hidiroglou (2018) provide a method for computing the variance of the resulting estimators given that rejective sampling is used.

For further information, please contact:
Mike Hidiroglou (michel.hidiroglou@canada.ca).

References

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

Chauvet, G., Haziza, D. and Lesage, E. (2017). Examining some aspects of balanced sampling in surveys. *Statistica Sinica*, 27, 313-334.

Dalenius, T., and Hodges, Jr., J.L. (1959). Minimum variance stratification. *Journal of the American Statistical Association*, 54, 285, 88-101.

Fuller, W.A. (2009). Some design properties of a rejective sampling procedure. *Biometrika*, 96, 933-944.

Gunning, P., and Horgan, J.M. (2004). A new algorithm for the construction of stratum boundaries in skewed populations. *Survey Methodology*, 30, 2, 159-166. Paper available at https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/2004002/article/7749-eng.pdf.

Hidiroglou, M.A., and Srinath, K.P. (1981). Some estimators of a population total from simple random samples containing large units. *Journal of the American Statistical Association*, 76, 375, 690-695.

Lavallée, P., and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 1, 33-43. Paper available at https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1988001/article/14602-eng.pdf.

Sub-project: Exploring clustering applications in outlier detection for administrative data sources

The outlier detection techniques currently available in Statistics Canada's edit and imputation generalized system (Banff) are highly effective in cases where the variable of interest follows a unimodal distribution, either on its own or within groups by class. Often with large administrative data sources such as international merchandise trade data, finding a set of class variables which can be used to satisfy this assumption is a challenge, and the effectiveness of the outlier detection is subsequently reduced. The goal of this project was to utilize unsupervised learning methods capable of handling a mixture of quantitative and qualitative variables, in order to exploit the high dimensionality of administrative data sets. More specifically, the goal was to examine the use of feature selection and hierarchical clustering to isolate modal distributions as a pre-treatment to the outlier detection in Banff, as well as examine the use of non-parametric density clustering methods for outlier detection alone in a comparison study against current methods.

Progress:

Studies were initiated to understand the various feature selection and clustering methods available, and to determine the best options for merchandise trade data. Based on these studies, three separate approaches were chosen to compare with the current method in Banff; these approaches use a combination of feature selection and clustering methods, either as a stand-alone outlier detection method, or as a pre-treatment to Banff. Programs were written to perform outlier detection using each approach, and tested on trade data. The results were compared directly to the current method. Presently, discrepancies between the results are in the process of being characterized. This work will be presented at the Joint Statistical Meetings 2018 and a paper will be written for the proceedings.

For further information, please contact:
Elizabeth Ayres (613-404-4102; elizabeth.ayres@canada.ca).

1.8 Divisional Research - Household Survey Methods Division (HSMD)

Any research sub-projects submitted for this project must be in line with the Methodology Branch's mandate, and specifically with the Household Survey Method Division's objectives. The sub-projects must aim to (1) develop innovative procedures to remain in step with changes in the context of increasing budgetary constraints and requests, (2) improve survey design efficiency, and thus reduce costs, and (3) ensure the best quality of the results produced by proposing innovative and effective statistical methods that are integrated into the agency's statistical programs.

Sub-project: Use of credit card data in household surveys

Credit or debit card data is one of the richest alternative data sources. Like many national statistical agencies, Statistics Canada has started exploring the use of alternate data sources and begun a process of acquiring credit and debit card data. The objective of this project is to identify potential uses of this type of data for household surveys.

Progress:

An agreement was signed with Destination Canada to receive raw payment processors data, which includes a portion of credit and debit card payments made by international visitors to Canada. The payment data, which is aggregated by Merchant Category Code, FSA (Forward Sorting Area) and Country of origin of the card, was received in November 2017. In collaboration with many divisions (including the Tourism and Centre for Education Statistics Division), the strengths and weaknesses of the data were identified. To assess the quality of the data, including coverage, various validation steps were completed. Moreover, a preliminary analysis was conducted to highlight the key features of the data. The payment data was then compared to data from International Travel Survey (ITS), and several modelling strategies, including small area estimation, were considered to produce tourism spending estimates. The preliminary results of the project as well as the challenges in using payment data to estimate tourism spending were presented in the HSMD Technical Committee (Gagnon and Gravel, 2018).

It should be noted that the scope of the analysis was limited to the International Travel Survey (ITS), as the current data set only includes electronic (credit and debit) card spending by international visitors to Canada. However, negotiations with three major credit card companies are underway to acquire domestic credit card data that can be used in other household surveys.

For further information, please contact:
François Gagnon (613-292-4645; francois.gagnon2@canada.ca).

Sub-project: An investigation into the use of sample matching for combining data from probability and non-probability samples

With increasing levels of nonresponse in household surveys, there is renewed interest in alternatives to the traditional way of conducting such surveys. One possible solution is to integrate data from web panels into probability samples, and benefit from the positive features of both types of survey. In order to explore that option, an experiment was conducted using data from an incentive-based digital platform along with survey data from the Canadian

Community Health Survey (CCHS). The goal of the project is to test sample matching method due to Rivers (2007) on this data.

Progress:

We calculated the sample matching (SM) estimates for different variables and compared SM with three estimates: CCHS estimate (Benchmark), unweighted panel estimate (naive estimate) and post-stratified panel estimate (improved naive estimate). The results suggested that performance of the SM estimator depends on variables of interest and varies by age groups. The difference between CCHS estimates and SM estimates could be due to many factors, such as population coverage, selection bias, non-sampling errors, mode effect or social desirability bias. The age distribution of the panel subscribers is skewed to the right and the compositional characteristics of the panel and CCHS respondents, such as marital status and education level, differ in some ways. The SM estimates are more biased in underrepresented age groups. The application used in this experiment is an incentive-based platform and respondents receive points for completing the surveys. This has an impact on the quality of data. Comparing estimates by age groups shows that estimates of the age group 18 to 34 are more biased even though this group is over-presented in the panel. One interpretation could be that the younger individuals participate to collect points (e.g., for a movie card), and do not answer to questions carefully. Moreover, social desirability bias is more likely to be found in the CCHS data than in the panel data.

The preliminary results of the project as well as the challenges in combining data from probability and non-probability samples were presented at the Advisory Committee on Statistical Methods (ACSM) in October 2017 (Chatrchi and Gambino, 2017). The final results will be presented in the SSC conference in June 2018.

For further information, please contact:
Golshid Chatrchi (613-854-1886; golshid.chatrchi@canada.ca).

