

Statistical Methodology Research and Development Program Achievements

Statistical Methodology Research and Development Program 2010-2011 Achievements



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 <i>Statistics Act</i> |
| ^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

© Minister of Industry, 2011

All rights reserved. Use of this publication is governed by the Statistics Canada [Open Licence Agreement](#).

An HTML version is also available.

Cette publication est aussi disponible en français.

Statistics Canada

Statistical Methodology Research and Development Program 2010-2011 Achievements

Archived Content

Information identified as archived is provided for reference, research or recordkeeping purposes. It is not subject to the Government of Canada Web Standards and has not been altered or updated since it was archived. Please "**contact us**" to request a format other than those available.

This report summarises the 2010–2011 achievements of the Methodology Research and Development Program 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 which would not otherwise be carried out during the provision of methodology services to those survey programs. This program also includes activities that provide client support in the application of past successful developments in order to promote the use of the results of research and development work. Contact names are provided for obtaining more information on any of the projects described. For more information on the program as a whole, contact:

Mike Hidioglou
(613-951-0251, mike.hidioglou@statcan.gc.ca).

Statistics Canada

Research Projects

Research, development and consultation in SRID

The Statistical Research and Innovation Division (SRID) was created within the Methodology Branch on June 21, 2006. SRID is responsible for researching, developing, promoting, monitoring and guiding the adoption of new and innovative techniques in statistical methodology in support of Statistics Canada's statistical programs. Its mandate also includes the provision of technical leadership, advice and guidance to employees elsewhere in the Program. This assistance is in the form of advice for methodological problems that arise in existing projects or during the development of new projects.

SRID is also jointly involved with other employees of the Program via research projects sponsored by the Methodology Research and Development Program on specific topics, e.g., estimation methods, imputation methods, small area estimation, use of administrative data and data collection. Two SRID members won prestigious honours in 2010: Jean-Francois Beaumont was named the best employee of the year at Statistics Canada for 2010 and Benoît Quenneville won the 2010 Statistics Canada Tom-Symons Award.

The Statistical Research and Innovation Division (SRID) was involved in many research, development and consultation projects in 2010-2011. In particular, SRID staff has made significant contributions in small area estimation, imputation and robust estimation, synthetic data generation and time series techniques. Details of the progress made are incorporated into the review of the research topics reported later.

In addition to participating in the research activities of the Methodology Research and Development Program (MRDP) as research project managers and active researchers, SRID staff was involved in the following activities:

- Staff advised members of the other divisions on technical matters on both an ad hoc and a formal basis. Examples of ad hoc advice include variance estimation for complex surveys, sample coordination, small area estimation and variance estimation for imputed data.
- Research project managers within SRID assisted team members of projects sponsored by the Methodology block fund by providing ideas on how to advance their research and regularly reviewing their work.
- SRID consulted with members of the Methods and Standards Committee, as well as with a number of other Statistics Canada managers, in determining the priorities for the research program.

- SRID continued to actively support the Survey Methodology Journal. Mike Hidiroglou has been the editor of the Survey Methodology Journal since January 2010, and three members of SRID are assistant editors.
- Staff continued activities in different Branch committees such as the Branch Learning and Development Committee, the Strategic Thinking Group and the Methodology Branch Informatics Committee. In particular, we have participated actively in finding and discussing papers of the month.
- SRID participated actively in the committee on the strategic thinking process and co-authored a paper presented at the Advisory Committee on Statistical Methods (Trépanier et Beaumont, 2010).
- SRID participated at the interchange with the U.S. Census Bureau in Ottawa in March 2011 (Jean-Francois Beaumont and Mike Hidiroglou).
- SRID made a presentation and wrote a paper on imputation in the context of the Economic Statistics China Project (Beaumont, 2011).
- Staff refereed several papers in statistical journals.
- Staff authored or co-authored 29 papers, many of which were presented at conferences (Statistical Society of Canada, Joint Statistical Meetings, 7th International Institute of Forecasters), or published in learned statistical journals such as *the Canadian Journal of Statistics*, *Survey Methodology*, *Handbook of Statistics*, *Statistica Sinica*, *Journal of Multivariate Analysis*, and the *Journal of the Royal Statistical Society*, *Eurostat*, *Journal of Indian Society of Agricultural Statistics*, *Workshop*, *Estudios de Economia Aplicada*, *Advances and Applications in Statistics*.

For further information, contact:

Mike Hidiroglou (613-951-0251, mike.hidiroglou@statcan.gc.ca).

Papers presented at conferences or published

1. Beaumont, J.-F. (2011). How to evaluate the impact of imputation on data quality? Paper written in the *SIMP II – Economic Statistics Document Series*, Document #48.
2. Beaumont, J.-F., and Bissonnette, J. (2011). Variance estimation under composite imputation: The methodology behind SEVANI. *Survey Methodology* (to appear).
3. Beaumont, J.-F., and Charest, A.-S. (2011). Bootstrap variance estimation with survey data when estimating model parameters. Paper under revision in *Computational Statistics and Data Analysis*.
4. Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2011). A unified approach to robust estimation in finite population sampling. Paper submitted to a refereed journal.
5. Beaumont, J.-F., and Patak, Z. (2011). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. Paper under revision in the *International Statistical Review*.
6. Beaumont, J.-F., and Patak, Z. (2010). Generalized Bootstrap in Prices Surveys. Paper presented at the Advisory Committee on Statistical Methods, April 2010.
7. Beaumont, J.-F., Haziza, D. and Bocci, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica* (to appear).
8. Bemrose, R., Meszaros, P. and Quenneville, B. (2011). Trend-cycle approach to estimate changes in Southern Canada's water yield. *Estudios de Economia Aplicada*, 28-3, 595-606.
9. Bemrose, R., Meszaros, P., Quenneville, B., Henry, M., Kemp, L. and Souldard, F. (2010). Using a trend-cycle approach to estimate changes in Southern Canada's water yield from 1971 to 2004. Environments Accounts and Statistics Division, Catalogue no. 16-001-M, no. 14. Won Statistics Canada Tom-Symons Award.
10. Bianconcini, S., and Quenneville, B. (2010). Real time analysis based on reproducing kernel Henderson filters. *Estudios de Economia Aplicada*, 28-3, 533-574.
11. Chen, Z.G., and Wu, K.H. (2010). Benchmark Forecast and Error Modelling. Presented at the 4th Annual International Conference on Mathematics & Statistics, Athens, Greece, June 2010.
12. Chen, Z.G., and Wu, K.H. (2011). Benchmarking using working error-model. *Advances and Applications in Statistics*, 19-2, 113-148.
13. Choudhry, G.H., Rao, J.N.K. and Hidiroglou, M.A. (2010). On sample size allocation for planned small domains. Submitted to *Survey Methodology*.
14. Choudhry, G.H., Hidiroglou, M.A. and Laflamme, F. (2011). Optimizing CATI call scheduling to minimize data collection costs. To be presented at *Joint Statistical Meetings* 2011.
15. Fortier, S., Quenneville, B. and Picard, F. (2010). Calendarisation methods and applications. Presented at ACSM October 2010.
16. Hidiroglou, M.A. (2010). Current Developments in Small Area Estimation at Statistics Canada. Invited Paper Presented at *Joint Statistical Meetings* 2010, Meetings held In Vancouver.
17. Jamroz, E. (2010). Report on Small Area Estimation Using a Linear Mixed Effects Model. Internal SRID report.

