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

Statistical Methodology Research and Development Program

Annual Report 2016/2017

Release date: November 3, 2017

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.

Standard table symbols

The following symbols are used in Statistics Canada publications:

. not available for any reference period
.. not available for a specific reference period
... not applicable
0 true zero or a value rounded to zero
0^s value rounded to 0 (zero) where there is a meaningful distinction between true zero and the value that was rounded
^{p} preliminary
^{r} revised
x suppressed to meet the confidentiality requirements of the *Statistics Act*
^{E} use with caution
F too unreliable to be published
* significantly different from reference category ($p < 0.05$)

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

This report summarises the 2016–2017 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. Acknowledgement are due to Dr. Pierre Lavallée who was in charge of the program for 2016-2017. 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).

The paper used in this publication meets the minimum requirements of American National Standard for Information Sciences – Permanence of Paper for Printed Library Materials, ANSI Z39.48 - 1984.
∞

Le papier utilisé dans la présente publication répond aux exigences minimales de l'"American National Standard for Information Sciences" – "Permanence of Paper for Printed Library Materials", ANSI Z39.48 - 1984.
∞

Statistical Methodology Research and Development Program Achievements

2016/2017

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 size 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 use such methods for the production of official statistics in a near future, if 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 1: On smoothing design variances in Small Area Estimation

In the basic Fay-Herriot methodology, the design variance of direct estimators is assumed to be known and fixed. In practice, it is never the case. An (approximately) design-unbiased estimator of the design variance can be used, but it is typically unstable due to small sample sizes. A common approach to handle this issue consists of smoothing the estimator of the design variance through an appropriate model, called the smoothing model. The goal of this project is to study some theoretical properties of smoothing and to compare a few alternatives via a simulation study.

Progress:

We have shown that, in general, the design variance is not only unknown but also random under the Fay-Herriot model. In addition, the goal is not to predict the unknown design variance but to estimate a smooth design variance. This has implications on the choice of explanatory variables to be used in the smoothing model. For instance, the explanatory variables should not include estimates from the survey, or any other random quantity, as is sometimes done in the literature.

We have also proposed a loglinear smoothing model with a back transformation that does not assume normality of the errors of the smoothing model. Some of these ideas have been described in Beaumont and Bocci (2016). We have completed a simulation

study to illustrate our findings and compared several choices of smooth design variance estimates via a simulation study. Our proposed approach performed well in general.

Sub-project 2: Empirical Best Linear Unbiased Prediction (EBLUP) and HB estimation under the Fay-Herriot area level model with estimated sampling variances

In small area estimation the Fay-Herriot model (Fay and Herriot, 1979) is widely used to obtain efficient model-based estimator for small areas. The sampling error variances are customarily assumed to be known in the model. However, assuming known sampling variances in the model is a very strong assumption. Rivest and Vandal (2002) and Wang and Fuller (2003) developed an EBLUP approach whereas You and Chapman (2006) developed a fully hierarchical Bayes (HB) model for modeling both the small area parameters and sampling variances at the same time. In this project, we consider two sampling variance models in HB framework and compare the HB estimators with the EBLUP estimator through a simulation study and real survey data analysis. In particular we compare the estimators under the situation when the model variance is relatively small compared to the sampling variances.

Progress:

We have finished the simulation study and applied both the EBLUP and HB approaches to LFS data. Our results show that HB approach performs better than the EBLUP approach, especially when the model variance is small compared to the sampling variances. A research paper has been finished and submitted to a journal for possible publication (You, 2016).

Sub-project 3: Analysis of Poverty Data by Small Area Estimation: a Book review

A book titled “Analysis of Poverty Data by Small Area Estimation”, edited by Monica Pratesi, was recently published by Wiley. This book contains 20 chapters which are written by different researchers and experts in the area of small area estimation (SAE) research and poverty data analysis. The book provides a comprehensive guide to implementing state-of-the-art SAE methods for poverty studies and poverty mapping.

Progress:

We have finished reviewing the book, and a book review has been written and published in the Survey Statistician (You, 2017a).

Sub-Project 4: Small area estimation using EBLUP, Pseudo-EBLUP and M-quantile approaches

Unit level small area models such as the nested error regression model are generally used to produce efficient model-based estimators in small area estimation. In this project, we consider to estimate small area means and quantiles using the EBLUP, pseudo-EBLUP and M-quantile approaches. In particular, we compare the estimates

of small area means and quantiles based on different scenarios including informative sampling and model mis-specification.

Progress:

We have finished the simulation study to compare the different estimators. A methodology branch seminar was given by a CANSSI student. A research paper was finished (Zhang and You, 2016).

Sub-project 5: HB small area estimation using unmatched log-linear models with different prior modeling on variance components and application to LFS unemployment rate estimation.

The Fay-Herriot model (Fay and Herriot, 1979) is a well-known linear mixed effects area level model used to improve small area direct survey estimates. In this project we consider a log-linear unmatched model (You and Rao, 2002) as an extension of the Fay-Herriot model with application in the Canadian LFS survey to estimate the local area unemployment rates. Following You and Chapman (2006), we also consider modeling the sampling variances directly. Particularly we consider and compare different prior models for the model variance and sampling variances under the unmatched models with application to the LFS data to evaluate the robustness of the HB estimates under different prior model specifications.

Progress:

Gibbs sampling programs have been developed to implement the procedure. We applied the proposed models to the LFS unemployment data. Our results show that the uniform prior and inverse gamma prior on the sampling variance lead to identical results in the LFS application. The log-linear Generalized Variance Function (GVF) modeling on the sampling variance performs slightly better, in particular for the areas with very small sample sizes. A research paper has been finished (You, 2017b).

Sub-Project 6: Bootstrap estimation of the Mean Square Error under the Fay-Herriot model

Estimation of the Mean Square Error (MSE) of the EBLUP estimator under the Fay-Herriot model requires several assumptions to hold, including the assumption that the errors of the linking model are homoscedastic. The estimated MSE is typically obtained through linearization and ignoring the variability due to not knowing the true smooth design variance. In this project, our goal is to develop a parametric bootstrap procedure that is robust against heteroscedasticity of the model errors and that accounts for the uncertainty in estimating the smooth design variance.

Progress:

A parametric bootstrap methodology has been developed and described in a document in progress. Some simulations programs were written to conduct Monte Carlo experiments. Those programs need to be refined due to efficiency considerations. The next step will be to conduct a simulation study to evaluate the properties of the proposed bootstrap approach and write a research paper.

Sub-Project 7: Development of a small area estimation methodology for the Monthly Survey of Manufactures

In the Monthly Survey of Manufactures (MSM), it would be desirable to obtain survey estimates of total sales at the industry level for each of 12 pre-determined Census Metropolitan Areas. The direct estimates at that level are not reliable. The goal of this project is to investigate the use of Small Area Estimation methods in order to produce sufficiently reliable domain estimates for the MSM. To achieve this goal, we use Goods and Services Tax (GST) data as our auxiliary variable.

Progress:

Our investigations were based on monthly data from the years 2009-2016. We first opted for the Fay-Herriot (area-level) model as the data at hand were subject to influential observations and a unit-level model may be highly affected by such observations. After testing different linking models and different smoothing methodologies, we chose to model take-some domain estimates using a linear linking model and a loglinear smoothing model. Outlier domains were observed in some of the months, which yielded unstable and possibly biased estimates. We thus developed a procedure that identifies outliers in both models using statistical tests based on standardized residuals. With the exception of outliers, our model diagnostics did not reveal any significant model failure. The small area estimates are under reviewed by analysts and are expected to be released in the next fiscal year. Our approach is documented in Bocci and Beaumont (2017).

Sub-Project 8: Local diagnostics for the Fay-Herriot model

Model validation in SAE estimation is a critical step. It gives confidence that the methodology used is reasonable (or not) with the data at hand. However, it does not tell the whole story if a user is interested in one particular domain. The goal of this project is to develop local diagnostics that better help understand the quality of a SAE estimator for a given domain.

Progress:

Four diagnostics have been developed and are currently being evaluated in a simulation study. They all use the model standardized residuals.

Sub-project 9: Challenges in fitting the Fay-Herriot model to survey data

The Fay-Herriot model consists of a sampling model and a linking model. The sampling model assumes that the direct survey estimator of the small area mean is equal to the small area mean plus a sampling error with mean zero and known variance. In the linking model, the small area mean is represented by a non-random linear term in the covariates, plus a random area effect.

The sampling model implies that the direct survey estimators are design-unbiased, but we often use as input calibration estimators like the Generalized Regression

(GREG) estimator. GREG are approximately design unbiased, but in some small area setups the GREG estimators are quite biased.

Another consideration is that we assume that the sampling variances are known. In reality, we do not know them, and we use a smoothed version of the estimated variances in lieu of the actual sampling variances. One of the challenges in fitting the model is the choice of smoothing variables.

In this project, we investigate via data simulation how the choice of input estimator and of smoothing variables affect the resulting Fay-Herriot EBLUP.

Progress:

We run simulations under different sampling designs using the SAS simulation macro developed by Estevao and Rubin-Bleuer (2016). The studies have shown:

1) GREG inputs to the Fay-Herriot model can produce small area estimators that are as biased as GREG and with very small reduction in the overall error (design-based MSE).
2) For GREG variance estimates, it is more difficult than for Horvitz-Thompson estimates to find smoothing variables that yield fitted variances close to the true variances.

Sub-project 10: Estimation of the design-based MSE for the small area unit level EBLUP estimator

Model-based estimators of the MSE of the EBLUP under the unit level model are design-biased estimators of the design-based MSE. We are investigating various design-based MSE estimators. The purpose of this project is to develop a new estimator of the design-based MSE that is reliable and approximately unbiased.