Sub-project: Model the income distribution to better predict the low-income portion

During the project to produce a historical series on income using data from the Survey of Labour and Income Dynamics (SLID) and data from the T1 Family File (T1FF), we realized that the distribution of family income from the SLID was different from the T1FF distribution, particularly for the low-income portion. We would like to model family income using a parametric distribution as well as survey data, and to use the modelled data to better “predict” the distribution in the low income percentiles of the T1FF.

The results would help to improve the quality of the T1FF income variables, and thereby indirectly improve the quality of the income variable for the Household Survey Frame Service.

Progress:

The project is currently in the documentation stage. The report will include a discussion on the use of the GB2 distribution (second-type generalized beta) for modelling income data from the SLID or the T1FF. We will describe the steps required to use this distribution in the different fields and will present an analysis of the results (Jocelyn and Tam, 2018).

For further information, please contact:
Wisner Jocelyn (613-862-0341; wisner.jocelyn@canada.ca).

Sub-project: Extending the Rao-Wu rescaling bootstrap method to two-stage sample designs without replacement

Implementing the Rao-Wu rescaling bootstrap method is justified for a survey whose design includes more than one stage, as long as the first-degree sampling is *with replacement*. Using this method is justified when sampling is done *without replacement* as long as the sampling fraction used is not significant. However, this method is sometimes used even when the first-degree sampling fraction is not negligible in certain strata. Very recently, in 2012, an extension of the Rao-Wu rescaling bootstrap method was proposed to fully consider the variance of an estimator under a multi-stage simple random design without replacement (Osiewicz and Pérez-Duarte, 2012). The authors of this paper also proposed a similar extension for cases where the first-stage sample is drawn using the Rao-Hartley-Cochran procedure (i.e., proportional to the size) (Osiewicz and Pérez-Duarte, 2012).

Progress:

We demonstrated that the proposed extension for two-stage simple random cases works theoretically by developing proof that the generalized bootstrap conditions described by Beaumont and Patak (2012) are satisfied. These results have been documented (Charlebois and Laperrière, 2017). We familiarized ourselves with the Rao-Hartley-Cochran sampling procedure, the estimation and variance estimation in this case using the articles by Rao, Hartley and Cochran (1962), Ohlsson (1989) and Chaudhuri, Dihidar and Bose (2006). We then looked into the draft Osiewicz and Perez-Duarte article that proposes the extension to the Rao-Hartley-Cochran (RHC) case. We programmed simulations on a data set from Särndal, Swenson and Wretman (1992) to confirm that the proposed extension works for the RHC case. It would be interesting to pursue the simulations to include various scenarios of two-stage sampling fractions, and then to test this method with data from a household survey.

For further information, please contact:
Christiane Laperrière (613-854-0571; christiane.laperriere@canada.ca).

Sub-project: To better understand non-response by using the Socio-Economic File (SEF)

Results from a 2015 evaluation by Justin Francis, Yves Lafortune and Jean-François Simard showed that the available information on the SEF highly coincides with information collected from responding households to the Labour Force Survey (LFS). Results also showed that even when the household composition was different, it was often the case that at least half of the persons aged 15 years or older still lived at the same address. The goal of this research project is to extend this evaluation to focus more on the LFS *non-respondents*. For about half of them, we are able to obtain survey responses at later months (which could be seen as a late response for other surveys with a larger collection window). We want to see if in such cases, the persons in the household are the same as on the SEF. In particular, is the concordance rate different than the one obtained from respondents? The results would be useful for the weighting and imputation stages, which are normally used to account for non-response.

Progress:

Appropriate files were prepared by linking selected data from LFS to the SEF. The linkage results were then examined by reproducing part of the analysis done for respondents in the Francis, Lafortune, and Simard (2015) study. In addition to comparing the linkage results between non-respondents and respondents, the linkage results of different non-response

categories (no contact, out of scope, etc.) were also assessed in order to determine whether different types of non-response might be more reliably linked to the SEF and how this might affect the formation of non-response adjustment groups. A report documenting the steps and analysis of the linkage between the LFS initial non-respondents and the SEF is in progress (Lafortune and Tam, 2018).

For further information, please contact:
Yves Lafortune (613-951-4772; yves.lafortune@canada.ca).

Sub-project: Confidence intervals for complex designs

The project has two purposes. The first is to provide recommendations for reporting the quality of estimates, proportions in particular. Concerns with the current practice of using CVs to measure the quality of proportions has led to the recommendation of releasing confidence intervals rather than quality indicators based on CVs. Given the latter recommendation, the second purpose of the project is to study methods of constructing confidence intervals. In 2016, a study examining several methods of constructing intervals for proportions was completed (Neusy and Mantel, 2016). The plan is to extend the study to other types of estimates, such as differences of proportions and weighted counts.

Progress:

Preliminary simulations were run to evaluate the performance of confidence interval methods for differences of proportions and weighted counts. Preliminary results seem to indicate that confidence intervals for weighted counts have the same type of performance issues as confidence intervals for proportions.

A presentation was given to the Methods and Standards Committee in November (Neusy, 2017), which recommended the use of confidence intervals for measuring and reporting the quality of estimates in terms of their sampling error. The committee approved the recommendation to adopt as best practice the use of confidence intervals for reporting quality.

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

Sub-project: Review the overall approach of using the residential telephone file as a sampling frame for household surveys

Household surveys such as the General Social Survey (GSS) and the Canadian Tobacco, Alcohol and Drugs Survey (CTADS) that are not address-based have been using the Residential Telephone File (RTF) as a sampling frame of telephone numbers for a number of years now. With the increase use of electronic questionnaires (EQ), the use of the RTF presents new challenges to such surveys as the EQ invitation sent by regular mail can sometimes not be linked to the right address. Furthermore, the use of the residual portion of the RTF has shown to be problematic in collection, leading to lower response rates and higher out-of-scope rates. However, this residual portion cannot be excluded based on earlier studies that show it significantly increases the coverage of the target population.

This project aims to review the use of the RTF, and in particular its residual portion, as a sampling frame in household surveys. Everything from coverage and weighting issues will be

evaluated. The idea is to come up with a recommendation for all surveys that use the RTF as a sampling frame. Of particular interest is to see if an address-based approach using the Dwelling Universe File (DUF) could be used to replace the current RTF approach.