18. Mazzi, G.L., Fortier, S. and Quenneville, B. (2010). Multivariate benchmarking - The direct versus indirect problem. Paper written for a chapter in a Handbook on seasonal adjustment to be published by *Eurostat*.
19. Quenneville, B., and Fortier, S. (2010). Benchmarking and temporal consistency. Paper written for a chapter in a Handbook on seasonal adjustment to be published by *Eurostat*.
20. Quenneville, B., and Fortier, S. (2011). Restoring accounting constraints in time series – Methods and software for a statistical agency. Sent for possible publication in an edited volume in honour of David Findley to be published by Chapman & Hall/CRC.
21. Quenneville, B., and Gagné, C. (2011). Testing time series data compatibility for benchmarking. Presented at ACSM April 2010; presented at the 7th International Institute of Forecasters' Workshop, Swiss, January 2011. Sent for possible publication in a special issue of the *International Journal of Forecasting*.
22. Quenneville, B., Picard, F. and Fortier, S. (2010). Interpolation, benchmarking and temporal distribution with natural splines. Invited presentation *Joint Statistical Meetings 2010*, Vancouver, Canada, August 2010. Sent for possible publication to the *Journal of the Royal Statistical Society Series C*.
23. Rao, J.N.K, Verret, F. and Hidiroglou, M.A. (2010). A weighted estimating equations approach to inference for two-level models from survey data. Statistics Society of Canada (SSC) Annual Meeting, *Proceedings of the Survey Methods Section*, May 2010.
24. Rao, J.N.K., Hidiroglou, M.A., Kovačević, M. and Yung, W. (2010). Role of weights in descriptive and analytical inferences from survey data: An overview. *Journal of Indian Society of Agricultural Statistics*, 64(2), 129-135.
25. Rubin-Bleuer, S. (2011). The proportional hazards model for survey data from independent and clustered super-populations. *Journal of Multivariate Analysis*, 102, 884-895.
26. Rubin-Bleuer, S. (2010). The proportional hazards model for survey data from independent and clustered super-populations (Additional material). SRID-2010-004-E.
27. Rubin-Bleuer, S., Yung, W. and Landry, S. (2010). Adjusted maximum likelihood method for a small area model accounting for time and area effects. SRID-2010-005-E.
28. Trépanier, J., and Beaumont, J.-F. (2010). Strategic thinking process of the methodology branch. Paper presented at the Advisory Committee on Statistical Methods, October 2010.
29. Verret, F., Hidiroglou, M.A. and Rao, J.N.K. (2010). Small area estimation under Informative sampling. SSC Annual Meeting, *Proceedings of the Survey Methods Section*, May 2010.
30. Xu, C., Chen, J. and Mantel, H. (2010). Smoothly Clipped Absolute Deviation in Analysis of Survey Data. SSC Annual Meeting, *Proceedings of the Survey Methods Section*.
31. You, Y., and Zhou, Q.M. (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology*, to appear.
32. You, Y. (2010). Small area estimation under the Fay-Herriot model using different model variance estimation methods and different input sampling variances. Methodology branch working paper, SRID-2010-003E, Statistics Canada, Ottawa, Canada.
33. You, Y. (2011). Hierarchical Bayes sampling variance modeling for small area estimation based on area level models. Internal Manuscript.
34. You, Y., Rao, J.N.K. and Hidiroglou, M.A. (2011). Benchmarking small area estimators under the Fay-Herriot model. Draft paper.
35. Yung, W., Rubin-Bleuer, S. and Landry, S. (2010). Small area estimation for business surveys. *Proceedings of the Survey Research Methods Section*, American Statistical Association, *Joint Statistical Meetings*.

MITACS

Again this year, Statistics Canada was involved in the National Program on Complex Data Structures by participating in the MITACS internship program. This program pairs academic statistics researchers and their doctoral students with mentors at Statistics Canada so that the students can conduct research on unsolved statistical problems that have been identified in the area of complex survey data. As has been the case for the past few years, Statistics Canada agreed to take up to three students, each for a period of approximately four months. In the spring of 2010, advertisements were circulated nationally for these placements and two suitable candidates were chosen, both of whom arrived at Statistics Canada in September 2010. One of the students, Zeinab Mashreghi from the University of Montreal, was researching bootstrap variance estimation in the presence of imputed data. The other student, Wei Lin from the University of Toronto, was researching the embedding of experiments within surveys. Some of her ideas have been incorporated into an embedded experiment comparing follow-up strategies for internet survey data collection. Both students made presentations and prepared written reports of their research progress (see Lin, 2010 and Mashreghi, 2010).

The three MITACS students from fiscal year 2009/2010 - Haocheng Li, Chen Xu, and Valery Dongo Joingo – all presented papers on their work at the 2010 SSC meeting. Xu also submitted a paper to the Proceedings of the Survey Methods Section (See Xu, Chen and Mantel, 2010).

For further information, contact:

Georgia Roberts (613 951-1471, georgia.roberts@statcan.gc.ca).

References

Dongmo Jiongo, V., Haziza, D. and Duchesne, P. (2010). Robust Small Area Estimation. Presentation in a Contributed Paper Session at the 2010 SSC Annual Meeting.

Li, H., and Yi, G. (2010). A Composite Likelihood Approach for Longitudinal Data with Nonignorable Non-monotone Missing Observations in Both Response and Covariates. Presentation in a Contributed Paper Session at the 2010 SSC Annual Meeting.