Progress:

We first considered a design-based MSE proposed by Jiang and colleagues, based on non-parametric bootstrap. We wrote the necessary code and showed that the non-parametric bootstrap MSE estimator overestimated the design-based MSE for areas with small sample size. Another design-based estimator of the MSE of the Best Linear Unbiased Prediction (BLUP) estimator developed by Jon Rao was shown through simulations to carry negative bias.

Sub-project 11: Estimation of the design-based MSE for the area level EBLUP

We know that model-based MSE estimators of the EBLUP estimators under an area level model are design-biased estimators of the design-based MSE and cannot identify outliers in the model. On the other hand, existing design-based estimators (are design unbiased but they not reliable. The purpose of this project is to develop a new estimator of the design-based MSE with better properties.

Progress:

A convex combination of the two estimators mentioned above was considered. It was showed, through simulations under various models, that the new combined MSE estimator is less biased and more reliable than the existing estimators of design-based

MSE. These results were presented by Professor Rao at the conference of the Statistical Society of Canada (SSC) in June, 2016 (Rao, Verret and Chatrchi, 2016).

Sub-project 12: Simulation SAS macro for small area estimation

When we produce small area estimates for a survey, we compare model-based MSE estimates with the variance of the direct estimator to make a decision. Model-based MSE estimators are design biased and do not detect outliers in the model relationship even if the model is correct. Model-based MSE estimates could be much smaller than the true error of the small area estimator under repeated sampling and can lead to the wrong conclusion on whether we should or should not release the small area estimates. The correct way should be to compare the variance of the direct estimator against the design-based MSE, which accounts for both variability and bias under repeated sampling. The purpose of this project is to develop a decision tool for producing small area estimates for any survey.

Progress:

Given a survey, this tool (or SAS macro) enables us to choose the best small area estimator for the data, detects small areas where the model might fail and provides an accurate range for area specific design bias and design MSE. We work with a finite population similar to that from where we draw the survey sample. This could be obtained by a non-parametric method of imputation, while ensuring that correlations between response and covariates are preserved. Given the finite population and the survey sampling design, the macro draws repeated samples and calculates design and model based measures of error for each small area estimator. In addition, it produces plots showing the distribution of the small-area estimates and the estimated mean square estimates for each estimator. The speed of the program makes it appropriate for use in actual applications to survey programs.

After the requirements were established, a simulation program was developed. With this program, one can input a population and select any number of samples under stratified simple random sampling without replacement. The simulation program permits to define the strata, calibration groups and the small areas quite generally. Also included are the calibration estimators, area-level EBLUP estimators with the Adjusted Density Maximization (ADM), Restricted Maximum Likelihood (REML) and Wang-Fuller methods of variance estimation and unit level EBLUP, You-Rao and Prasad-Rao pseudo-EBLUP and synthetic estimators. This macro can be used not only for making decisions on particular survey populations, but also for practical research applications.

Sub-project 13: The pseudo-EBLUP estimator for a weighted average with an application to the Canadian Survey of Employment, Payrolls and Hours.

The paper “The pseudo-EBLUP estimator for a weighted average with an application to the Canadian Survey of Employment, Payrolls and Hours” by Susana Rubin-Bleuer, Leon Jang and Serge Godbout was revised and published.

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

References

Fay, R.E., and Herriot, R.A. (1979). Estimation of income for small places: An application of James-Stein procedures to census data. *Journal of the American Statistical Association*, 74, 268-277.

Rivest, L.-P., and Vandal, N. (2002). Mean squared error estimation for small areas when the small area variances are estimated. *Proceedings of the International Conference on Recent Advances in Survey Sampling*, July 10-13, 2002, Ottawa, Canada.

Wang, J., and Fuller, W.A. (2003). The mean square error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 718-723.

You, Y., and Chapman, B. (2006). Small area estimation using area level models and estimated sampling variances. *Survey Methodology*, 32, 97-103.

You, Y., and Rao, J.N.K. (2002). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics, 30, 1*, 3-15.

1.2 Developmental Research - Record-linkage

Record linkage brings together records within the same file or across different files. It is an important tool for exploiting administrative data, including the creation of registers or analytical files and the maintenance of sampling frames in regular surveys. The research has three components. The first direction is the exploration of new methods for linking data. The second direction focuses on solutions for measuring linkage errors and the quality of linked data. The third direction is analysis and estimation with linked data.

Sub-project: Prototype for estimating linkage errors

The goal is developing a SAS prototype for estimating linkage errors including false positives and false negatives, for a probabilistic linkage, with and without clerical-reviews. The methodology has been described by Dasylva, Abeysundera, Akpoué, Haddou and Saidi (2016) and Fournier (2016).

Progress:

The SAS prototype takes as input the potential pairs, their linkage weights, the clerical-review sample size, and a clerical review sample. To date the following features have been implemented.

- Estimation of the mixing-proportion: This is the probability that a potential pair is matched. The estimated mixing-proportion provides the basis for estimating the error rates without any clerical-reviews. It is computed via the numerical

maximization of a partial log-likelihood, with the PROC OPTMODEL procedure under some assumptions.

- Validation of the input weights: The actual (unknown) mixing-proportion is a number between 0 and 1. This constraints provides the basis for validating the user-specified weights. This validation is done when the mixing-proportion is computed. A warning message is given to the user when the related partial log-likelihood has no stationary point between 0 and 1.
- Selection of a clerical-review sample: A sample for clerical-review may be selected according to different sample designs that include the following:
    a. Stratified Simple Random Sampling without replacement;
    b. Poisson sampling.
- Estimation of error rates without clerical-reviews: Model-based estimates are computed from all the potential pairs, under the assumptions used to estimate the mixing-proportion.
- Estimation of error rates with clerical-reviews: Point and variance estimates are computed using G-EST. Confidence intervals are also provided including intervals based on the normal approximation and on the bootstrap. The prototype also offers calibration estimators with an improved precision, when the clerical-review sample is not selected optimally, including point, variance and interval estimates.
- Confidence intervals for small proportions when estimating with clerical-reviews: Confidence intervals for small proportions have been implemented, based on the method proposed by Clopper and Pearson (1934).
- Estimation for clerical-review errors: Clerical errors are estimated when the user inputs a clerical-sample where each pair is reviewed by many conditionally independent reviewers. An expectation-maximization algorithm has been implemented to estimate the clerical errors and produce error rates that are adjusted for clerical errors.

This prototype is useful for measuring and improving the quality of linked data. It also reduces the costs of measuring the quality of linked data, both by providing error estimates that do not require any clerical-reviews and by optimizing the selection of the sample when clerical-reviews are used. The plan for the next version of the prototype is to add variances and confidence intervals for the situation the error rates are estimated without clerical-review.

Sub-project: Improving the E-M algorithm for G-LINK

A prototype E-M algorithm was developed for G-LINK in SAS, including interactions (Dasylva, Abeysundera, Akpoué and Quadir, 2015). It will be enhanced with hypothesis tests, diagnostics and model selection features. It will also be evaluated in scenarios involving survey data with a complex sample design, or administrative files that have a nonuniform coverage.

For further information, please contact:
Abel Dasylva (613-854-1918, abel.dasylva@canada.ca).

Sub-project: Evaluating the quality of preprocessing

In record-linkage, the common wisdom is that preprocessing has a major impact on the quality of linked data. However, there are few quantitative studies on this question. There is also no standard definition of what constitutes preprocessing within the agency or in the literature. The original goal was developing a methodology for evaluating the quality impact of preprocessing as well as comparing different preprocessing methods using this methodology. However, it quickly became clear that it was necessary to first define what constitutes preprocessing for record-linkage.

Progress:

Following a literature review, several prominent papers were identified, regarding the impact of preprocessing on data quality and linkage quality, including Randall, Ferrante, Boyd and Semmens (2013), and Ong, Mannino, Schilling and Kahn (2014). The paper by Randall et al. (2013) is of particular interest because it specifically addresses a fundamental question that is: Does preprocessing actually provide any significant benefits regarding the linkage quality? The authors propose a methodology based on synthetic data, which also includes several metrics for measuring the impact of preprocessing on the final linkage quality. The results show that preprocessing has a limited impact and that certain preprocessing operations are even detrimental. These findings are inconsistent with previous work, where preprocessing has been often said to have a very large impact on the final quality. However, it was noted that the validity of the study critically depends on what exactly qualifies as preprocessing. Furthermore, the synthetic data may not be sufficiently representative.

In parallel, an extensive environmental scan has been conducted to identify on-going efforts and existing solutions within the agency and elsewhere.

A generic model of preprocessing for record-linkage was developed. This model was positively received by important stakeholders. It has also served as a reference point for a corporate LEAN process on preprocessing for record-linkage.

The model and the findings have been described in a report, Holness, Sobrino, Dasylva, Trudeau and Pelletier (2017). The next steps are to propose a methodology to evaluate the impact of preprocessing activities on the quality of record-linkage and apply it to the Social Data Linkage Environment (SDLE). Also planned is the exploration of machine learning solution to preprocessing.

For further information, please contact:
Paul Holness (613 864-0176, paul.holness@canada.ca).

Sub-project: Adjustment for linkage errors

This project aims at comparing different methodologies for adjusting linkage errors through simulations.

Progress:

Chipperfield, Bishop and Campbell (2011) have proposed a maximum likelihood method to adjust for linkage errors when analysing contingency tables and performing

logistic regression with linked data. The logistic regression linkage error adjustment was implemented in the format of a distributable SAS macro and used in a simulation study under the assumption that the only source of linkage error are the incorrect links or false positives. Simulation results showed that the estimators are unbiased and exhibit smaller variance than those obtained from clerically reviewed data. Results also amply demonstrated the large bias of the estimators of the naïve method (using linked data without any adjustment for linkage errors) even when the linkage error is as low as 0.05. The findings have been summarized in a report, Chu, Saidi and Dasylva (2017).