Progress:

An evaluation of the new DUF was completed in order to update some of the coverage numbers that were derived in 2012. Improvements to the Household Survey Frame Service (HSFS) products lead us to think that the number of non-mailable addresses would be reduced significantly since 2012, perhaps enough so that a survey with no CAPI could consider using the DUF as an address-based sampling frame for electronic questionnaires-Computer-assisted telephone interview (EQ-CATI) surveys. Although the non-mailable addresses are indeed less of an issue on the newer DUF, there are still some geographical areas where using the DUF would present significant under-coverage unless some CAPI could be considered. Some simulations were done to evaluate the potential bias in estimates from excluding these non-mailable dwellings from the target population.

Results of the study were presented at the HSMD Technical Committee (Caron and Chabot-Hallé, 2018). A number of approaches were presented and discussed to support the use of an address-based sampling. These included the use of targeted CAPI for problem areas and the use of neighbourhood mail to reach non-mailable dwellings. It has been recommended that some of these approaches be pilot-tested in the near future.

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

Sub-project: Smart meter electricity data for Manitoba

The ultimate objective is to produce estimates of residential electrical energy consumption at the Census dissemination-area level for Manitoba (a little over 2,000 census dissemination areas). We began receiving monthly electricity use data from Manitoba Hydro in 2015. We would like to understand the coverage of the data as it pertains to households. The first goal is to assign a census dissemination area to each meter and then compare the number of meters (hydro data) with the number of households (2016 Census data) within each dissemination area.

Progress:

After trying various options to obtain the dissemination areas (DAs), the best solution was found to be the use of the ARCGIS software to overlay the points from Manitoba Hydro and census boundary files. Other options investigated included a link with the address register and the use of the Postal Code Conversion File (PCCF); they would less successful due to a lack of civic address in rural area or impreciseness.

With the geography issue resolved, we compared the number of meters with the number of households for May 2016 (cross-sectional analysis). What we found was that if a DA was composed of mostly single detached dwellings, there is an approximate 1:1 relation between the number of electrical meters and the number of households. This relationship does not hold as well, unfortunately, when the DA includes apartment buildings. One of the challenges is to find the number of households that are associated with a particular meter.

One surprising finding was that meters associated with apartment buildings were often classified not as residential meters but as commercial meters. This is because the meter is connected to a business. This may suit the purposes of Manitoba Hydro, but it means that this electricity usage is not going to be included in residential use estimates.The overall results are documented in a presentation (Duddek, 2018).

For further information, please contact:
Christopher Duddek (613-862-9234; christopher.duddek@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, 127-148.

Chaudhuri, A., Dihidar, K. and Bose, M. (2006). On the feasibility of basing Horvitz and Thompson's estimator on a sample by Rao, Hartley, and Cochran's scheme. *Communications in Statistics - Theory and Methods*, 35, 12, 2239-2244.

Francis, J., Lafortune, Y. and Simard, J.-F. (2015). What Can Be Learned by Combining the Socio-Economic File (SEF) and the Labour Force Survey (LFS)? Internal document, Household Survey Methods Division (HSMD).

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

Ohlsson, E. (1989). Variance estimation in the Rao-Hartley-Cochran procedure. *Sankhya, Series B,* 51, 3, 348-361.

Osiewicz, M., and Pérez-Duarte, S. (2012). Flexible and homogeneous variance estimation in a cross-country survey under confidentiality constraints. Q2012 Conference, Athens, Greece.

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

Rao, J.N.K., Hartley, H.O. and Cochran, W.G. (1962). On a simple procedure of unequal probability sampling without replacement. *Journal of the Royal Statistical Society, Series B,* 24, 2, 482-491.

Rivers, D. (2007). Sampling for web surveys. *Proceedings of the Joint Statistical Meeting*, Salt Lake City, Utah, 2007.

Särndal, C.-E., Swenson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*. Springer Series in Statistics.

1.9 Divisional Research - Social Survey Methods Division (SSMD)

Sub-project: CANCEIS course for methodologists

Budget was allocated to the E&I unit to build a CANCEIS course that would be offered to methodologists that are interested in learning CANCEIS or improving their knowledge of this

system that is now part of the StatCan tool box. That course would be added to the list of courses within the Methodology branch.

Progress:

Two CANCEIS courses have been put together, one in English, the other in French. The format for each course is a two-day class, assisted with computers. The participants will learn how to develop deterministic and donor imputation modules, and become more familiar with the various parameters CANCEIS has to offer. Dates have been set for the fall of 2018, and we plan to offer each course once a year (based on demand).

For further information, please contact:
Lyne Guertin (613-862-6772; lyne.guertin@canada.ca).

1.10 Divisional Research - International Co-operation and Corporate Statistical Methods Division (ICCSMD)

Sub-project: Development of 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 program to implement these methods. This will allow us to study their effectiveness by applying them to some of our surveys.

Progress:

Specifications were written and we have now completed modules for the calculation of the estimated conditional bias for two-phase sampling with stratified srswor at each phase and two-stage sampling with srswor at first stage and Poisson sampling at second stage. Modules for single stage sampling are being developed from these two more complex programs. Two modules have also been developed for establishing coherence of the domain robust estimates for totals by producing domain coherent estimates satisfying various types of constraints. One module is for single dimension coherence while the other handles coherence in two dimensions such as province by North American Industry Classification System (NAICS). These modules can handle missing values (for cells with no sample) and structural gaps (for cells with no population). You can request that the final estimates meet one or more of the following constraints: (1) the estimated totals must satisfy bound constraints (2) the estimated totals must satisfy linear relationships between variables of interest (y1=y2+y3), and (3) the estimated totals for each margin (dimension) must be the sum of the corresponding cell estimates.

Sub-project: Consultation on the use of small area estimation methods

We provided consultation and support to methodologists exploring the possibility of using small area estimation methods in different surveys. In particular, we provided support for combining payment processor data (aggregated credit/debit transactions) with data from the International Travel Survey. The goal is to estimate total spending for each tourism region using payment data as the auxiliary information in the small area models.

Progress:

We provided information and existing manuals, when appropriate, to users on the small area estimation system component of the generalized estimation software. We had meetings to offer suggestions on how to improve the models given the particularities of data sets. In addition, we provided SAS and R code to perform explorations beyond the small area system. These consultations will most likely continue, particularly the one with the International Travel Survey, until the valid models for use in small area estimation are established.

Sub-project: Statistical capacity building

Different courses were prepared or delivered during the year:

- A course on robust estimation was given in French and English at Statistics Canada.
- A course on survey weighting was delivered at Statistics Canada to a delegation from Senegal.
- A course on weighting and robust estimation was given at the ISI meeting in Morocco.
- Development of a course on small area estimation began.

Consultations on different methodological topics were provided to methodologists in the Branch and, occasionally, outside the Branch. Examples of topics covered: sampling design issues, weighting, nonresponse treatment, bootstrap variance estimation, data analysis and small area estimation.