Lin, W. (2010). Optimum Sample Allocation to Embedded Experiments within a Survey. Report submitted at end of MITACS internship.

Mashreghi, Z. (2010). Bootstrap Variance Estimation in the Presence of Imputed Data. Report submitted at end of MITACS internship.

Xu, C. (2010). Penalized Likelihood Methods for Variable Selection in Analysis of Survey Data. Presentation in a Contributed Paper Session at the 2010 SSC Annual Meeting.

Development of computer tools

A substantial amount of research has been done on several topics over the last few years. However, very few of the most innovative ideas get implemented in statistical programs. There are a number of reasons to explain this fact, including: i) methodologists responsible for statistical methods programs often do not have time to read and understand the newest findings; ii) no available statistical software package implements these new ideas; and iii) there is no time to develop appropriate software in a production context.

The main goal of this project is thus to support the development and maintenance of documented computer tools that implement research ideas with the potential to benefit statistical programs. This support will help ensure that state-of-the-art statistical theory and survey methods are developed and used by Statistics Canada's programs. This is, in fact, the mandate of the Methodology Branch.

Note that the goal of this project is not to develop generalized systems. Once a computer tool has matured and come into use by many statistical programs, it should be considered to be integrated in the family of generalized systems.

SEVANI – Research and Development on Variance Estimation in the Presence of Imputation

SEVANI is the System for the Estimation of Variance due to Non-response and Imputation. In the current year, we have made progress both in the development of the System and in the associated research. We have accomplished the following development tasks:

- i. improved the robustness (version 2.3.1) and computer efficiency (version 2.3.2) of the System; and
- ii. almost completed the development of a prototype that can be used to estimate the variance when a ratio is estimated. The current version of SEVANI (version 2.3.2) only handles totals and means. This prototype is being developed for use by the Travel Survey of Residents of Canada. It assumes that, when missing, the variables in the numerator and denominator have both been imputed simultaneously using donor imputation with a common donor for both variables.

Note that the above development work has been undertaken following users' requests. We have also continued to offer support to our users on the methodology and use of the System, and a seminar was given for the Business Special Surveys and Technology Statistics Division.

On the research side, we have accomplished the following:

- i. revised a paper on SEVANI's methodology for composite imputation (Beaumont and Bissonnette, 2011), which has been accepted in *Survey Methodology*;
- ii. studied different nonparametric options for the estimation of model variances to improve computer efficiency over PROC TPSPLINE, which is currently implemented in SEVANI. The new options are planned for implementation in version 2.4; and
- iii. researched a bootstrap procedure to estimate the variance for T1 data (Oyarzun and Nambeu, 2011) when nearest-neighbour imputation is used to fill in missing values. The goal of this research is to find a faster nonparametric alternative to the use of PROC TPSPLINE in SEVANI in the context of census data. The intuition behind this work is that a standard iid (independent and identically distributed) bootstrap should be valid for censuses. So far, the results do not support this intuition but research continues to determine an appropriate bootstrap methodology for this scenario.

Outlier detection tool

Several outlier detection methods currently in use at Statistics Canada have been combined into one tool developed in SAS. The tool helps methodologists select the most appropriate outlier detection method for their surveys and/or optimize the parameters of a specific method. It also makes it possible to visualize the data easily and compare detection methods.

In the past year, the top contributors method was added to the seven others already included. Development work was done to convert traditionally univariate or bivariate methods into multivariate methods. A presentation was given as part of a Business Survey Methods Division seminar to inform potential users of the tool's existence and make it available on request. A number of methodologists expressed interest and used the tool to analyze their survey data. A summary document that explains the procedure for using the tool and the theory underlying its methods is now available. In response to feedback from users, improvements were made and the methods were adjusted to stabilize the system.

BOOTVAR

BOOTVAR is a SAS tool that estimates variances using bootstrap weights. We developed version 3.2. This version allows the user to specify the number of degrees of freedom (the number of primary sampling units minus the number of strata) for confidence interval calculation and hypothesis testing. In previous versions, an infinite number of degrees of freedom and the normal (or chi-square) approximation were used. The documentation was updated accordingly. There is also an SPSS version of BOOTVAR. No changes were made for that version.

For further information, contact:

Jean-François Beaumont (613-951-1479, jean-francois.beaumont@statcan.gc.ca).

Sampling and estimation

This topic includes the following research projects:

- Controlled random sampling
- Indirect sampling from skewed populations
- Generalization of the Taylor linearization method for total variance estimation
- Sample coordination
- Mean bootstrap properties
- Calibration for variance estimation
- Calibration adjustment to globally minimize the bias due to various non-response phases
- Generalized bootstrap
- Robust estimation

Controlled random sampling

Hedayat and Robieson (1998) and later Chang, Wang and Huang (2004) showed how to build a sample design that is equivalent to simple random sampling without replacement (SRSWOR) but that excludes some possible samples to prevent the selection of certain undesirable samples. Their construction is based on a partition of the

space of size n samples so that a family of sample designs can be constructed in such a way that first- and second-order inclusion probabilities can be represented using a system of linear equations. Solving that system generates a design that is (weakly) equivalent to SRSWOR. We plan to extend those results by considering an operational research problem in which an objective function has to be minimized and the constraints would be of the same type but might be expressed in terms of inequalities.

We were able to simplify the proof of Hedayat and Robieson's initial result and show that non-trivial objective functions must necessarily be non-linear functions of inclusion indicators. The general form of the optimization problem has changed considerably since the project's inception. In fact, we substantially extended the scope of the initial results covering the SRSWOR case, and they now apply to instances involving any first- and second-order inclusion probabilities. The form of the general optimization problem is

Find a distribution $\beta_0, \dots, \beta_n$ that maximizes $E[B \mid B \leq E(B)]$
 Under the constraints and $E(B) = N_1\pi_1$

$$\sum k(n-k)\beta_k \leq N_1N_2\pi_1\pi_n \left[1 - \alpha \frac{N-n}{n(N-1)} \right]$$

where B is the random variable with distribution $\beta_0, \dots, \beta_n$, N is the size of the total population, N_1 is the number of undesirable units in the population, $N_2 = N - N_1$, π_1 and π_n are the inclusion probabilities of undesirable units and desirable units respectively, α is a parameter that takes values in the interval $[0, 1]$, and n is the sample size.