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

Sub-project: Bayesian solutions

The goal is exploring the Bayesian approach for record-linkage. In theory, one likely advantage is the ability to compute the (conditional) joint probability that many pairs are matched, instead of being limited to estimating the marginal match probability of each pair. In that regard, the solution developed by Steorts, Hall and Fienberg (2013) is an interesting starting point. In this method, a Metropolis-Hastings algorithm is used to obtain the posterior distribution of a matrix comprising of indicator variables, where each such variable corresponds to the match status of a record-pair. A first step would be to extend this approach to consider many partial agreement levels, including a level for perfect agreement and another for disagreement, as in G-LINK; Statistics Canada generalized record-linkage system.

Progress:

A literature review has been conducted. The findings have described in a presentation, Dasylva and Labrecque-Synnott (2017).

For further information, please contact:
Félix Labrecque-Synnott (613 863-8137, felix.labrecque-synnott@canada.ca).

Sub-project: Linkage of businesses

This project aims at developing an effective methodology for linking businesses including the following features:

a. Standardization of business names and addresses for the Business Register (BR).
b. The elaboration of similarity measures and continuity indicators based on business names, addresses as well as any other information, e.g., labour tracking information.
c. The measure of linkage errors using clerical-reviews, logistic regression or some other model as appropriate.

This methodology is to improve linkage to the Business Register (BR), including longitudinal linkages such as the National Accounts Longitudinal Microdata File (NALMF); a database created by linking tax data and to the BR.

Progress:

A methodology was developed and presented to the technical committee in the Business Survey Methods Division (BSMD). It was also presented at the 2016 Symposium (Oyarzun and Wile, 2016) at the record-linkage workshop (Oyarzun, 2016), and in various other forum. The details are as follows.

Standardization of business names: After a thorough internal consultation within the Methodology Branch on current practices, a generic strategy was derived. This strategy was further improved by incorporating a few features from the Census solution for standardizing business names. The improved strategy was applied to the BR.

Standardization of addresses: A strategy for standardizing addresses has been developed in consultation with the Address Register within the Geography division, including the use of a geocode.

Similarity measures: A methodology was developed to compare business names and addresses and applied to tax records. For business names, the comparison is based on the Generalized Edit Distance (GED). Based on this distance, a first score is assigned. For addresses, the comparison is based on the geographical distance using Global Positioning System coordinates and it produces a second score.

Linkage decisions and related errors: For each record-pair, the linkage decision is based on its score and thresholds, to determine whether the pair is automatically accepted, rejected or to be resolved manually. A simple methodology was developed to quantify the likelihood that a given pair is matched based on its score.

For further information, please contact:
Javier Oyarzun (613 302-8454, javier.oyarzun@canada.ca).

Sub-project: Machine learning for record linkage

Record Linkage is the process of identifying pairs of records that refer to the same entity. Probabilistic record linkage, which was originally coined by Newcombe, Kennedy, Axford and James (1959) and formalized later by Fellegi and Sunter (1969), is currently widely used at the statistical agencies and health care services for matching records across databases which lack unique identifiers (e.g., SIN, DIN, etc.) for linking. Although Fellegi-Sunter classifier is deemed as optimal, the methodology requires multiple iterations to determine the prudent separation of definite and rejected pairs and to resolve the pairs in the grey zone (in between upper and lower thresholds), which is time consuming and laborious.

To get an educated guess of the weight thresholds and help reduce the burden of clerical review, this project aimed at classifying the pairs based on alternative classification algorithms called machine learning. The methodology applied different machine learning classifiers which mainly fall into two different categories – unsupervised and supervised. Unsupervised classifiers are based on clustering,

graphing techniques, while supervised approaches underline the principle of regression modelling where subset of the pairs (i.e. training data) requires true match status which is used to train the classifier and to predict the match status of all pairs.

Progress:

A review of the application of machine learning algorithms in record linkage was performed with a focus on the following three algorithms:

- Two unsupervised solutions: K-means clustering and Farthest-First classification.
- One supervised solution: Support Vector Machine (SVM).

This project also reviewed the following software to get an insight into the implementation of machine learning algorithms in record linkage:

- FEBRL (Freely Extensible Biomedical Record Linkage), a record-linkage software package developed in PYTHON and used at the Australian Bureau of Statistics.
- SCIKIT-LEARN, a PYTHON based machine learning package.
- RECORD LINKAGE package built in R.

K-means clustering and Farthest-First algorithms of FEBRL and SVM of SCIKIT-LEARN were used to classify the pairs of G-Link tutorial and NPHS-DEPOT, an SDLE (Social Data Linkage Environment) project. Considering Fellegi-Sunter classification as gold standard, K-means clustering and Farthest-First classification exhibited about 99% of sensitivity and specificity for G-Link tutorial and about 94% and 99% respectively for NPHS-DEPOT project. Supervised classifier SVM showed nearly 100% of sensitivity and specificity with sampled training data set where Fellegi-Sunter match status was deemed as true match status.

As the empirical results were quite promising, a prototype of estimation of automated weight thresholds (Quadir, 2017) based on K-means clustering (using SAS PROC FASTCLUS) has been built and its implementation in G-Link is currently underway.

The study also revealed that the reviewed open-source applications (FEBRL, RECORDLINKAGE) are typically unable to handle medium-to-large linkage projects for their memory-dependent processing, so they would not be very helpful for accomplishing the record linkage projects at the agency. As the backbone of our generalized system (G-Link) is SAS, building prototype of machine learning classifiers for record linkage in SAS is the next step.

For further information, please contact:
Tanvir Quadir (613 863-7806, tanvir.quadir@canada.ca).

Sub-project: New building blocks for record-linkage

This project critically looks at the potential of specific technology components or methods for helping the agency meet its future record-linkage needs, in relation to major initiatives such as the census transformation project and the much anticipated

use of big data sources. These components include innovative and massively scalable blocking strategies (Christen, 2012, Chap. 4), graph databases (Harron, Goldstein and Dibben 2016, Chap. 7), and the protocol suite developed by the World Wide Web Consortium (W3C) for distributed databases that are linked dynamically (Mitchell and Wilson, 2012), including the Resource Description Framework (RDF) and the Web Ontology Language (OWL). The goal is formulating a set of recommendations regarding the use of these building blocks in a manner that is compatible with G-LINK.

Progress:

The completed review (Mizdrak, 2017a, and Mizdrak, 2017b) has identified ways to enhance G-LINK, including blocking methods based on machine learning and graph databases for effectively managing multi-file linkages, such as when building a register from administrative files. Some of the identified solutions have been recommended for a proof of concept and a further evaluation against specific requirements. These findings have been summarized in two reports.

For further information, please contact:
Predrag Mizdrak (613 617-8563, predrag.mizdrak@canada.ca).

References

Chipperfield, J., 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.

Christen, P. (2012). *Data Matching*, New York: Springer.

Clopper, C.J., and Pearson, E.S. (1934). The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26, 404-413.

Dasylva, A., Abeysundera, M., Akpoué, B. and Quadir, T. (2015). Expectation-Maximization (E-M) Algorithms for Probabilistic Record Linkage: Methodology, internal report.

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

Harron, K., Goldstein, H. and Dibben, C. (2016). *Methodological Developments in Data Linkage*. Chichester: John Wiley & Sons, Inc.

Mitchell, I., and Wilson, M. (2012). Linked data: Connecting and exploiting big data, Fujitsu white paper, available at www.fujitsu.com\uk.

Newcombe, H.B., Kennedy, J.M., Axford, S.J. and James, A.P. (1959). Automatic linkage of vital records. *Science*, 130(3381), 954-959.

Ong, T.C., Mannino, M.V., Schilling, L.M. and Kahn, M.G. (2014). Improving record linkage performance in the presence of missing linkage data. *Journal of Biomedical Informatics*, 52(2014), 43–54.

Randall, S.M., Ferrante, A.M., Boyd, J.H. and Semmens, J.B. (2013). The effect of data cleaning on record linkage quality. *BMC Medical Informatics and Decision Making*, 13:64.

Steorts, R.C., Hall, R. and Fienberg, S.E. (2013). A Bayesian Approach to Graphical Record Linkage and Deduplication. Preprint in arXiv: http://arxiv.org/abs/1312.4645.

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 generalised systems, the team also take on development research related to disclosure control, data visualisation, variance estimation and other survey methods on which the systems are built upon.

Sub-project: Disclosure control - Protecting sensitive data when prior information has been released

The current disclosure control process does not explicitly take into consideration data that is either revised or released as part of an on-going program. In this sub-project, two specific scenarios are considered: (i) controlling the disclosure of the current month after releasing the results of prior months, and (ii) controlling the disclosure of a revised release given the release of preliminary results for the same reference cycle.

Progress:

To take into consideration when prior information has been released, a set of approaches to use the weighted cost of cells that favours the suppression of previously suppressed cells, and/or to favour the publication of previously published cells was investigated. Using the real data, GenSys identified that strictly favouring the suppression of previously suppressed cells best met the success criteria under certain assumptions. The study is summarized by Wright (2016). The team also proposed a risk assessment and threshold approach based of the inherent sensitivity of a cell in the context of a revised release given the release of preliminary results for the same reference cycle. Finally, risks in the context of another scenario -- for which the definitions of the domains of publication were refined – were identified.

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 simulated data. Testing using real data has stated. Overview presentations explaining the method and its potential to analysts were prepared and given. A more technical version of the material was presented to the Advisory Committee on Statistical Method (Gray, Stinner, Wright and Thomas, 2016) and prepared for the annual Statistical Society of Canada meeting (Stinner, 2017). Due to the high interest and promising results, future work on this is given a high priority.

Sub-project: Data visualization for use with survey sampling

This project aims to explore, identify and propose the use of data visualization methods at the various stages of survey process. This study is co-investigated with Data Analysis Resource Centre (DARC).