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

Sub-project: Automated methods of data collection, retail locations and shopping centres in Vancouver

To further enhance the data collected by the Monthly Retail Trade Survey (MRTS), shopping centre type was added as an attribute to the frame of retail store in the Vancouver region, using a mostly manual approach of cross-referencing with information available online. This research sub-project proposes a technology solution to streamline this process; reduce the instances of manual intervention; allow similar work to other regions and explore data collection alternative and possible frame validation or improvements.

The proposed pilot solution is based on the requirements gathered with stakeholders throughout Statistics Canada. The pilot-solution is a new web-based application that integrates Google Services Application Programing Interface (API), web scraping as the front end and SAS as a database management tool (back end). It is based on work done by Chang (2018) and Chen (2018). The new application uses reference data collected manually by Olineck (2017) from the Vancouver CMA from June 2017 and recently extracted web-based data to develop a prototype. The prototype leverages automated business services from Google Web Services, Near-by Search and Text Search and the web-scraping tools such as Scrapy and Selenium to extract information from the Internet. Together these technologies can provide direct access to publicly available data and reduce the instances of manual intervention in data acquisition.

Progress:

We have shown that, in general the prototype application effectively identified a comparable number of retail locations as the manual process in a more efficient and cost-effective manner.

In addition, the application completed the collection process in a matter of hours compared to several months for the manual process. These results suggest that web-based collection methods offer greater operational efficiencies when compared to manual processes. The project also helped identified and resolve many operations challenges with the use of the various technologies.

For further information, please contact:
Paul Holness (613-866-0367; paul.holness@canada.ca).

Sub-project: A Generic Statistical Business Process Model (GSBPM)

The current version 5.0 (2014) of the GSBPM has been adopted by over 50 jurisdictions including Statistics Canada. The first version 1.0 of the model was introduced in 2008. Since that time the model has undergone significant changes in terms of its strategic focus and design. However, over the past 10 years, we have witnessed a tremendous increase in the development, consumption and integration of data. The demand for real-time access to detailed data is growing at an enormous rate. Along with the increased volume of data, there has been a noted increase in the variety of input data sources including the Internet of things, web scraping, scanner data, electronic questionnaires and a further increase in the use of administrative data. The increased volume and variety of data have been further accelerated by the evolution and proliferation of a host new technologies which have increased the velocity at which we are generating data. These changes are driving the need for improved data management, particularly in the preparation of official statistics.

Progress:

This paper examines the Generic Statistical Business Process Model (GSBPM) and discusses how changes in the mapping of its components could improve the transparency of the data lifecycle and provide the basis for improved data management. The proposed model extends the functionality of existing sub-processes and breaks down the complex “Processing Phase” into its discrete components, “Profile & Discover,” Cleanse & Transform” and “Integrate” (join, link and model). It captures the output from each phase along with performance measures and data quality indicators. The result is a more flexible, transparent and accessible framework for producing official statistics. One that encourages end-to-end automation and the implementation of powerful new data science techniques to help improve data quality, drive program efficiency and build a more effective community of practice. The model was presented at the ModernStats workshop (Holness and Mayda, 2018).

For further information, please contact:
Paul Holness (613-866-0367; paul.holness@canada.ca).

2 Support Activities

2.1 Record Linkage Resource Centre (RLRC)

The objectives of the Record Linkage Resource Centre (RLRC) are to provide consulting services on 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.

Mixmatch exclusion and conversion table prototypes were integrated into Version 3.4 of G-Link and were tested on linkage projects of various sizes by the RLRC. In order to improve the Fellegi and Sunter classification and to reduce the burden of a manual review, the RLRC developed and tested a series of SAS macros to generate automated weight thresholds using various techniques (unsupervised automatic learning technique called K-mean automatic classification, two-phase automatic learning technique (K-mean followed by Probit modelling), a technique founded on the theory of extreme values to measure tail ranges, and finally, two techniques based on the linked and unlinked weight distributions of the profiles). Our record linkages using data from the project on the “tobacco court case” helped us to analyze software performance and the solutions to provide. The work on these data has contributed to developing more systematic and theoretically more coherent approaches for defining and adjusting record linkages on servers and on the SAS Grid platform.

The new alpha version of G-LINK 3.4 boasts many improvements, including an interface for loading data and conversion, exclusion and look-up tables. The RLRC contributed to the development of the G-LINK 3 user guide.

The RLRC assessed the SAS enterprise miner version in local mode to study the implementation of unsupervised (K-mean) and supervised (support vector machine, decision trees, etc.) automatic learning algorithms in order to classify pairs into two groups (true pairs and true non-pairs).

The inventory of record linkages done by the Methodology Branch was updated in 2017 and the results were presented.

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

2.2 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 subtopics with emphasis on the following:

- Consultation in time series (including course development);
- Time series processing and seasonal adjustment;
- Support for G-Series (benchmarking and reconciliation);
- Modelling and forecasting;
- Trend-cycle estimation.

Progress:

Consultation in time series

As part of the Time Series Research and Analysis Centre (TSRAC) mandate, consultation was offered as requested by various clients. Topics most frequently covered in the review period were related to the identification of break in series, interpolation and forecasting for various programs (education, justice, tourism) application of seasonal adjustment in various situations (for example, System of National Accounts, small area estimates and capacity utilization estimates for manufacturing) and specific applications of benchmarking and reconciliation. TSRAC members also contributed to the development of a directive and guidelines for time series continuity, jointly with the System of National Accounts (Statistics Canada, 2018).

TSRAC members continued their participation in various internal analytical and dissemination groups such as the Forum for Daily analysts and the forum on seasonal adjustment and economic signals. TSRAC members also met with international visitors (Manpower Singapore) and participated in exchanges both in person and virtually with other agencies including the Office for National Statistics, Eurostat, the US Census Bureau, the Bureau of Labor Statistics, and the Bureau of Economic Analysis to discuss current issues with seasonal adjustment and other time series tools and techniques under development. Several papers under consideration for journals and other publications on topics of benchmarking, reconciliation and seasonal adjustment software packages were also reviewed by TSRAC staff.

Time series processing and seasonal adjustment

This project monitors activities related to the support and development of the Time Series Processing System. Seasonal adjustment is done using X-12-ARIMA and X-13-ARIMA-SEATS (for analysis and development or production) or SAS Proc X12 (for production).

Several enhancements were implemented in the Time Series Processing System (TSPS) in order to prepare outputs needed for diagnostic reports and to make the system more robust and flexible in processing environments as required for the SAS grid (Ferland, 2017).