Using numerical routines (SAS proc optmodel), we obtained designs that considerably reduce the number of possible undesirable units in a sample. These methods can be used to generate designs for samples of any size, not just small sample sizes. We also developed a simple way of introducing these new designs. These projects will be the subject of a presentation at the SSC Annual Meeting in June 2011 (see León, 2011).

In the coming year, we will carry out simulations to put these methods through their paces and see how they perform with survey data. Various approaches are possible depending on the available auxiliary information and the objectives.

Indirect sampling from skewed populations

Enterprise-level estimates can be produced from an establishment-level survey using the generalized weight-share method. However, establishments are known to be a skewed population. Applying the conventional weight-share method to such a population could generate very high variances. The project's goal is to propose an alternative weight-share method so as to reduce the variance. We propose nine different approaches, all of them largely based on an economic weighting of relations between the survey frame (the population of establishments) and the target population (the population of enterprises that are made up of establishments). We will compare the nine methods using an actual population from the Business Register (BR).

The population was extracted from the BR, and a sample design was developed. The algorithm of the generalized weight-share method was programmed for each of the proposed methods:

1. Conventional approach
2. Weights proportional to selection probabilities
3. Weights proportional to size (number of employees)
4. Weights proportional to size (revenue)
5. Low-optimality weights (with Poisson sampling), after Deville and Lavallée (2006)
6. Low-optimality weights (with simple random sampling), after Deville and Lavallée (2006)
7. Low-optimality weights (with Poisson cluster sampling), after Deville and Lavallée (2006)
8. Exact calculation of enterprise inclusion probabilities
9. Selection of one representative establishment for each enterprise

We carried out simulations in which we produced estimates of totals and their variance. Our analyses showed that methods 5, 6, 7 and 8 yielded the most promising results. The article should be completed shortly (see Lavallée and Labelle-Blanchet, 2011). We are planning to give presentations on the theory and the results, in particular at the next annual meeting of the Statistical Society of Canada.

Generalization of the Taylor linearization method for total variance estimation

A method of estimating the sampling variance by Taylor linearization was developed by Demnati and Rao (2002, 2004, 2007). The method deals with two sources of error: sampling error and non-response error. The purpose of this research project is to examine a generalization of the method to cover additional sources of error. Essentially, we plan to generalize the linearization by adding the variables of interest, control totals and other variables associated with the sources of error in order to extend the concepts included in the total variance. In this project, we hope to define the theoretical bases for the estimation of total variances for some simple cases (sampling variance with non-response; sampling variance with control totals, including known precision) and construct a simulation study to check the validity and precision of the suggested method. We also want to compare this method with Demnati and Rao's proposed method for estimating the variance when there is non-response.

A simulation program was developed, and an initial series of tests was conducted. The tests, which used an estimator of the total reweighted for non-response, explored three components of variance: sampling error, non-response error and error due to noise in the data (to simulate capture or response errors, for example). The method was validated through an analysis of the estimated parameter's rates of inclusion in the 95% confidence interval. The results of the tests were sufficiently positive to warrant continuing the project.

The same program was then modified for a simulation with a second series of tests. These tests, which use a ratio estimator of the total with calibration on a control total for which the CV is known, look at two components of variance: sampling error and error due to the control total used for calibration. We are in the process of validating the method through an analysis of the estimated parameter's rates of inclusion in the 95% confidence interval.

The hypotheses, theoretical bases, methodology, results and analysis will be clearly described in a working paper (see Demers and Godbout, 2011). Comments will be added to the programs so that they can easily be reused in the future.

At some point, we plan to give a presentation on the project at a Division seminar. The method may also be tested on other, more complex estimators and compared with Demnati and Rao's proposed method for estimating the variance when there is non-response.

Sample Coordination

Statistical agencies often need to coordinate the selection of samples drawn from overlapping populations so as to maximize or minimize the expected overlap, subject to constraints determined by the survey designs. For instance, maximizing the expected overlap between repeated samples can increase the precision of estimates of change between two occasions and reduce the costs of first contacts. Meanwhile, minimizing the expected overlap can avoid overburdening respondents with multiple surveys.

During the fiscal year 2010-2011, we proposed the development of a new sample coordination method that attains the first-order inclusion probabilities for an updated survey and maximizes the expected overlap of samples between the two occasions. We are searching for a method that has good conditional properties for positive coordination and yields efficient estimators of change between the two survey occasions. Research on the impact of maximizing the overlap of surveys on the variance estimator of change will also be carried out (see Qualité and Tillé, 2008).

To deal with a manageable number of variables, we resolved the problem into two steps, without compromising on our goals, that is, we obtained a solution with good conditional properties and a maximum expected overlap of samples. In the first step, we grouped the original samples into configurations, defined two objective functions (LP and LQ) that represent the overlap of configurations, and used linear programming techniques to solve the problem at this level. We compared the LP and LQ solutions numerically to the solution proposed by Kish and Scott (JASA 1971, 66) on a medium-size and large-size stratum of the updated survey. The results provided by LQ were very satisfactory, and we presented them at the 2010 Joint Statistical Meeting. The presentation was well received.

In the second step, we considered three options for selecting ultimate samples within each configuration (pps systematic, Sampford pps, and a new LP method), all of which attain the maximum overlap within each configuration. With the pps solution, one cannot easily estimate the second order inclusion probabilities in the updated survey. The Sampford method does not have this drawback; however, the deviations from the maximum overlap for each initially selected sample ("row-errors") are not always small. The LP solution that we propose at this level minimizes the "within" variance of row errors and gives the probabilities of selection of any sample in the updated survey. In conclusion, the combination of the LQ problem at the level of configurations and our LP

solution within each configuration gives an overall solution that minimizes the total variance of row-errors and maximizes the expected overlap, subject to the constraints imposed on the second design.

Our theoretical and numerical results have been included in a technical report that is now under revision by the team members (see Schiopu-Kratina, Fillion, Mach and Reiss, 2011). We intend to make it a Statistics Canada technical report first and then to submit it to a refereed journal.

Research on the impact of maximizing the overlap of surveys on the variance estimator of change will also be carried out (Qualité and Tillé, 2008). This research is also useful for controlling the overlap of rotating samples.