Progress:

After a targeted literature review, consultation with internal users/developers and exploratory work with various software, a report illustrating some possibilities was produced (Fung, 2017). Examples of simple and uncommon charts for categorical data were also added in DARC's training material.

Sub-Project: Sampling variance of the calibration estimator under bound restrictions

The linearization technique used to provide an approximation of the variance of the calibration estimator does not consider the operational constraints (non-negativity, multiple of design weight, etc.) that can be imposed on the calibration weight. While the impact on the variance should be negligible under certain conditions (the sample sizes are quite large and the number of active constraints is small). These conditions are not always respected in practice. The purpose of this project is to expand the linearization procedure to consider the bound restrictions on calibration weights and evaluate the impact on the variance estimate.

Progress:

The theoretical foundations have been formulated. They are founded on the concept of active constraint in operational research and the standard Taylor linearization technique. A summary prototype has been constructed. Initial simulations were conducted but they were non-conclusive in terms of the impact on the variance (it is possible that the data set used was insufficient for adequately determining the impact of the method). However, the simulations confirmed that the approach would considerably increase the processing time.

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:

Generalized Systems designed and developed 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). This approach harmonizes the identification of domains and questionnaire flows for both G-Est and G-Confid, and facilitates troubleshooting when inconsistencies occur.

In preparation of a future development initiative, the team gathered user input on Banff from both users at Statistics Canada via focus groups, and also external users via questionnaires (Finch, 2017).

Finally, the team provided continuous support to users, updated and delivered training presentation in various forum and met with international delegates to discuss on-going and future development of the generalised systems.

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 2016-17 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 continuing projects from last year while the other ones are new initiatives. Other data collection research initiatives are conducted by or in collaboration with the Collection Planning and Research Division.

Sub-Project: Developing a prioritization score for assigning cases to interviewers as they become available in a computer-assisted telephone interview (CATI) survey

The objective of this project is to develop a procedure for calculating a prioritization score to identify which case will be interviewed using the Integrated Collection and Operations System (ICOS) when an interviewer becomes available in the CATI surveys. The algorithm has three components. The first takes into account information already recorded regarding arrangements for meetings with the respondent, 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 under way, 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 resident can be contacted within a given time slice.

Progress:

The general form initially proposed for calculating the score has been modified slightly to improve control over the selection process to determine the case for the next

interview attempt (Kleim, 2016). A method based on Bollapragada and Nair (2010) that seemed promising to predict the probability that a household resident can be contacted in a given time slice has been tested. Unfortunately, the results observed were rather disappointing (Miville and Ubartas, 2016). A second method in which modelling this probability relies on the use of auxiliary information available from the individual and on a self-learning nature in real time has been considered (Sallier and Kleim, 2017). This method is actually an adaptation of a hierarchical Bayes model applied in the medical field (McCormick, Rudin and Madigan, 2012). This approach is still being explored.

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

Sub-Project: Segmentation project

In order to derive efficient communication and operation strategies for increasing response rate for social surveys, one could segment the population into homogeneous groups (clusters or segments) using socio-demographic and collection-type (response rate, response mode) metrics and variables and leverage on the clusters' characteristics to develop strategies. The challenge for social surveys is to find a comprehensive database containing timely information for the Canadian population including response metrics derived from social surveys. This project has for objective to create segments or clusters of population profiles that will help increase response rates on two fronts. Learning more about our potential respondents will help us derive efficient communication strategies (publicity, letter of introduction, etc.) and also help the organization develop efficient collection strategies by understanding our potential respondents (difficult to reach population, easy to contact, high response propensity, language, etc.).

Progress:

As proof of concept, tests were performed using 2011 National Household Survey (NHS) data and 2011/2012 Canadian Community Health Survey, Survey of Financial Security and Survey of Household Spending data. First, logistics regression models were run to identify household characteristics associated with household surveys nonresponse and then groups of geographical areas (currently Census Subdivision (CSD)) were created. These groups show similar prevalences for the identified characteristics. Initial results are promising but come with a strong limitation that they come from a 15% success linkage between NHS and survey data. An investigation on the use of applying an importance factor to the clustering strategy (i.e., making some variables more important than others) has been performed. The strategy and the initial results were documented and presented in various forums (Halladay, Brisebois and Yang, 2017; Brisebois, 2016). The methodology will be further developed and evaluated using information from the 2016 Census.

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

Sub-Project: Response mode preference

The issue of declining response rates is a major challenge for Statistics Canada and many other statistical agencies. One way to address this issue is offering respondents several ways of completing questionnaires (for instance, online or using paper) with the hope of increasing response rates. The trade-off for this, however, is cost: electronic questionnaires (EQ) are more costly to set up, and paper questionnaires (PQ) are more costly to collect and process. In order to evaluate appropriately the costs and the benefits of a multimode collection strategy, it is important to determine which factors affect the choice of a response mode and to understand how much we can expect to sway respondents to choose EQ over PQ by encouraging them to fill out questionnaires online.

Progress:

The extent to which the mode of contact affects the mode of response can be evaluated using the 2014 Canadian Census Test. A sample of dwellings was taken and randomly assigned to different panels; some panels were sent an invitation letter to complete a questionnaire online and others were sent a paper questionnaire directly. The response mode choice from the Census Test could then be compared with the choice made during the 2011 Census, while controlling for contact mode. Census Test response rates in different panels, by mode of response, show that our mode of initial contact is a very strong factor in determining the mode of response and allow us to quantify this effect.

The analysis was completed and the results documented (Romanyuk and Boulet, 2017).

For further information, please contact:
Yuliya Romanyuk (613-862-1193, yuliya.romanyuk@canada.ca).

References

Bollapragada, S., and Nair, S.K. (2010). Improving right party contact rates at outbound call centers. *Production and Operations Management*, Vol. 19, No. 6, November–December 2010, 769-779.

McCormick, T.H., Rudin, C. and Madigan, D. (2012). Bayesian hierarchical rule modeling for predicting medical conditions. *The Annals of Applied Statistics*, Vol. 6, No. 2 (June 2012), 652-668.

1.5 Prospective Research - Non-probabilistic approaches

Sub-project: Alternative data and sample matching

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 the use of Web panels and the problem of making these surveys probability surveys. To do this, he proposed the application of *Sample Matching.* In this method, a probability sample is drawn from a sample frame, and the sample is linked to the Web panel respondents using statistical matching. With statistical matching, each individual in the probability sample is matched with a panellist based on given characteristics; but exact matches are not required.

The purpose of this research progress is two-fold. The first goal is to assess to possibility of using Sample Matching for some of the programs of Statistics Canada. This approach would be used, for example, for programs where response rates obtained through traditional collection methods are relatively low. The second goal is related to the exploration of alternative data sources.

Progress:

This research project was presented to the Advisory Committee on Statistical Methods (ACSM) on May 15, 2016. There was a detailed presentation on sample matching. There was also a comparison with two other methods used in similar situations: indirect sampling (Lavallée, 2002; 2007) and model imputation. A test was done to compare the quality of the estimates derived from each of the three methods. This simulation used the Labour Force Survey (LFS) to validate the concepts. The LFS sample from Prince Edward Island used a representative sample derived from the frame, and the sample from the rest of Canada represented the Web panel. The results showed that the three methods can be used, in practice, to produce estimates of acceptable quality.

Following the comments received during the presentation to the ACSM, research continued using new perspectives. According to an article from Bethlehem (2014), it would be possible to make the data from Web panels usable in a probabilistic context by calibrating them on known totals to correct possible issues of representativeness. Lee (2006) also proposed a reweighting method. He uses a propensity score to adjust the error due to non-participation in the panel.

Two new sample matching applications were tested. The first concerns the use of data from the National Household Survey (NHS) and the Labour Force Survey (LFS). In this application, the LFS served as a pseudo-Web panel, while the NHS served as a population to select probabilistic samples. The second application concerned health care. A test was conducted to see if certain Canadian Community Health Survey (CCHS) questions could be asked through a Web panel. Sample matching would be used in this case to link information gathered by the CCHS Web panel to produce estimates.

Several program-sponsored initiatives linked to exploration and the adequate use of alternative data sources are under way. In particular, there has been progress in the use of scanner data for consumer prices, for which automated classification of information presents a challenge, the exploration of data from credit and debit card

providers for household surveys, and the identification of data sources for health surveys. These last two projects have, among other things, helped identify several operational issues related to the acquisition of such data. Two methodological challenges in relation to big data are also present. Two groups have been established to study, moderate or eliminate the impact of those issues: a working group on the issues and best practices regarding the use of administrative data and a reading group on big data and data science.

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

References

Bethlehem, J. (2014). Solving the nonresponse problem with sample matching? *Statistics Netherlands Discussion Paper*, The Hague/Heerlen (Netherlands), 2014.

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

Lavallée, P. (2002). *Le Sondage Indirect, ou la Méthode généralisée du partage des poids.* Édition de l'Université de Bruxelles (Belgium) and Éditions Ellipse (France), 215 pages.

Lavallée, P. (2007). *Indirect Sampling*. Springer, New York.

Lee, S. (2006). Propensity score adjustment as a weighting scheme for volunteer panel Web survey. *Journal of Official Statistics*, 22, 329-349.

1.6 Prospective Research - Operational Research

Operational research is a vast branch of mathematics which encompasses many diverse areas of minimization and optimization. As part of prospective research, we are exploring *if* and *how* operational research could be beneficial in the context of survey methodology and official statistics.

Sub-project: Operational Research

The goal of the project is fully explorative and no specific problems are expected to be solved at this point. As part of the prospective work, we expect to:

- Document and identify areas in survey methodology where approaches related to operational research and optimisation are used, or could be useful;
- Review similar uses and research and development (R&D) work in a few other agencies, if applicable;
- Identify areas of knowledge gap and/or needed improvements;
- Identify areas of future research/explorative work.