Initial work was done to empirically compare X12-ARIMA with the other methods for seasonal adjustment such as SEATS and other model-based methods. In addition, state space models were used to approximate each method and provide insight into the similarities and differences between the methods. This work will be presented at an upcoming conference of time series specialists.

An evaluation of interaction effects between trading days and holidays was also completed to determine if trading day effects can reasonably be assumed constant across different calendar months, notably those such as July or December with important holidays which may not exhibit the same change in level, given the different consumer habits surrounding certain holidays (Verret and Matthews, 2018).

The quality assurance process of Seasonal Adjustment was further developed, with an emphasis on clarifying client expectations, and documenting available methods in different situations (Matthews, 2018). Summary tools were prepared to help guide clients on resource requirements and what conditions need to be met to consider seasonal adjustment methods.

Support for G-Series (benchmarking and reconciliation)

This project entails the support and development of G-Series 2.0, which includes PROC BENCHMARKING and PROC TSRAKING, two SAS procedures, as well as the Macro TSBALANCING, which solves multi-dimensional reconciliation problems through a numerical approach. TSBALANCING was introduced with the release of version 2.0 and the training materials for reconciliation were updated to include this methodology. The macro has now been implemented for production for one survey and use will be expanded in other upcoming development work. Some of the challenges related to the application of these methods in the context of seasonal adjustment and some applied solutions were documented in Fortier and Ferland (2017).

A decommissioning plan was developed to guide the transition of production projects throughout the agency to the new version of G-Series. According to this plan, as of March of 2019, no internal production applications would be using earlier versions G-Series.

In terms of the methods available for benchmarking, research was done to identify and compare methods suitable for benchmarking of stock variables (Leung, 2018). Several options were considered, including benchmarking of first differences, and intermittent benchmarking to individual periods (closing inventories at end of year, for example). Of those identified, the most appealing is the direct BI method, which will be developed as a module in the Time Series Processing System. This module would include flexibility for more challenging aspects such as additive or multiplicative adjustments and the behaviour of the adjustment outside of the span of available benchmarks.

Modelling and forecasting

The recently acquired software SAS Forecast Studio (SAS/HPF) continued to be explored and used for various projects related to time series modelling. It proved to be an efficient preliminary tool to evaluate breaks in series and to detect large time series outliers. Forecasting techniques were applied in a number of projects to detect breaks in series in the absence of a formal parallel run. It has also been used to impute key units for periodic non-response and to forecast preliminary estimates for individual domains.

Specific research was conducted to become more familiar with state-space models. Benchmarking and reconciliation were expressed within this framework, as well as splines and other time series models (Picard, 2018; Picard, 2018b). A number of advantages were identified, including the ability to handle and treat missing data, the efficiency of the model estimation via a Kalman filter (relative to ARIMA and other models), and the somewhat sophisticated seasonal structures that can be specified. Before the approach could be used to develop seasonal adjustment for a production setting, a number of challenges would need to be addressed, such as the extent of revisions from adding new periods, and methods to make the method robust to extreme observations.

Additional analysis was explored, to estimate the effect of extreme weather on economic time series. While weather effects are not inherently included in seasonal adjustment, average weather patterns (climate) do tend to account for some of the periodic movements represented by the seasonal component. The approach that has been proposed by the Office

for National Statistics is very much in line with X12-ARIMA, using an ARIMA model with specific weather data-based regressors to estimate the effect of a given dimension of the weather on an economic time series. A paper was presented at the 2017 Joint Statistical Meetings to describe the approach and present examples using Statistics Canada data (Matthews and Patak, 2017). The approach was applied to monthly data on retail sales to explore if it would perform well on a larger scale, with more automation (Aston and Patak, 2018). In addition, progress was made toward simplifying the access to weather data used in the analysis. Weather data was also used in a project to explore modelling of tourism counts and found to be a useful explanatory variable. In both of these cases, the weather effects were validated by subject-matter experts and contributed to a better understanding aggregate data.

Analysis was carried out to explore modelling of high-frequency (daily) data, with daily frontier counts used for tourism. Development is required in general to detect and estimate patterns in higher frequency data, especially when the number of occurrences of the cycle is small. In this project, we were able to integrate monthly patterns from monthly data and weekly patterns from daily data, as well as integrating extraneous weather variables to explain the series. An evaluation of the model used demonstrates some potential to produce advance indicators or nowcasts of tourism counts on a more timely basis, using the daily data accumulated for the partial month. This work has been submitted for a talk at the 2018 Statistics Canada Methodology Symposium. A summary of initial work related to weather and daily data can be found in Fortier, Matthews and Patak (2017).

Trend-cycle estimation

Following up on the addition of trend-cycle lines added in *The Daily*, further documentation was prepared for external users, which is now available on the Statistics Canada website (Statistics Canada, 2018b). This documentation includes the precise weights used in the moving averages so that results can be reproduced by external users. In addition, the comparative study originally used to select the method used was re-applied to recent estimates and the findings confirmed the original choice. The expansion of trend-cycle estimates to other programs will be considered further. The use of trend-cycle estimates in the detection of residual seasonality is also currently being explored.

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

2.3 Research Data Centres and Confidentiality support

The research data centres (RDC) provide researchers with access to microdata from population and household surveys in a secure university setting. They are operated under the provisions of the *Statistics Act* in accordance with all the confidentiality rules and are accessible only to researchers with approved projects who have been sworn in under the Act as “deemed employees.” The role of the methodologist is to provide support to the RDC analysts and researchers on vetting requests. Methodologists also develop survey-specific guidelines whenever a new survey becomes available in the RDCs.

Progress:

A Methodology Expert Panel (MEP) on the creation of Public Use Microdata File (PUMF) was put in place. The team of methodologists has the mandate to review, guide and recommend

the approval of PUMF to the Microdata Release Committee (MRC). The MEP reviewed about half a dozen PUMFs this year. They also refined their process and prepared a seminar.

General support is provided to clients and other methodologists on disclosure control. The team in the Methodology Branch has been actively involved on a working group that has created a simple checklist for managers to assess the risk of disclosure from their products. Development is also sought on specific projects related to disclosure control, such as the Companion and Confid-on-the-fly.

The Companion was further developed to the point where it is ready for piloting in selected survey program areas. It has also added the capacity to advise managers on measures for reducing the risk of disclosure from their products that contain sensitive statistical information. Our continued communication with the Australian Bureau of Statistics concerning Confid-on-the-fly has renewed their efforts to resolve the reported issues with errors in multinomial regression and with parameters needed for the creation of perturbation tables required to produce its confidentialized outputs. A synthetic version of the MacKay file (data from the Survey on Financing and Growth of Small and Medium Enterprises) was produced. Preliminary investigations were also made into the methodology and workings of the R Synthpop package.