Mean bootstrap properties

Mean bootstrap is a resampling method that is similar to the Rao-Wu standard bootstrap. In this method, $R \times B$ bootstrap replication is first generated using the Rao-Wu method. Each of the B mean bootstrap weights is then calculated by taking the mean of R Rao-Wu bootstrap weights. Finally, the B mean bootstrap weights are used to estimate variance. The advantage of this technique is that it reduces the chances of having nil bootstrap weights. Nil bootstrap weights can be a problem when a ratio is estimated for a small area: they can result in division by zero. They can also cause convergence problems, for example, in the case of logistic regression models.

During the year, we carried out simulation studies that showed that the mean bootstrap tends to be more biased than the Rao-Wu standard bootstrap in the estimation of a median, particularly for small to moderate sample sizes. The bias appears to increase as the number of replicates, R , increases. We also compared the mean bootstrap with the method described by Saavedra (2001). Saavedra's method also has a tendency to be more biased than the Rao-Wu standard bootstrap. In addition, we showed that no bootstrap method seems to work very well for estimating the variance of a median for a discrete variable for which some values occur very frequently in the population.

For more details, see Mach, Beaumont, Bosa and Lafortune (2010).

Calibration for variance estimation

Calibration can be used to estimate variances. See Singh, Horn and Yu (1998) or Thériberge (1999), for example. In this project, the variance will be viewed as a weighted total. For a population of size N with a variable of interest y_i $i = 1, 2, \dots, N$, the variance is a weighted sum of z_k $k = 1, 2, \dots, N^2$ with $z_k = y_i y_j$ if $k = ri + ji$, $j = 1, 2, \dots, N$. Calibration could be used to estimate the weighted sum of the z_k .

With calibration, the distance between the vector of calibrated weights of the units in a sample s , $\mathbf{w}_s$, and the vector of Horvitz-Thompson weights, $\mathbf{a}_s$, is minimized (under a constraint involving auxiliary data). This distance can be expressed as $(\mathbf{w}_s - \mathbf{a}_s)' \mathbf{U}_s (\mathbf{w}_s - \mathbf{a}_s)$, where $\mathbf{U}_s$ is a positive diagonal matrix. Because the $\mathbf{z}_k$ are correlated, we would have to generalize calibration so that the matrix $\mathbf{U}$ is positive semi-definite but not necessarily diagonal.

The project is to generalize calibration where, in the general regression setting, we would say that the covariance matrix of the vector of interest, under the model, is not necessarily diagonal.

Godambe and Joshi (1965) proved the existence of a lower bound for model expectation of the variance of an unbiased estimator. They assumed that the model had a diagonal variance-covariance matrix. Their result was generalized to a model with a positive definite variance-covariance matrix (see Thériberge, 2011). However, the lower bound is valid only for the family of linear estimators. We also simplified the form of the lower bound. In particular, it is no longer necessary to calculate the root of matrix $\mathbf{U}$. The proof that it is a lower bound was also simplified.

We are planning to generalize the lower bound for a model with a positive semi-definite variance-covariance matrix and show that the bound is valid for the family of unbiased estimators (not just unbiased linear estimators). We will also need to describe a calibration estimator that reaches that bound.

Calibration adjustment to globally minimize the bias due to various non-response phases

The Canadian Health Measures Survey (CHMS) collects health indicators by taking direct measurements for a representative sample of the Canadian population. It is divided into a number of collection stages, each of which results in several phases of non-response: the attempt to contact the household, the individual questionnaire, the

clinic visit and the measurements made on the specimens. Key information is obtained at those stages, including household composition, questionnaire responses, and the direct measurements taken at the clinic.

It would therefore be useful in this context to make a calibration adjustment for all non-response phases at once rather than model each phase independently (as is currently done using logistic regression models). This would make it possible to keep key estimates for each phase intact. In addition, even though the CHMS has a large number of measured variables, the number of calibration constraints to consider would be limited by the number of target domains, of which there are 10 (5 age groups cross-tabulated with gender at the national level).

To do this, this project will develop a multiple non-response phase version of the technique for selecting calibration constraints that minimize non-response bias, developed by Särndal and Lündström (JOS, 2008). In addition, we will implement the improvements to the method that were made in another calibration project (Verret and Kevins, 2010) to control multicollinearity and avoid calibrating for groups with too few respondents.

A presentation entitled "Calibrating Complex Survey Weights to Adjust for Refusals to Record Linkage" was given at the 2010 Annual Meeting of the Statistical Society of Canada in Québec City. The article was written and published in the proceedings. It provided a description of the work done in the previous fiscal year for one non-response phase.

Generalized bootstrap

The generalized bootstrap technique can be used to estimate the variance of estimators for general sampling designs. In the context of this methodology, bootstrap weights are defined so that the first two (or more) design moments of the sampling error are tracked by the corresponding bootstrap moments. Most bootstrap methods in the literature can be viewed as special cases. We have studied in greater depth the case of Poisson sampling, which is often used to select samples in Price Index surveys, and we have written a paper that was presented to the Advisory Committee on Statistical Methods (Beaumont and Patak, 2010). A more detailed version of our paper has been submitted to the *International Statistical Review* and is currently under revision (Beaumont and Patak, 2011).

Robust estimation

We have developed a unified approach to robust estimation in finite population sampling that is based on the conditional bias of a unit (bias conditional on a particular unit being selected in the sample). The conditional bias is used as a measure of influence as it is closely related to the influence function in classical statistics. Unlike the influence function, the concept of conditional bias is straightforward to extend to the design-based framework. Our robust estimator is obtained by reducing the impact on the sampling error of the most influential sample units. It reduces to the estimator developed by Kokic and Bell (1994) when a stratified simple random sampling design is used. We have conducted a simulation study that shows the good properties of our estimator and have refined a paper that was submitted for publication in a refereed journal (Beaumont, Haziza and Ruiz-Gazen, 2011).

For further information, contact:

Pierre Lavallée (613-951-2892, pierre.lavallee@statcan.gc.ca).

References

Chang, H.-J., Wang, C.-L. and Huang, K.-C. (2004). Simple Random Sample Equivalent Survey Designs Reducing Undesirable Units From a Finite Population. *Statistical Papers*, Vol. 45, 287-295.

Demnati, A., and Rao, J.N.K. (2002). *Estimateurs de la variance par linéarisation pour des données d'enquêtes avec des réponses manquantes*. Methodology Branch Working Paper, Statistics Canada, SSMD-2002-007 E/F, November 2002.

Demnati, A., and Rao, J.N.K. (2004). Linearization variance estimators for survey data. *Survey Methodology*, Vol. 30, No. 1, pp. 17-26.