Progress:

The longer term goals of the explorative work were established. Through previous and information discussions of the working group on operational research, some areas in survey methodology where approaches related to operational research and

optimisation are used or could be useful have been identified. The review of similar uses and R&D work started by, amongst other activities, attending the 22nd International Conference on Computational Statistics. A pertinent application related to sample allocation was identified and will be further explored. Challenges related convergence and numerical precision were identified in the tool currently used. To address one area related numerical precision in optimisation problem, an alternative formulation of a specific problem related to sample allocation was proposed and is being implemented.

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

1.7 Prospective Research - Measurement Devices

Sub-project: Evaluation of different models of activity monitors for the objective measurement of sedentary behaviors, physical activity and sleep in the CHMS

An activity monitor is used by the Canadian Health Measures Survey (CHMS) at Statistics Canada to objectively measure Canadian's physical activity and sedentary behaviors. The model of device used is the Actical accelerometer, worn at the waist during day time. Since this model was chosen for CHMS, the accelerometer's technology has tremendously evolved (Troiano, McClain, Brychta and Chen, 2014). Also, researchers are now interested in objectively measuring sleep duration and quality as well. The recently released Canadian 24-Hour Movement Guidelines for Children and Youth (Tremblay, Carson, Chaput, Connor Gorber, Dinh, Duggan and Janssen, 2016) integrate recommendations for physical activity, sedentary behaviour and sleep into a single guideline.

The main objective of this research project was to evaluate different models of activity monitors, while taking into account the CHMS objectives and constraints, and to provide information to Health Statistics Division (HSD) regarding the activity measurement by the CHMS in the future.

Progress:

Two reports were written, in collaboration with Didier Garriguet and Rachel Colley (Health Analysis Division), and shared with the HSD clients. The first one (Michaud, Colley and Garriguet, 2016) included an extensive literature review of scientific papers to provide objective information on the evolution of activity monitors' technology and on the direction the field of activity measurement was going. Recommendations on the model of device, monitor location and type of output data were provided. Based on this, the Actigraph accelerometer worn at the waist for 7 days (24h/day) was selected. The changes are planned to be applied for the Cycle 7 of CHMS.

The second report (Michaud, Colley and Garriguet, 2017) gave clients more information on the Actigraph device and also provided the experimental design for the pilot study that is planned to be done during the Cycle 6 of CHMS. The pilot study will consist of testing the feasibility of the recommended Actigraph model, and

comparing the estimates between the current (Actical) and new (Actigraph) device and between the current (day-wear) and new (24h-wear) protocol.

For further information, please contact:
Isabelle Michaud (613-762-9747, isabelle.michaud@canada.ca).

References

Tremblay, M.S., Carson, V., Chaput, J.P., Connor Gorber, S., Dinh, T., Duggan, M. and Janssen, I. (2016). Canadian 24-hour movement guidelines for children and youth: An integration of physical activity, sedentary behaviour, and sleep 1. *Applied Physiology, Nutrition, and Metabolism*, *41*(6), S311-S327.

Troiano, R.P., McClain, J.J., Brychta, R.J. and Chen, K.Y. (2014). Evolution of accelerometer methods for physical activity research. *British Journal of Sports Medicine*, *48*(13), 1019-1023.

1.8 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: Developing quality measures for statistics derived from survey and non-survey data

Many programs are now synthesizing information collected from multiple sources. These sources may include “expert opinion” type estimates that are based on a combination of surveys, other industry sources and subject matter expertise/judgment. Estimating variances or a quality measure in that context is challenging. The goal of this research is to follow up on recommendations received from the Technical Committee of the Business Survey Method Division where a pseudo-bayesian approach to assess the variance of these partial-survey and/or non-survey estimates was proposed.

Progress:

During the review period, the problem was stated in the Bayesian context (Chen and Jamrov, 2016) with a continuous prior distribution that needs to be approximated with a discreet variant. An empirical-Bayesian framework has been fully developed that requires further Taylor linearization. Resampling methods on the observed data are considered to derive the discreet prior distribution.

For further information, please contact:
Sanping Chen (613-854-2466, sanping.chen@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, geometric and Lavallée-Hidiroglou (1988), and to see the impact of extended algorithms to aim for optimal solutions.

Progress:

A paper entitled “Stratification of Skewed Populations: A Comparison of Optimization-Based Versus Approximate Methods” was written and submitted to the International Statistical Review journal. Minor comments were received and the revised version of the article was re-submitted to that journal on March 15, 2017.

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

Sub-Project: Survey of Employment, Payrolls, and Hours (SEPH)

The Survey of Employment, Payrolls, and Hours is currently using a calibration procedure that is theoretical sound but has a number of weaknesses in practice. They include: Multiple weights for each sampled unit and negative weights resulting in negative estimates. The objective of the research was to suggest ways of simplifying the existing estimator, with the following objectives: (i) Have one weight per establishment; (ii) reduce the number of negative weights,(iii) Improve outlier detection and treatment; (iii) identify “true” outliers using standard available methods; (iv) use the Generalized Estimation System (GES2.0) for estimation and variance estimation.

Progress:

The SEPH existing estimation methodology was standardized, and a report describing this was written. An extensive study was carried out to compare four different estimation and weighting methods. These included: (i) Horvitz-Thompson estimator (HT); (ii) Generalized regression estimator fixed at national level (GREG1; calibration at industry group); the Generalized regression estimator fixed at sub-national level (GREG2; calibration at industry group by province grouping); and the current estimator (iv) The generalized regression estimator with domain-specific calibration (GREG3; calibration at NAICS4 industry group by province). An extensive study on these estimators revealed that GREG2 is better than GREG1 when there are enough responding units in the calibration group (more A level quality estimates). However, GREG2 yields poorer quality estimates (F level) because there are currently not enough units at the calibration level, which is calibration industry by provincial grouping. The current SEPH estimation and variance outputs can be produced using the Generalized Estimation System (GES 2). This was made possible via a number of suggestions. Work also started on how to improve the existing outlier detection and treatment method. It is based on work done by Beaumont, Haziza, and Ruiz-Gazen (2013).

The results of this work will enable to achieve the above mentionned objectives of this subproject.

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

Sub-Project: Exploration of the practical issues associated with calibration in business surveys

Calibration is the default estimation method in the Integrated Business Statistics Program's estimation system. This method was effective for the first surveys that were integrated and that primarily measured income and expense variables. With more and more different surveys being integrated, some practical issues were raised. Thus, a working group was established to gain a better understanding of the current issues and provide possible solutions.

Progress:

Studies were initiated to understand the calibration issues for certain IBSP surveys. Estimates using the calibration method were compared with estimates without calibration for the auxiliary variable, as well as for the primary variables of interest and by industry. The correlation between the variables of interest and the auxiliary variable has also been studied. The results of these studies have been documented (Nolet-Pigeon, 2017a; 2017b).

For further information, please contact:
Marie-Claude Duval (Marie-Claude.Duval@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.

Lavallée, P., and Hidiroglou, M. (1988). On the stratification of skewed populations. *Survey Methodology*, 14, 1, 33-43.

1.9 Divisional Research - Household Survey Methods Division (HSMD)

The research sub-projects submitted to this project must be in keeping with the Methodology Branch's mandate, and more specifically, with the Household Survey Methods Division's objectives. Thus, the sub-projects must aim to (1) develop innovative procedures to keep up with changes while budgetary constraints and demands increase; (2) improve the efficiency of survey plans, and thereby reduce costs; (3) ensure the highest quality results by proposing innovative and effective statistical methods that are integrated in the agency's statistical programs.

Sub-Project: Modeling the income distribution to better predict in the low income portion

The empirical distribution of family income based on the Survey of Labour and Income Dynamics (SLID) data and on the family file developed from personal tax returns (T1FF) differ, particularly at the extremes. The objective of this project is to model family income using a parametric income distribution model by using information from multiple sources to better "predict" income in files where this information is not available.

Progress:

The generalized beta of the second kind (GB2) distribution was chosen to model income from the Survey of Labour and Income Dynamics (SLID), the Canadian Income Survey (CIS) and the T1FF. Some adjustments to the concepts and the family unit definition were made for the purposes of computing equivalent income from the different sources. Optimization procedures in SAS were tested to derive the maximum likelihood estimates of the parameters of the GB2 distribution.

Based on the comparison of summary measures (mean, standard deviation, and median), QQ-plots, and Gini coefficients, the GB2 distribution appears to well represent the income distribution implied by the SLID/CIS and T1FF data.

Comparisons of the entire income distribution between different years of the SLID/CIS and the T1FF, respectively were made by looking at the GB2 parameters and at graphs of the GB2 density. In general, for both SLID/CIS and the T1FF, there was an increase in the scale parameter leading to a decrease in the mode and a slight shift of the income distribution to the right. However, the shape parameters of the GB2 did not change much over time. Comparisons of the GB2 parameters were also done between SLID/CIS and the T1FF. The scale parameters were generally comparable, but the shape parameters showed greater discrepancies. Therefore, some modification to the T1FF GB2 parameters may be necessary in order to use the T1FF GB2 distribution to predict income for SLID/CIS.

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

Sub-Project: Extension of the Rao-Wu Rescaling Bootstrap method to two-stage sample designs without replacement

Implementing the Rao-Wu Rescaling Bootstrap method is warranted for a survey with a sample design with more than one stage provided that the first-stage sample is with a replacement. Using this method is warranted when this sampling is done without a replacement provided that the sampling fraction used is negligible, or a modification taking into account the first-stage sampling fraction is applied. However, it has been established that this modification would not take into account the total variance, which increases as the first-degree sampling fraction increases. Recently, in 2012, an

extension of the *Rao-Wu Rescaling Bootstrap* was proposed to full take into account an estimator's variance under a multi-stage plan (Osiewicz and Pérez-Duarte, 2012).