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

2.4 Quality Secretariat

The mandate of the Quality Secretariat is to promote and support the use of sound quality management practices across Statistics Canada.

The projects can be split into various subtopics with emphasis on the following:

- Capacity building;
- Quality indicators.

Progress:

Capacity building

In 2016/2017, the Quality Secretariat developed a two-day pilot course on quality management for middle managers in subject-matter fields. The feedback was so overwhelmingly positive that it was clear that the Quality Secretariat did not have the capacity to offer such training to all potential participants. Instead, this fiscal year, the material was re-packaged into modules and made available to all Statistics Canada employees on the Internal Communications Network.

Another vehicle for capacity building is the Quality Guidelines. A revision of the Guidelines was undertaken this year. The most significant improvements are in terms of accessibility and relevance. The revised Guidelines are aligned with Version 5 of the Generic Statistical Business Process Model (GSBPM), offering succinct and specific quality assurance practices for activities throughout the statistical process, and covering sub-processes such as data integration and preparation of national accounts.

Other federal government departments have enquired about the quality management tools used at Statistics Canada. We openly share the Quality Assurance Framework and the Quality

Guidelines, but these tools are very StatCan-centric. As the journey toward good quality management is long, with many small steps, the Quality Secretariat developed a Data Quality Toolkit, intended for anyone outside Statistics Canada who produces or uses data. The objective of the toolkit is to raise awareness about quality assurance practices. It offers two checklists, one for self-assessment by data producers, and the other to help users assess the fitness-for-use of a data set.

Quality indicators

There is growing need for tools to evaluate the sources of error and the magnitude of that error on administrative data. The Quality Secretariat worked with the Agriculture Division on providing the metadata for a preliminary admin-based release for the province of Alberta. For this situation, the established quality indicators for survey-based Statistics Canada releases were not applicable. We thus ensured that the note to users provided the requisite available metadata (e.g., timeliness), a warning about the lack of a true accuracy indicator, and the need to proceed with caution.

The Quality Secretariat was also approached by the Canadian Housing Statistics Program to discuss the data quality statement for their release in December 2017. It is clear that not only do we need new methods for measuring quality of administrative data, we also need new indicators and metadata to report the quality of data products composed in part or in whole of non-sample survey data.

These two collaborations led to the launch of a research project into methods to measure and report on the quality of non-sample survey data. The most technical challenge will be to develop a quantitative measure of accuracy, incorporating both variability (noise) and bias. The anticipated outcome of this research project is that metadata about data quality will facilitate informed decision-making by data users.

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

2.5 Data Analysis Resource Centre

The Data Analysis Resource Centre (DARC) is a team of statistical consultants and researchers in 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 analysts and researchers from academia or the research data centres (RDCs).

Progress:

Consultations

As part of DARC's mandate, consultation was offered as requested by various clients. Specific consultation services were provided to Statistics Canada's analysts from a dozen different divisions. These various consultations covered topics on the use of bootstrap weights, estimation of medians and their standard errors, tests of independence and other statistical

tests, estimation of confidence intervals, variance estimation for age-standardized rates, helping with SUDAAN and SAS SURVEY procedures, etc.

The group also provided services to other methodologists. These consultations included questions on Poisson regression, survival analysis, degrees of freedom for variance estimation, analyzing older General Social Survey (GSS) cycles, for which the mean bootstrap weights are given, together with the new GSS cycles that come with the "standard" bootstrap weights, etc.

External consultations were also delivered to a variety of clients from other federal and provincial governments. The requests included using STATA for analyzing Canadian Health Measures Survey (CHMS) data, statistical testing using the 2016 Survey on Sexual Misconduct in the Canadian Armed Forces, statistical inference using a census of employees with high non-response, etc.

Finally, expert advice was given to the analysts and researchers from the RDCs. The topics included bootstrap variance estimation, combining survey cycles, multiple imputation using SAS, longitudinal analysis and weights, etc.

Training

The team presented the newly redesigned special seminar for recruits, "Analyzing Data from a Survey with a Complex Design."

The course Statistical Analysis of Survey Data, Module 2, "Linear, Logistic and Generalized Logistic Regression Analysis", was redesigned and presented in the fall.

Several other training activities were developed and / or presented, in particular at the annual RDC analysts conference in November (with new presentations on structural equation modeling and weighting). An introductory course on descriptive analysis using survey data and presentations at the Data Interpretation Workshop were also delivered.

Collaboration with analysts

An article co-authored by Isabelle Michaud, in collaboration with Dr. Jean-Philippe Chaput and Suzy L. Wong (Health Analysis Division [HAD]), entitled "Duration and quality of sleep among Canadians aged 18 to 79", was published in the September issue of *Health Reports.* (Health Reports, Vol. 28, no. 9, pp. 28-33, September 2017 • Statistics Canada, Catalogue no. 82-003-X).

Another article entitled "The effect of reallocating time between sleep, sedentary or active behaviours on obesity and health in Canadian adults", coauthored by Rachel Colley, Isabelle Michaud, and Didier Garriguet was accepted for publication in *Health Reports*.

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), 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.

Progress:

Questionnaire review and consultation

During the review period, the QDRC responded to requests for expert reviews and consultation services on a variety of survey topics. Some survey themes included the Canadian Forest Services' questionnaire, the New Tourism Vision (various visitor exit surveys for three northern territories), the Newfoundland and Labrador Tourism Visitor Exit Survey, PRASC – Organisation of Eastern Caribbean States (OECS) Advocacy Project, Organisation for Economic Co-operation and Development (OECD) – Survey on Policy Responses to New Forms of Work, and Statistics Canada's Client Satisfaction Survey.

Training

The QDRC presented the three-day "Questionnaire Design" workshop (Course 410) three times in 2017/2018 (June, September and March). The course is offered as part of Statistics Canada's Training and Development Program.

Other work

The QDRC is an active member of several ongoing Statistics Canada committees related to questionnaire development, including a team which reviews and applies a LEAN process to the development of EQ for efficiency purpose.

For further information, please contact:
Paul Kelly (613-371-1489; paulkelly2@canada.ca).