Demnati, A., and Rao, J.N.K. (2007). Linearization Variance Estimators for Survey Data: Some Recent Work. Proceedings of the International Conference on Establishment Surveys (ICES) III, June 18-21, 2007, Montréal, Canada.

Deville, J.-C., and Lavallée, P. (2006). Indirect sampling: The foundations of the generalized weight share method. *Survey Methodology*, Vol. 32, No. 2, 165-176.

- Godambe, V.P., and Joshi, V.M. (1965), Admissibility and Bayes estimation in sampling finite populations, 1. *Annals of Mathematical Statistics*, 36, 1707-1722.
- Hedeyat, A.S., and Robieson, W.Z. (1998). Exclusion of an Undesirable Sample for the Support of a Simple Random sample. *The American Statistician*, Vol. 52, No. 1, 41-43.
- Kish, L., and Scott, A. (1971). Retaining Units after Changing Strata and Probabilities. *Journal of the American Statistical Association*, 66, 461-470.
- Kokic, P.N., and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 419-435.
- Mach, L., Reiss, P.T. and Schiopu-Kratina, I. (2006). Optimizing the Expected Overlap of Survey Samples via the Northwest Corner Rule. *Journal of the American Statistical Association*, 101, 1671-1679.
- Matei, A., and Skinner, C. (2009). Optimal sample coordination using controlled selection. *Journal of Statistical Planning and Inference*, 139, 3112-3121.
- Matei, A., and Tillé, Y. (2006). Maximal and minimal sample co-ordination. *Sankhyā*, 67, 590-612.
- Nedyalkova, D., Pea, J. and Tillé, Y. (2008). Sampling Procedures for Coordinating Stratified Samples: Methods Based on Microstrata. *International Statistical Reviews*, 76, 368-386.
- Qualité, L., and Tillé, Y. (2008). Variance estimation of changes in repeated surveys and its application to the Swiss survey of value added. *Survey Methodology*, 34, 173-181.
- Reiss, P.T., Schiopu-Kratina, I. and Mach, L. (2003). The use of Transportation Problem in Coordinating the Selection of Samples for Business Surveys. *Proceedings of Survey Methods Section, Statistical Society of Canada Annual Meeting*, June 2003.
- Saavedra, P. (2001). An Extension of Fay's Method for Variance Estimation to the Bootstrap. *Proceedings of the Survey Research Methods Section*, American Statistical Association.
- Särndal, C.-E., and Lundström, S. (2008). Assessing auxiliary vectors for control of nonresponse bias in the calibration estimator. *Journal of Official Statistics*, Vol. 24, 167-191.
- Singh, S., Horn, S., and Yu, F. (1998). Estimation of variance of general regression estimator: Higher level calibration approach. *Survey Methodology*, 24, 41-50.
- Théberge, A. (1999). Extensions of calibration estimators in survey sampling. *Journal of the American Statistical Association*, 94, 635-644.

Small area estimation

Issues in the application of Penalized Spline models for small area estimation

In this project we consider an extension of the classical Fay-Herriot (1979) model, namely, the Penalized Splines (PS) model, used to accommodate departures from linearity in the linking model.

In order to accommodate a non-linear functional relationship, we consider a mixed linear linking model that can be viewed as a combination of fixed effects, random spline effects and an area random effect. We call this model the Penalized Splines (PS) model.

The small area means are estimated by the empirical best linear unbiased prediction (EBLUP) estimators. However we cannot apply the usual methods of estimation for the mean squared error (MSE) of the EBLUP because the variance structure is not block diagonal. Previously, we had proposed a parametric bootstrap procedure to estimate the MSE of the EBLUP estimators and investigated its properties in terms of its Monte Carlo relative bias (Rubin-Bleuer, Dochitoiu and Rao, 2009a and Rubin-Bleuer, Dochitoiu and Rao, 2009b).

This year, we continued to investigate the properties of the bootstrap procedure with a more extensive simulation (Jamroz, 2010). We concluded that when the number of areas modeled is moderate, say 50, the average of the

relative bias of the bootstrap MSE over groups of 10 areas with the same sampling variance decreases as the number of knots increases (for 50 areas, a reasonable maximum number of knots is 20). However, even with a large number of knots, the relative bias of each individual bootstrap MSE can be up to 35%, which makes for a very poor estimator. In the literature, the examples refer to modeling a number of observations larger than 250 and they recommend a number of knots of 40. In that case, the bootstrap estimator may have better relative bias, but this is not a common scenario in business survey data. We continue to investigate other forms of bootstrap estimators to better approximate the MSE.

Time-series and cross-sectional small area models of the Yu-Rao type

The Survey of Employment, Payrolls and Hours (SEPH) is a monthly survey designed to produce estimates of levels and month-to-month trends of payrolls, employment, paid hours and earnings. SEPH parameters are currently estimated by the Generalized Regression (GREG) estimator. In an attempt to obtain estimators with improved precision, the SEPH program is planning to use a composite estimator RC, which is a Generalized Regression (GREG) estimator that borrows strength across time by using information from previous occasions of the survey as auxiliary data. The purpose of this project is to examine different models that incorporate the same information than the RC estimators for small area estimation in SEPH. We use a cross-sectional and time series model of the Yu-Rao type (Rao, JNK, 2003) and the empirical best linear unbiased estimator (EBLUP). Last year, we compared the EBLUP with the GREG estimator when the variance components were estimated by the method of moments and by the method of Wang and Fuller (2003). Our results are contained in Yung, Rubin-Bleuer and Landry (2009).

This year, we concentrated on investigating the best method of estimation for the model variances. We looked at design-based and model-based properties. For the design-based Monte-Carlo mean squared error (MSE), we used the simulated SEPH population and sampling covariance matrices obtained via design-based Monte-Carlo methods: we drew 500 independent samples following the SEPH sampling design and calculated the MSE by the average of the corresponding sampling variance estimates. To calculate the model-based MSE, we simulated several populations with differing sampling covariance matrices.

We estimated the variance components by the adjusted maximum likelihood method (ADM) originally developed by Li and Lahiri (2010) for the Fay-Herriot (1979) model. We adapted it to this time series and cross-sectional model.

We compared the Wang and Fuller (2003) and the restricted maximum likelihood (REML) methods of variance estimation with the ADM method in terms of the design-based Montecarlo MSE (with 1,000 independent samples).

The results of this design-based study were presented as an invited paper at the Joint Statistical Meetings in 2010.