Progress:

A document was written to put in context the problem with the current *Rao-Wu Rescaling Bootstrap* method, which does not correctly estimate the variance in the two-stage sample design without a replacement where the first-stage sampling fraction is significant (Charlebois and Laperrière, 2017). This document justifies exploring the proposed extension by Osiewicz and Pérez-Duarte (2012).

Simulations have been conducted using a data set provided in Särndal, Swenson and Wretman (1992), extreme cases in which the *Rao-Wu Rescaling Bootstrap* was implemented twice and failed to accurately estimate the variance. With a very large sampling fraction in the first stage and a small sampling fraction in the second stage, both implementations were far from the true variance while the extension gives exactly the right value. Other simulations were conducted on an artificially-created data set and the same phenomenon was observed. These simulations also revealed that under certain extreme conditions, the extension can result in a negative bootstrap weight.

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

Sub-Project: Confidence Intervals for Proportions

The purpose of the project is to identify methods of constructing confidence intervals for proportions that perform reasonably well for complex survey data and that are easy to implement. The most commonly used method, the Wald interval does not perform well for small sample sizes and for proportions near zero or one. Many alternative methods have been proposed in the literature, of which a few have been adapted for complex survey data (e.g., Kott and Carr, 1997; Korn and Graubard, 1998). Limited empirical studies have been conducted to evaluate the performance of these adapted methods (e.g., Liu and Kott, 2009). The project involves conducting simulation studies to evaluate the performance of confidence interval methods, and providing guidelines and recommendations based on the simulation results.

Progress:

The results of a simulation study conducted in 2015-2016 were documented in Neusy and Mantel (2016). The simulation study evaluated the performance of the Wald interval, the modified Wilson interval, the modified Clopper-Pearson interval, the logit transformation interval and the bootstrap percentile interval under simple random sampling without replacement, stratified random sampling and two-stage sampling.

Guidelines for constructing confidence intervals and disseminating proportions were finalized (Neusy, 2017). The recommended methods for constructing confidence intervals for proportions are the modified Wilson interval, the modified Clopper-Pearson interval and the logit transformation interval (in order of preference).

Following a recommendation from our Advisory Committee on Statistical Methods, a simulation was run to compare the performance of the modified Wilson interval and

the empirical likelihood interval for proportions. We observed that the modified Wilson interval seems to perform better than the empirical likelihood interval for small sample sizes under simple random sampling without replacement.

Finally, another simulation was run to evaluate the performance of confidence interval methods under a two-stage design similar to that of several Household Survey Methods Division (HSMD) surveys, where households are selected using simple random sampling in the first stage, and one adult per household is randomly selected in the second stage. The recommended methods once again performed better than the Wald interval.

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

Sub-project: A test of sample matching using existing data

With increasing levels of nonresponse in household surveys, there is renewed interest in alternatives to the traditional way of conducting such surveys. Rivers (2007) proposed the sample matching approach, and showed that under certain assumptions, matching from a sufficiently large and diverse web panel provides results similar to a simple random sample. The goal of this study was to test the sample matching method proposed by Rivers at Statistics Canada.

Progress:

Sample matching methodology was simulated using data from two different household surveys. The population of the study consisted of the 2011 National Household Survey (NHS) respondents and the respondents of the Canadian Labour Force Survey (LFS) were treated as a pseudo-web sample. Demographic information from both surveys were used as auxiliary information in the matching process, and the percentage of the respondents employed during the reference week was considered as the variable of interest. To have the same reference period for both surveys, only the May 2011 LFS data were considered.

The simulation study was conducted under various conditions, and the performance of the method was evaluated by comparing the sample matching estimates with the NHS data. The simulation results showed that three factors impact the matching estimates: the size of the random sample drawn from the population, the matching variables and the sampling method. Moreover, the results suggested that the ratio between the sample size and the size of the web panel affects the absolute bias (AB) and the root mean square error (RMSE) of the estimates.

The overall results were satisfactory as the absolute bias and the root mean square error of the sample matching estimates were fairly small when the right auxiliary information was used for matching the data. However, it should be noted that the methodology used in this project had some limitations. We used survey data (LFS) to mimic the process, but the LFS data do not have the same characteristics as data from a panel. The LFS has a high response rate and it is not subject to self-selection bias. Also, the quality of the data used in this study is not comparable with a web survey, as the LFS data and the NHS data are edited and imputed. Hence, we may

get a different outcome using data from a web panel. The plan is therefore to investigate this possibility in the next fiscal year using data from a real web sample.

The results of this project were presented at the Inference from Non Probability Samples (INPS) conference in March 2017.

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

References

Korn, E.L., and Graubard, B.I. (1998). Confidence intervals for proportions with small expected number of positive counts estimated from survey data. *Survey Methodology*, 24, 193-201.

Kott, P.S., and Carr, D.A. (1997). Developing an Estimation Strategy for a Pesticide Data Program. *Journal of Official Statistics*, 13, 367-383.

Liu, Y.K., and Kott, P.S. (2009). Evaluating alternative one-sided coverage intervals for a proportion. *Journal of Official Statistics*, 25, 569-588.

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

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.

1.10 Divisional Research - Social Survey Methods Division (SSMD)

The research sub-projects submitted to this project must be in keeping with the Methodology Branch's mandate and particularly the Social Survey Methods Division's objectives.

Sub-Project: Small area estimation under an informative sample design

The research aims to study small area estimation with unit-level models under an informative sample design. The proposed estimators in this research will be based on augmented models of design variables. They will have the advantage of being simpler and easier to use than the corrected estimators for the informative design of Pfeffermann and Sverchkov (2007).

Progress:

Simulation studies have been conducted for various sampling plans. They have provided a better understanding of the impact of plan features on small and medium area estimation and have allowed us to find plan variables to add to the model that

gives good point estimators. However, the studies also revealed that the estimation of the mean square error (MSE) of these estimators needs work. Moreover, the results of the point estimate were presented at the 2016 SAE conference in the Netherlands (Rao, Verret and Chatrchi, 2016).

For further information, please contact:
François Verret (613-862-6638, francois.verret@canada.ca).

Sub-Project: Theoretical domain framework

The Theoretical Domains Framework (TDF) is a tool used in the health care field to understand and change complex behaviours. This project is attempting to apply the TDF in the context of household survey non-response at Statistics Canada, to better understand the barriers to survey response.

Progress:

Work has been ongoing on several avenues: understanding the relevance of each of the TDF domains in the survey context, understanding how existing materials provided to individuals selected for surveys may fit into these domains, and learning about the process of designing interventions to address barriers. A consultation with Dr. Brehaut and his colleague, who have expertise working with the TDF, took place in December. Some questions inspired by the TDF were integrated into focus group testing on the Survey of Household Spending (SHS). Results of the SHS test will be analyzed and documented.

For further information, please contact:
Maggie Wu (613-863-8564, margaret.wu@canada.ca).

Reference

Pfeffermann D., and Sverchkov M. (2007). Small-area estimation under informative probability sampling of areas and within the selected areas. *Journal of the American Statistical Association*, Vol. 102, No. 480, 1427-1439.

2 Support Activities

2.1 Record Linkage Resource Centre (RLRC)

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

Progress:

We continued to provide the development team of G-Link with support and participate in the Methodology-System Engineering Division (SED) Record Linkage Working Group meetings. RLRC team met with SED bi-weekly and tracked the minutes which could be deemed as potential source of current/past fix/bugs/improvements of G-Link. RLRC also provided internal and external G-Link users with support when help/comments/suggestions regarding G-Link were sought at G-Link_info through JIRA tickets.

Prototype of Mixmatch exclusion and conversion tables have been tested through G-Link user rule and delivered to SED for the integration with G-Link 3.4. To help improve the Fellegi-Sunter Classification and reduce the burden of clerical review, RLRC developed a SAS macro to generate automated weight thresholds using unsupervised machine learning technique called K-means clustering. Our record linkages with health data helped us document performance and issues pertaining to management and developers. We also worked on tobacco litigation linkages and used the linkages as an opportunity to field test new G-Link features and develop more systematic and theoretically coherent approaches of defining and adjusting record linkages under servers and SAS Grid.

RLRC prepared the G-LINK 3.3 tutorial for the release of G-LINK 3.3 and contributed to the user guide of G-Link 3.3, the methodology of MixMatch string comparators and EM (Expectation Maximisation) algorithm. The new version of G-LINK 3.3 contains several enhancements including implementation of EM algorithm under the assumption of conditional independence to calculate weights, Mixmatch string comparators.

RLRC evaluated the version 4.0 of Febrl (Freely Extensible Biomedical Record Linkage) to explore the implementation of unsupervised and supervised machine learning algorithms for classifying the pairs into definite and rejected clusters.

The list of record linkages carried out in the Methodology Branch was updated in 2016 and the results 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 sub-topics with emphasis on the following:

- Consultation in Time Series (including course development);
- Time Series Processing and Seasonal Adjustment;
- Support to G-Series (Benchmarking and Reconciliation);
- Modeling 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 reviewed period were related to the identification of break in series, application of seasonal adjustment in various situations (System of National Accounts, local level estimates, new Labour Survey Division (LSD) surveys, etc.) and specific contexts for benchmarking and reconciliation.

TSRAC members continued their participation in various analytical and dissemination groups such as the Forum for the Daily analysts and the Forum on seasonal adjustment and economic signals. TSRAC members also met with various international visitors to discuss time series issues and refereed papers for external journals.

Time Series Processing and Seasonal Adjustment

This project monitors high-level 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 expand on the available diagnostics produced by the system, to improve the reporting capabilities for clients. Other enhancements were made to

improve functionality and processing performance as needs arise within relevant projects.