2.7 Statistical Consultation Group

Sub-project: Reducing non-response Bias with replacements

The Teaching and Learning International Survey (TALIS) is a large-scale survey that focuses on the working conditions of teachers and the learning environment in schools. In this survey, samples of schools are selected at the first stage for each participating country. For each of these sampled schools, a pair of replacement schools is selected so that if a sampled school refused to participate, then the corresponding replacement school would be invited to participate survey so that total nonresponse would be minimized at the school level. Although this approach is supported in the literature (Chapman, 1976 and 1982; Platek, Singh and Tremblay, 1978), the investigation carried out by Chapman (1982) failed to validate the appropriateness of this approach as the studies were not carried out under ideal conditions. This research focuses on comparing the TALIS's current approach with three different approaches that are teacher imputation method, conditional probability estimation method and triplet clustering method using simulation studies to investigate the appropriateness of using replacements.

Progress:

Goal of this research to identify the appropriateness of the use of replacements in TALIS. This research is carried out in three parts. The first part focuses on identifying the characteristics of response rates for participating countries in TALIS. Based on the results of the analysis, appropriate assumption was arrived to carry out the next step. The second part pertains to the development of equations for above-mentioned approaches which was completed. The third and final part of the work requires simulating schools with similar characteristics as the schools from participated countries in TALIS and producing estimates to compare these above mentionned approaches. An extensive SAS program was created to simulate the required samples and produced the estimations. An R program was created to produce the graphical display of these results.

Documentation of the method is completed but the results of these analyses are soon to be added.

For further information, please contact:
Ahalya Sivathayalan (613-302-6647; Ahalya.Sivathayalan@canada.ca).

References

Chapman, D.W. (1976). A Survey of nonresponse imputation procedures. *Proceedings of the Social Statistics Section,* American Statistical Association, Part I, 245-251.

Chapman, D.W. (1982). Substitution for Missing Units. Bureau of Census.

Platek, R., Singh, M.P. and Tremblay, V. (1978). Adjustment for nonresponse in surveys. In *Survey Sampling and Measurement* (Ed., N. Krishnan Namboodiri), Academic Press, New York, San Francisco, London, 1978.

2.8 Knowledge Transfer - Statistical Training

In 2017-2018, there was a pause in the automatic scheduling of the large curriculum of courses under the statistical training umbrella. This time was used to think about new strategies to enable us to achieve our objective for capacity building and talent development.Generally speaking, the curriculum is now split into thematic blocks under the responsabilities of the appropriate resource centres (and relevant activities are reported in their respective section).For core survey methods, a small group of experts is reviewing and reorgonising the content, with new courses scheduled to start in 2018-2019. Moreover, two year-long activities focussing on a very active learning approach were successfully piloted for machine learning and data science.These learning groups were a hybrid between a reading and working group tasked with evaluating and organizing other learning activities for their peers.

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

2.9 Knowledge Transfer – *Survey Methodology*

Survey Methodology is an international journal available at http://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 2017 issues (43-1 and 43-2) were released in both PDF and HTML version. The June issue contains 7 regular papers. The December 2017 issue contains one special paper discussing the past, present and future of sample surveys followed by four short discussions of the paper, two regular papers and one short note.

Most of the journal's historical papers are available online (including all papers published after December 1981 (volume 7-2).) Older papers can still be obtained upon request. A subset of the papers published before issue 7-2 were selected based on their relevance, have been prepared, translated and added to the website.

From April 2017 to March 2018, Survey Methodology pages were viewed 48,500 times and 64,800 copies of papers were downloaded. 43 papers were submitted for publication.

A new online tool to manage the editorial process called Scholar One was tested and is being implemented. A special edition for the 9th colloque francophone sur les sondages (2016) will be released in December 2018. Some papers presented at a conference titled "*Contemporary Theory and Practice in Survey Sampling: A Celebration of Research Contributions by J.N.K. Rao*" have been selected and will be released in a special issue published in collaboration with the *International Journal of Statistics*.

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

3 Research papers sponsored by the Methodology Research and Development Program

Aston, J., and Patak, Z. (2018). Analysis of Retail Sales and Weather. Internal report.

Beaumont, J.-F., and Bocci, C. (2017). Recent experiments in Small Area Estimation. Paper presented at the Advisory Committee on Statistical Methods, September 2017, Statistics Canada.

Bocci, C., and Beaumont, J.-F. (2017). Small Area Estimation Experiments using CARROT data. Internal document, Methodology Branch, Statistics Canada.

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

Beaumont, J.-F., Estevao, V. and Haziza, D. (2018). Robust Estimation Prototype, Methodology Specifications, Statistics Canada.

Caron, P., and Chabot-Hallé, D. (2018). Use of the Household Survey Frame Service (HSFS) in Household Surveys. Presentation to the Household Survey Methods Division (HSMD) Technical Committee, February 2018.

Chang, C. (2018). Automated Methods of Shopping Centre Data Collection: Testing on Vancouver. Working paper No. 1. Ottawa: Statistics Canada.

Charlebois, J., and Laperrière, C. (2017). An extension of the Rao-Wu bootstrap to two-stage sample designs. Internal document (draft), Household Survey Methods Division (HSMD).

Chatrchi, G., and Gambino, J. (2017). CCHS sample matching experiment. Advisory Committee on Statistical Methods, October 2017.

Chen, H. (2018). Automated Methods of Shopping Centre Data Collection: Retail Locations and Shopping Centres in Vancouver. Working paper No. 2. Ottawa: Statistics Canada.

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

Dasylva, A. (2017). Dual System Estimation (DSE) with Linkage Errors. International Cooperation and Corporate Statistical Methods Division (ICCSMD) seminar.

Dasylva, A. (2017). Estimation of errors without clerical reviews in a deterministic linkage. Presentation to Household Survey Methods Division (HSMD) Technical Committee in Nov. 2017.

Dasylva, A. (2018). A new methodology for capture-recapture with linkage errors. International Cooperation and Corporate Statistical Methods Division (ICCSMD), internal report.

Dasylva, A., and Goussanou, P. (2018). Nonparametric Estimation of Thresholds and Errors in Record-Linkage. International Cooperation and Corporate Statistical Methods Division (ICCSMD), internal report.

Duddek, C. (2018). Comparing Manitoba Hydro meter data to the 2016 Census. Presentation to the Working Group #3 Dwelling Occupancy Status – Indicators and Model, 26 April 2018.

Favre-Martinoz, C., Haziza, D. and Beaumont, J.-F. (2018). Robust estimation for skewed distributions: A general approach. Paper submitted for publication.

Ferland, M. (2017). What's new in TSPS V3.06? Internal working paper.

Fortier, S. and Ferland, M. (2017). Seasonal adjustment and the balancing problem – Statistics Canada's experience, presented at the 2017 World Statistical Congress, organised by the International Statistical Institute, Morocco.