In the model-based project, we studied the model-based properties of the ADM variance estimator.

We showed that the ADM estimator is $\sqrt{m}$ -consistent; we obtained a second order approximation of the bias of the ADM estimator, a second order approximation of the MSE of the EBLUP, and an unbiased estimator of the MSE following the same argument of Datta and Lahiri (2000).

At present we are looking at a model-based simulation to study the finite properties of the ADM method of variance estimation. Preliminary results show that the procedure of combining the REML and ADM estimators (the latter to be used when REML produces a zero estimate) yields the most efficient EBLUP estimator. We plan to present the model-based study at SAE2011.

Small area estimation under informative sampling

We compared via simulations small area estimation procedures based on unit-level models, namely, the unweighted EBLUP (Battese, Harter and Fuller, 1988), the weighted pseudo-EBLUP (You and Rao, 2002) and a procedure proposed by Pfeiffermann and Sverchkov (2007) to take design informativeness into account. We considered informative two-stage sampling designs where all the first stage units (small areas) are selected with certainty and the second stage units are selected using the Rao-Sampford nPS sampling scheme (Rao 1965, Sampford 1967). Various degrees of design informativeness were considered and this was controlled through a tuning parameter used to generate probabilities of selection similar to the parameter used in the simulation of Asparouhov (2006). We also considered a version of the model used to construct the EBLUP and pseudo-EBLUP that included the second stage weights as an additional auxiliary variable to attempt to make the resulting estimators more robust to design informativeness. At SSC 2010, we reported the bias and MSE of the point

estimators and the relative bias of the MSE estimators. To pursue the work presented at the SSC, we did another simulation with probabilities of selection generated using a scheme similar to the one found in Pfeiffermann and Sverchkov (2007).

The project also developed into a study of the estimation of two-level model parameters under informative sampling. Consistent point estimators were derived using a weighted estimating equations approach. It was also shown that these estimators can be obtained through maximisation of the weighted composite-likelihood. Jon Rao presented this at SSC 2010 along with the results of additional simulations done using the first setting described above. In these simulations, the newly developed point estimators performed very well compared to competitors such as the REML, the estimators of Korn and Graubard (2003) and two estimators of Aparouhov (2006). Only nested error mean models were simulated, but estimators for nested error linear models were developed for future simulations.

Variance component estimation methods for the Fay-Herriot model

The Fay-Herriot model is the most popular model used in small area estimation to improve direct survey estimates. The Fay-Herriot model has two variances, namely, a model variance associated with the area random effect and a sampling variance associated with the sampling errors. The sampling variance can be estimated from the survey data and can also be smoothed using various methods. The model variance is unknown and needs to be estimated from the model. In small area estimation, various methods have been proposed to estimate the model variance. In this project, we study the various estimation methods including the fitting constant (FC) method, REML method, FHI (Fay-Herriot Iterative) method, Wang and Fuller's method (Wang and Fuller, 2003), and the most recently developed ADM (adjusted density maximum) method (Li and Lahiri, 2010). We conduct a simulation study along the line of Rivest and Vandal (2002) to evaluate the methods of estimation variance components and small area parameters under the EBLUP approach. We also study the effect of input sampling variances under these different estimation methods for the model variance, particularly for the MSE estimation when the direct sampling variance estimates are used, as in Wang and Fuller (2003). This will give us an indication of the proper use of estimation methods under sampling variance smoothing and modeling. The sampling variances will be generated from a chi-square model as in Rivest and Vandal (2002), and then the small area data will be generated from a Fay-Herriot model. We have finished the simulation study and obtained interesting results. We compared the different variance component estimation methods and their impact on the MSE estimation. The estimation methods are coded in S-Plus. A Methodology Branch working paper (You, 2010) has been finished and documented. The results of this study can be used as a reference guide for choosing which method to use to estimate the model variance in practice as well as for the small area estimation system development at SRID.

Benchmarking model-based small area estimators

In small area estimation, it is important and desirable for survey sampling users to benchmark model-based small area estimates so that the benchmarked model-based estimates add up to the total of direct estimates at the same level or higher. You and Rao (2002) proposed a pseudo-EBLUP estimator for a unit level model with self-benchmarking. For area level models, Wang, Fuller and Qu (2008) proposed a method for adding weighted sampling variances as another auxiliary variable in the regression model to achieve self-benchmarking of the EBLUP estimates. They also applied the You and Rao (2002) method to the area level model to obtain the You-Rao (YR) self-benchmarking estimator at the area level. In this project, we obtain the MSE estimator for the YR self-benchmarking estimator under the Fay-Herriot model. We will compare the two self-benchmarking estimators, namely, the Wang-Fuller-Qu (WFQ) augmented EBLUP estimator and the YR estimator, through a simulation study. We are particularly interested in the MSE estimation of the benchmarking estimators compared to the usual EBLUP estimator. We have finished a simulation study that compares the EBLUP, YR and WFQ estimators in terms of MSE estimation. A draft paper has been written up (You, Rao and Hidirolou, 2011). The paper will be finalized and submitted to a journal for possible publication.

Hierarchical Bayes sampling variance modeling for small area estimation

In small area estimation, when area level models are used to improve the direct survey estimates, the sampling variances are usually assumed to be known in these area level models. This is a strong assumption, as the sampling variance is also based on the direct survey estimates. A smoothing approach is usually employed to smooth the direct sampling variance estimates, and smoothed sampling variance estimates are obtained using external generalized variance functions and models. Alternatively, we can model the sampling variance estimates and use various prior models for the true sampling variances. The advantage of the modeling approach is that we can combine the sampling variance models and the small area models for the direct survey estimates together.

The integrated model borrows strength for small area estimates and sampling variance estimates simultaneously, unlike the smoothing approach, where the smoothed sampling variance estimates are obtained separately from the small area models. In this project, we consider the Fay-Herriot model as the basic studying model and study various sampling variance models under a hierarchical Bayes (HB) framework. We have started the project by extending the work of You and Chapman (2006) for sampling variance modeling under the Fay-Herriot model. We have reviewed and investigated some models for sampling variances. We have finalized the integrated models and developed the Gibbs sampling programs for the HB inference for the integrated model under different sampling variance models. We will compare the different models and the sensitivity of the small area estimates to the choice of sampling variance models through the analysis of different small area survey data. A draft paper (You, 2011) is in progress. We will extend the sampling variance integrated model to the small area spatial models (You and Zhou, 2011) and unmatched nonlinear models (You and Rao, 2002).