An additional project was undertaken this year to explore different methods for estimating the variance of seasonally adjusted data. A literature review and comparative study was conducted and a presentation was made to the Advisory Committee on Statistical Methods. Further research was suggested and is currently underway in the section. Based on the feedback from the Advisory Committee on Statistical Methods (ACSM), the leading methodologies being considered are based on replication techniques, and linearization.

Support to 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. Various papers and presentations are were given for 2016-2017 to publicize the new functionality and the new module will also be incorporated into the training material offered for raking. The new functionality is being considered for use in a number of projects, including particular areas of the SNA, and the reconciliation for the Quarterly Financial Survey. New training materials will be included in the reconciliation course to explain the macro, and compare it to the PROC TSRAKING.

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 of G-Series.

Modeling and forecasting

The recently acquired software SAS Forecast Studio (SAS/HPF) continued to be explored and used for various projects related to time series modeling. It proved to be an efficient preliminary tool to evaluate breaks in series and to detect large time series outliers. Statistics Canada's recent experience was documented and shared in various presentations and applied in a number of projects to detect breaks in series in the absence of a formal parallel run. Further exploration of the capacities and theory underlying the software were explored, and several team meetings on the topic were held to discuss the underestimation of variance, as well as the similarity between the models that are available.

Additional analysis was explored, to estimate the effect of extreme weather on economic time series. While not inherently included in seasonal adjustment, 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 working group was built to communicate the analysis throughout the agency, led by the Time Series Research and Analysis Centre, and a paper will be

presented at the 2017 Joint Statistical Meetings to describe the approach and show examples using Statistics Canada data. In addition, an Analysis Project has been approved to prepare an analytic paper jointly with Retail and Service Industries Division based on applying this analysis to monthly retail trade data.

Another area that is currently being researched within the division is that of model-based seasonal adjustment methods. While SEATS, a decomposition based solely on ARIMA models, is growing in popularity it has not been widely adopted outside of Europe. Comparative analysis was conducted using data from Statistics Canada to identify circumstances where the results are better, worse, or similar to current results using this methodology. State space models are also a promising framework which could be applied to seasonal adjustment, and preliminary work was conducted to familiarize with these techniques, and explore how they could be used. These models have the capacity to take into account the variance of the unadjusted data, and ultimately provide variance estimates for the resulting seasonally adjusted estimator.

Trend-Cycle Estimation

Trend-cycle lines were added in *The Daily*, Statistics Canada's official release bulletin for many key economic indicators. This was the results of a multi-year project where various methods were reviewed, first in the literature and then in a simulation study. The results - which favor the use of a trend-cycle line in our published graphs using a variant of the Dagum and Luati (2009) – were adapted in production. In the 2016-2017 period, the communication plan was expanded to increase the transparency of the method used, and respond to inquiries from external clients. A post-mortem analysis is planned for the 2017-18 period, including the collection of feedback from internal and external users, and to explore other areas within the organization that could consider the release of trend-cycle estimates.

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

Reference

Dagum, E.B., and Luati, A. (2009). A cascade linear filter to reduce revisions and false turning points for real time trend-cycle estimation. *Econometric Reviews*, 28:1-3, 40-59.

2.3 Research Data Centres and Confidentiality support

The Research Data Centres provide researchers with access, in a secure university setting, to microdata from population and household surveys. 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 Research Data Centers (RDC) analysts and researchers on

vetting requests. Methodologists also develop survey specific guidelines whenever a new survey becomes available in the RDCs.

Progress:

A Metholodoogy Expert Panel (MEP) on the creation of Public Use Microdata File (PUMF) was put in place. It consists of 4 methodologists headed by Peter Wright. The team has the mandate to review, guide and recommend the approval of PUMF to the Microdata Release Committee (MRC). The MEP reviewed about half dozen PUMF this year. They also refined their process and prepared a seminar.

General support is provided to clients and other methodologists on disclosure control. Development is also sought on specific projects related to disclosure control such as The Companion and Confid-on-the-fly.

We developed and tested the initial SAS programs required to implement the proposed score function for the Companion. For Confid-on-the-fly, an issue detected in the course of the evaluation of the multinomial regression is still unsolved. The Australian Bureau of Statistics continues to provide help to resolve this situation. Methodology research work on understanding how to create the perturbation tables required for creating confidentialised outputs from the Confid-on-the-fly was completed. A document was written detailing the necessary steps required for categorical and continuous data.

Material for an advanced confidentiality courses was completed.

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 sub-topics with emphasis on the following:

- Redesign of the Quality Assurance Reviews;
- Process Quality Management;
- Corporate Performance Indicators.

Progress:

Redesign of the Quality Assurance Reviews

The Quality Secretariat has been conducting quality assurance reviews at Statistics Canada since 2006. The scope of the reviews under the original model was quite broad, and a wide range of quality improvement opportunities was discovered. However the workload for Reviewers was quite high compared to the perceived impact or value of the outputs. In 2015 senior management agreed to take a pause from

doing quality reviews, and the Quality Secretariat was tasked with researching and proposing a new model that would be effective, efficient, relevant, feasible, affordable, and sustainable through time.

We researched how quality assurance is assessed at other national statistical offices, and were most impressed by ASPIRE (A System for Produce Improvement, Review and Evaluation) developed in 2012 by Biemer and Trewin and implemented at Statistics Sweden. An ASPIRE inspired model was developed for use at Statistics Canada. The model assesses quality risk mitigation for the 6 dimensions of quality (accuracy, relevance, timeliness, coherence, interpretability and accessibility) and uses a questionnaire to be completed by the manager of the statistical process as well as objective assessment by reviewers external to the program being reviewed but internal to Statistics Canada. The draft model was reviewed by former Reviewers and managers of programs that had been reviewed under the former review process, and the revised model was presented to the Methods and Standards Committee. We met with the Director of Program Evaluation and ensured that there is no overlap between Program Evaluation and Quality Reviews; in fact, the scope of the two assessments is complementary. The material is ready to be pilot tested once a suitable program is chosen. The Quality Review questionnaire will be available for self-assessment by anyone on the Internal Communications Network (ICN).

Process Quality Management

Managing the quality of statistical products has been extensively researched, explored, documented and discussed. Now attention is shifting to how to manage the quality of a statistical process.

An international working group known as the “Leadership Expert Group on Quality” has proposed dimensions of process quality that look at the integrity of the process rather than characteristics of the products it produces. Researchers at IStat (the national statistical office of Italy) have refined those quality dimensions for application to infrastructure processes in the context of producing official statistics. The Quality Secretariat has brought these ideas into the context of the objectives and principles of Statistics Canada’s Corporate Business Architecture, and is proposing 6 dimensions of process quality for Statistics Canada: efficient; reproducible, reliable, robust and secure; flexible; effective; identified, transparent and corporately supported; integrated.

These concepts were shared with the Methods and Standards Committee, however so far no further work has been done. The Quality Secretariat will continue to monitor development in other national statistical offices and in multilateral organizations. More feedback on the proposed process quality dimensions will be sought from the Methodology, Information Technology, and Subject Matter expert communities.

Corporate Performance Indicators

A suite of approximately 40 corporate performance indicators is managed by Finance Division. The Quality Secretariat is the “indicator owner” for 7 indicators related to quality: timeliness, punctuality, accuracy (coefficients of variation), accuracy (response rates), and 3 for relevance.

In June 2016 the first annual set of indicators for 2015-2016 was delivered to Financial Reporting Division. Some domain definitions and indicator methodology were revised for the 2016-2017 indicators. The processes for gathering inputs and processing the indicators was streamlined and refined. We are on track to deliver the 2016-2017 indicators on time (mid-May). We were expecting a request to provide sub-annual indicators or even to feed a dashboard where managers could see their indicators at any point throughout the year, but this request has not yet been made. Once these requests are clearly specified, the Quality Secretariat will make the necessary revisions to our system.

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

2.5 Data Analysis Resource Centre

The Data Analysis Resource Centre (DARC) is a team of statistical consultants and researchers within the Methodology Branch. The main goals of DARC are to give advice on the appropriate use of data analysis tools and methods, and to promote best practices in this area. DARC's services – which focus mainly on survey, census or administrative data – are available to the employees of the Agency or other departments, as well as analysts and researchers from academia or Research Data Centres (RDCs).

Progress:

Consultations

As part of the Data Analysis Resource Centre (DARC) mandate, consultation was offered as requested by various clients. Specific consultation services were provided to Statistics Canada's analysts from a dozen of different divisions. These various consultations covered topics on the use of weights and normalized weights in different studies, bootstrap variance, testing of hypotheses, statistical significance with large data sets, dealing with missing data in analysis, regression models, coding categorical variables, variance estimation for age-standardized rates, helping with SUDAAN and Stata code, etc.

The group also provided services to other methodologists. These consultations included questions on confidence intervals for proportions, testing of hypotheses, multiple comparisons, using normalized weights, logistic regression, etc.

External consultations were also delivered to a variety of clients, ranging from other government ministry to academia and including specialised private analytical firms.

Finally, expert advices were exchanged with analysts and researchers from the Research Data Centres (RDC). The topics included bootstrap variance estimation, significance testing, combining cycles of the Canadian Community Health Survey (CCHS), mediation models, trajectory analysis, latent variables, multiple imputation, variance estimation for medians, using SAS SURVEY procedures, Stata commands, using weights and normalized weights in different types of analysis, multi-level models, etc.

Written material to support analysis at RDCs

The document « Introduction aux effets médiateurs et modérateurs » which is intended to be a tool for analysts who are considering using these models in their analysis was completed (Michaud and Mach, 2017). The document has been translated and will be distributed soon.