Fortier, S, Matthews, S. and Patak, Z. (2017). Impact of atypical weather on seasonal adjustment and time series analysis, presented at the 11th International Conference on Computational and Financial Econometrics, London.

Freeman, J., and Haddou, M. (2017). Stratification Algorithms: A comparison Study. Statistics Canada, internal report.

Gagnon, F., and Gravel, R. (2018). Challenges related to the use of payment data to estimate inbound tourism spending. Presentation to the Household Survey Methods Division (HSMD) Technical Committee, April 2018.

Guérendel, C. (2017). Estimation des erreurs de couplage d'enregistrements sans vérification manuelle. International Cooperation and Corporate Statistical Methods Division (ICCSMD), internship report.

Hidiroglou, M.A. (2017). Identifying outliers for SEPH. Internal document, Business Survey Methods Division (BSMD).

Hidiroglou, M.A. (2018). Minimum sample size requirements given the use auxiliary data. Internal document, Business Survey Methods Division (BSMD).

Hidiroglou, M.A., and Émond, N. (2018). Modifying the Hidiroglou-Berthelot (HB) method. Internal document, Business Survey Methods Division (BSMD).

Hidiroglou, M.A., and Kozak, M. (2018). Stratification of Skewed Populations: A Comparison of Optimisation-based versus Approximate Methods. *International Statistical Review*, 86, 1, 87-105.

Hidiroglou, M.A., Beaumont, J.-F. and Yung, W. (2018). Development of a Small Area System at Statistics Canada. To be submitted to the special 2019 *Survey Methodology* issue in Honour of J.N.K. Rao's contributions to small area estimation.

Holness, P., and Mayda, J. (2018). Presentation at ModernStats workshop, Geneva, Switzerland.

Jocelyn, W., and Tam, M. (2018). Modelling the income distribution to better predict the low-income portion. Internal document (draft), Household Survey Methods Division (HSMD).

Lafortune, Y., and Tam, M. (2018). How valid is SEF information for non-respondents? A partial answer provided by the results of the linkage between the LFS initial non-respondents and the SEF. Internal document (draft), Household Survey Methods Division (HSMD).

Leung, J. (2018). Benchmarking/Quarterization of Balance Sheet Totals. Internal working paper.

Logan, G., Jang, L. and Hidiroglou, M.A. (2017). Calibration for Domain Totals for Business Surveys. Proceedings of the Joint Statistical Meetings 2017 Survey Research Methods Section Baltimore, Maryland, July 29 - August 3, 2017.

Lola, P. (2017). Optimisation de la préparation des données pour le couplage d'enregistrements. International Cooperation and Corporate Statistical Methods Division (ICCSMD), internship report.

Matthews, S. (2018). Current Challenges with Quality Assurance of Seasonal Adjustment. Presentation prepared for the Seasonal Adjustment Practitioners Workshop, Washington, April 2018.

Matthews, S., and Patak, Z. (2017). Weather Adjustment of Economic Data – Beyond Seasonal Adjustment. Proceedings of the Joint Statistical Meetings of the American Statistical Association, Baltimore, Maryland.

Moffat, E. (2018). Using Machine Learning Algorithms on Realistic Synthetic Data for Record Linkage. International Cooperation and Corporate Statistical Methods Division (ICCSMD), internal report.

Neusy, E. (2017). Measuring and Reporting the Quality of Estimates. Presentation to the Methods and Standards Committee, November 20, 2017.

Olineck, C. (2017). Where did Vancouverites Go to Shop in 2016: A Snapshot of Vancouver Retail Store Sales by Shopping Centre Type. Ottawa: Statistic Canada.

Picard, F. (2018). Proc SMM to impute and interpolate. Internal working document.

Picard, F. (2018b). SSM for Seasonal Adjustment. Internal working document.

Rao, J.N.K., Rubin-Bleuer, S. and Estevao, V.M. (2018). Measuring uncertainty associated with model-based small area estimators. *Survey Methodology* (to appear).

Statistics Canada (2018). Guidelines on Maintaining Time Series Continuity in Economic, Social and Environmental Statistics Programs - Draft.

Statistics Canada (2018b). Details on calculation of trend-cycle estimates at Statistics Canada, http://www.statcan.gc.ca/eng/dai/btd/trend-cycle.

Stefan, M., and Hidiroglou, M.A. (2018). A procedure for estimating the variance of the population mean in rejective sampling. Submitted to *Statistica Sinica.*

Stefan, M., and Hidiroglou, M.A. (2018). Benchmarked Estimators for a Small Area Mean under a One-Fold Nested Regression Model. Submitted to *International Statistical Review*: accepted pending revisions.

Stefan, M., and Hidiroglou, M.A. (2018). Local polynomial estimation for a small area mean under informative sampling. Submitted to *Survey Methodology*: accepted pending revisions.

Stinner, M. (2017). Disclosure Control and Random Tabular Adjustment. 2017 Proceedings of the Annual Meeting of the Statistical Society of Canada, Survey Section. To be published.

Stinner, M. (2018). Disclosure Control and Random Tabular Adjustment. 2018 Proceedings of the Federal Committee on Statistical Methodology. To be published.

Sun, M. (2018). Evaluation of Machine Learning Algorithms for Record Linkage. International Cooperation and Corporate Statistical Methods Division (ICCSMD), internal report.

Tsui, J., Dasylva, A. and Chu, K. (2017). Optimal Estimating Equation for logistic Regression with Linked Data. International Cooperation and Corporate Statistical Methods Division (ICCSMD), working paper.

Verret, F., and Matthews, S. (2018). Some Discussion on Calendar Effects in X12-ARIMA. Presentation prepared for the Seasonal Adjustment Practitioners Workshop, Washington, April 2018.

Wright, P. (2017). Disclosure control that accounts for survey realities: assessing the risk using G-Confid. Paper presented at the Joint UNECE/Eurostat Work Session on Statistical Data Confidentiality, Skopje, 20-22 September 2017.

You, Y. (2017). Small area estimation under benchmarking and model misspecification. Research report, International Cooperation and Corporate Statistical Methods Division (ICCSMD).

You, Y. (2018). Report on the R package of the Fay-Herriot and spatial-temporal Fay-Herriot models with LFS application. Research report, International Cooperation and Corporate Statistical Methods Division (ICCSMD).

You, Y. (2018). Small area estimation using linear and log-linear models when sampling variances are estimated with MSM application. Research report, International Cooperation and Corporate Statistical Methods Division (ICCSMD).