Provision of Training

A full-day workshop “Bridging the gap: Turning classical statistics experience directly into a working knowledge of survey data analysis” was developed and presented at the 2016 Annual Meeting of the Statistical Society of Canada in St. Catharines, Ontario. The workshop was recently given to employees of the Canadian Revenue Agency (CRA) in Ottawa and, because of high demand, is scheduled to be offered to other CRA employees this winter.

The team redesigned the DARC lecture for the Data Interpretation Workshop (DIW). The main objective of the redesign was to make the course material more accessible and engaging for the DIW participants and this initiative was very welcome by the DIW management. The new DARC lecture is presented in two half-day sessions, the first session is entitled “Key Aspects of Survey Data Analysis”, and the second covers four staples of statistical analysis (confidence intervals, p-values, multiple comparisons and statistical tests). We presented the redesigned DARC lecture at the English DIW in September. It was well received by the DIW management as well as the participants.

Collaboration with analysts

One of the collaboration with Children's Hospital of Eastern Ontario (CHEO) involving the analysis of three cycles of Canadian Health Measures Survey (CHMS) were summarized in scientific papers. One deals with sleep in children and adolescents (Michaud and Chaput, 2016); the other with sleep in adults and older adults (Chaput, Wong and Michaud, 2017).

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

2.6 Knowledge Transfer - Statistical Training

The statistical training program offered internally currently includes thirty courses that cover various topics related to survey methodology, data analysis methods and time series. Development and coordination of the program courses are the responsibility of the Statistical Training Committee whose mandate is as follows:

1. Assist the Methodology Branch’s divisions with determining the statistical training needs.
2. Coordinate the development of new statistics courses.
3. Find instructors who can give statistics courses.

4. Coordinate logistical support for courses (including preparation of the schedule and registration process).
5. Collect feedback from participants and follow up as needed.

Progress:

In the 2016-2017 fiscal year, 42 courses were offered for a total of 106 days of training. A total of 386 participants (359 from Statistics Canada and 27 from outside) attended those courses. In all, that makes 1129 participant days, which is an increase of 8% from 2015-2016. This is the second highest total (in terms of participant days) since the Statistical Training Program started.

During that period, the redesign of the 0438B course (Statistical Analyses with Survey Data – Module 2) continued. The new H-0435A course "Introduction to Confidentiality and Statistical Disclosure Control with Emphasis on Household and Social Surveys" was offered for the first time in 2016-2017.

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

2.7 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 2016 issues (42-1 and 42-2) were released in both PDF and HTML version. The volume 42 includes 14 papers and 3 short notes in total.

A large number of historical papers representing about 15 years were also released online in May 2016. Electronic copies of any paper published since December 1981 (Volume 7-2) are now available. Older papers can still be obtained upon requests. A subset of those selected based on their pertinence is being prepared for electronic dissemination.

From April 2016 to March 2017, *Survey Methodology* pages were viewed 47,000 times and 51,000 copies of papers were downloaded. 48 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) is in preparation.

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

Beaumont, J.-F., and Bocci, C. (2016). Small Area Estimation in the Labour Force Survey. Paper presented at the Advisory Committee on Statistical Methods, May 2016, Statistics Canada.

Bocci, C., and Beaumont, J.-F. (2017). SAE methodology applied to the Monthly Survey of Manufacturing. Internal document, Methodology Branch, Statistics Canada.

Brisebois, F. (2016). *Segmentation of the Canadian Population to Derive Efficient Communication and Operation Strategies for Household Surveys,* Internal presentation.

Buskirk, T., and Kreuter, F. (2016). A small course on big data for survey researchers, workshop given at the 2016 Symposium.

Chaput, J.P., Wong, S.L. and Michaud, I. (2017). Duration and quality of sleep among Canadians aged 18 to 79. *Health Reports*, Vol. 28, no. 9, September 2017.

Charlebois, J., and Laperrière, C. (2017). An Extension of the Rao-Wu Rescaling Bootstrap for two-stage sample designs, Internal Document – Draft.

Chen, S., and Jamrov, E. (2016). Bayesian Approach with a discreet prior, Draft working paper.

Chu, K., Saidi, A. and Dasylva, A. (2017). A simulation study of the method for linkage-error-adjusted logistic regression based on clerical review of Chipperfield et al., Internal report.

Dasylva, A., Abeysundera, M., Akpoué, B., Haddou, M. and Saidi, A. (2016). Measuring the quality of a probabilistic linkage through clerical-reviews, presentation given at the 2016 Symposium.

Dasylva, A., and Labrecque-Synnott, F. (2017). Bayesian solutions for record-linkage problems, Internal presentation.

Estevao, V., and Rubin-Bleuer, S. (2016). SAS simulation macro for small area estimation. International Cooperation and Corporate Statistical Methods Division (ICCSMD), Internal report, Statistics Canada.

Finch, D. (2017). *Banff User Review Report*. Internal Report.

Fournier, L. (2016). Prototype for estimating linkage errors, Internal presentation.

Fung, H. (2017). *Example of a graphical analysis using New Graduates Survey data*. Internal Report.

Gray, D., Stinner, M., Wright, P. and Thomas, S. (2016). *Current Developments in Disclosure Control for Business Surveys at Statistics Canada*. Technical Report Presented at Statistics Canada's Advisory Committee on Statistical Methods, October 24-25, 2016.

Halladay, A., Brisebois, F. and Yang, M.J. (2017). *Segmenting the Canadian Population to Remedy Non-Response in Statistics Canada Household Surveys*, Statistical Society of Canada annual meeting.

Haziza, D., and Beaumont, J.-F. (2017). Construction of weights in surveys: A review. *Statistical Science* (accepted).

Holness, P., Sobrino, R., Dasylva, A., Trudeau, R. and Pelletier, C. (2017). Generic preprocessing model for record-linkage, Internal report.

Kleim, G. (2016). An alternative approach to the calculation of a prioritization score in ICOS, Internal document.

Michaud, I., and Chaput, J.P. (2016). Are Canadian children and adolescents sleep deprived? *Public Health*, *141*, 126-129. http://dx.doi.org/10.1016/j.puhe.2016.09.009.

Michaud, I., and Mach, L. (2017). Introduction aux effets médiateurs et modérateurs. Training material, soon to be available in English.

Michaud, I., Colley, R. and Garriguet, D. (2016). Preliminary Report: Evaluation of different models of activity monitors for the objective measurement of sedentary behaviours, physical activity and sleep in the CHMS. Internal document, October 2016.

Michaud, I., Colley, R. and Garriguet, D. (2017). Second Report: Objective measurement of sedentary behaviours, physical activity and sleep in the CHMS. Internal document, February 2017.

Michaud, I., Henderson, M., Legault, L. and Mathieu, M.E. (2016). Physical activity and sedentary behaviour levels in children and adolescents with type 1 diabetes using insulin pump or injection therapy–The importance of parental activity profile. *Journal of Diabetes and its Complications*. http://dx.doi.org/10.1016/j.jdiacomp.2016.11.016.

Miville, H., and Ubartas, C. (2016). Test sur l'algorithme de priorisation des appels – application de l'approche de Bollapragada et Nair, Internal document.

Mizdrak, P. (2017a). Summary of blocking methods, Internal report, Statistics Canada.

Mizdrak, P. (2017b). Graph databases in record-linkage, Internal report, Statistics Canada.

Neusy, E. (2017). Disseminating the Quality of Proportions. Internal Document.

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

Nolet-Pigeon, I. (2017a). *Résultats préliminaires des études sur les enjeux pratiques liés au calage pour le Programme intégré de la statistique des entreprises*. Internal document.

Nolet-Pigeon, I. (2017b). *Résultats préliminaires des études sur les enjeux pratiques liés au calage pour le Programme intégré de la statistique des entreprises, Part 2*, Internal document.

Oyarzun, J. (2016). Business record-linkage, presentation given at the 2016 Record linkage workshop.

Oyarzun, J., and Wile, L. (2016). An Overview of Business Record Linkage at Statistics Canada: How to link the unlinkable, presentation given at the 2016 Symposium.

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

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

Rao, J.N.K., Verret, F. and Chatrchi, G. (2016). Small Area Estimation under Informative Sampling. Paper presented at SAE2016 Conference, Maastricht, Netherlands, Aug. 17-19, 2016.

Romanyuk, Y., and Boulet, C. (2017). Effect of Contact Strategy on Response Mode Selection. Statistics Canada, Internal document, draft.

Rubin-Bleuer, S., Jang, L. and Godbout, S. (2016). The pseudo-EBLUP estimator for a weighted average with an application to the Canadian Survey of Employment, Payrolls and Hours. *Journal of Survey Statistics and Methodology*. http://jssam.oxfordjournals.org/cgi/reprint/smw013?ijkey=txjkdjo6EbImaxD&keytype=ref.

Salier, K., and Kleim, G. (2017). Création d'un score de priorisation dans les enquêtes ITAO - Étude de la modélisation des règles hiérarchiques bayésiennes, Internal Presentation.

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.

Wright, P. (2016). *Specification to adjust the weighted cost of cells as a function of the outstatus of the previous release of the results of the same cycle*. Internal Report.

You, Y. (2016). Small area estimation when sampling variances are unknown and estimated with application in LFS. Research paper submitted to a journal for possible publication.

You, Y. (2017a). A book review: Analysis of Poverty Data by Small Area Estimation (Eds. M. Pratesi), New York: John Wiley & Sons, Inc. 2016. *The Survey Statistician*, January 2017.

You, Y. (2017b). Report on Hierarchical Bayes small area estimation using log-linear area level models with LFS application. International Cooperation and Corporate Statistical Methods Division (ICCSMD), Internal report.

Zhang, Q., and You, Y. (2016). Small area estimation using EBLUP, pseudo-EBLUP and M-quantile approaches: Review and simulation study. International Cooperation and Corporate Statistical Methods Division (ICCSMD), Internal report.