Development of Small Area Estimation System

This project has involved review and documentation of small area estimation methods and their implementation in a SAS prototype program. The current system produces the model-based small area estimates based on both the area level and unit level models. The area level model is currently based on the Fay-Herriot model (Fay and Herriot, 1978) and the unit level model is based on the nested error regression model (Battese, Harter and Fuller, 1988). More complex models such as the unmatched models and time series models will be included in the system later.

In summary, we have finished the following steps for the system.

1. Area level estimation based on the Fay-Herriot model is finished. The input sampling variances can be unsmoothed or smoothed. Four variance component estimation methods, namely, REML, FHI, WFI and ADM, are implemented. Benchmarking procedures are developed to ensure the robustness of the model-based estimates. Model evaluation procedures including various diagnostics and tests have been specified and implemented in the system. Methodology documentations and system specifications for the Fay-Herriot model have been finished.
2. EBLUP estimation based on the unit level nested error regression model has been implemented. Model diagnostics and tests for the unit level model are under development. Documentations and methodology specifications based on the unit level model are in progress.

We are currently doing evaluation for the unit level models. We will also implement the pseudo-EBLUP approach (You and Rao, 2002) based on the unit level model taking account of the survey weights. The pseudo-EBLUP estimates are design-consistent and satisfy certain benchmarking conditions. Also, hierarchical Bayes (HB) inference method with Gibbs sampling will also be introduced to the system to deal with more complex models and provide the users with more options to choose small area estimation methods and models to improve the direct survey estimates from both the area level and unit level models.

For further information, contact:

Susana Rubin-Bleuer (613-951-6941, susana.rubin-bleuer@statcan.gc.ca).

References

- Asparouhov, T. (2006). Generalized multi-level modeling with sampling weights. *Communications in Statistics - Theory and Methods*, 35, 439-460.
- Battese, G.E., Harter, R.M. and Fuller, W.A. (1988). An Error Component Model for Prediction of County Crop Areas Using Survey and Satellite Data. *Journal of the American Statistical Association*, 83, 28-36.
- Datta, G., and Lahiri, P. (2000). A unified measure of uncertainty of estimated best linear unbiased predictors in small area estimation problems. *Statistica Sinica*, 10, 613-627.
- Fay, R.E., and Herriot, R.A. (1979). Estimation of Income from Small Places: An Application of James-Stein Procedures to Census Data. *Journal of the American Statistical Association*, 74, 269-277.
- Hall, P., and Maiti, T. (2006). Nonparametric estimation of mean squared prediction error in nested error regression models. *The Annals of Statistics*, Vol. 34, N04, 1733-1750.
- Korn, E.L., and Graubard, B.I. (2003). Estimating variance components using survey data. *Journal of the Royal Statistical Society B*, 65, 175-190.

- Li, H., and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. *Journal of Multivariate Analysis*, 101, 882-892.
- Opsomer, D., Claeskens, G., Ranalli, M.G., Kauermann, F. and Breidt, F.J. (2008). Nonparametric small area estimation using penalized spline regression. *Journal of the Royal Statistical Society, Series B*, 70, 265-286.
- 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.
- Rao, J.N.K. (1965). On Two Simple Schemes of Unequal Probability Sampling without Replacement. *Journal of the Indian Statistical Association*, 3, 173-180.
- Rao, J.N.K. (2003). *Small Area Estimation*. New York: John Wiley & Sons, Inc.
- 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.
- Rubin-Bleuer, S. Dochitui, D. and Rao, J.N.K. (2009a). Bootstrap estimation of the mean squared error in small area models for business data. Long abstracts, Small Area Estimation 2009, International Statistical Institute Satellite Conference.
- Rubin-Bleuer, S. Dochitui, D. and Rao, J.N.K. (2009b). Bootstrap estimation of the mean squared error in penalized spline small area models. Internal SRID report, 2009.
- Sampford, M.R. (1967). On Sampling without Replacement with Unequal Probabilities of Selection. *Biometrika*, 54, 499-513.
- Wang, J., and Fuller, W.A. (2003). The mean squared error of small area predictors constructed with estimated area variances. *Journal of the American Statistical Association*, 98, 716-723.
- Wang, J., Fuller, W.A. and Qu, Y. (2008). Small area estimation under a restriction. *Survey Methodology*, 34, 29-36.
- 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). A Pseudo-Empirical Best Linear Unbiased Prediction Approach to Small Area Estimation Using Survey Weights. *The Canadian Journal of Statistics*, Vol. 30, No. 3, 431-439.
- Yung, W., Rubin-Bleuer, S. and Landry, S. (2009). Small area estimation: A business survey application. SSC Annual Meeting, *Proceedings of the Survey Methods Section*, June 2009.

Data analysis research (DAR)

The resources assigned to Data Analysis Research are used for conducting research on methodological problems related to analysis that are currently identified by analysts and methodologists, as well as on problems that are anticipated to be of strategic importance in the foreseeable future. People doing this research also contribute to knowledge transfer and exchange. They also gain experience through publishing technical papers, and by giving seminars, presentations, and courses.

Meta-Analysis of survey data

The recent increase in straightforward access to data has led to a proliferation of analysis that groups or pools multiple study results. This ready access, combined with user-friendly computer software, has led researchers to try to borrow strength from the many different observational and non-randomized studies in an attempt to either improve precision or address research questions for which the original studies were not designed. Researchers have started to employ many different techniques including meta-analysis; however, the analysis is often done without reference to a generalized framework or a systematic review and often without an understanding of the methodological differences between surveys and experiments.

Work has been done on investigating an appropriate framework for the meta-analysis of survey data. This framework has been presented at the Joint Statistical Meetings in Vancouver (See Fox, 2010). The framework was refined over the rest of the year and is in the process of being written into an article for peer reviewed submission (See Fox, 2011).

Estimation and simulation uncertainty: Assessing model and parameter assumptions in microsimulations

A literature review into methods to assess parametric uncertainty in Monte Carlo and microsimulation methods was completed. As each parameter in a microsimulation comes with its own uncertainty, we need to incorporate this uncertainty into the simulation directly. We have chosen to use a latin-hypercube sampling approach for selecting the parameters within the confidence region for each parameter. We will use this range of estimates in separate sub-samples. We will calculate the variance due to the parametric uncertainty by noting that the variance is comprised of the variability between and within sub-sample components. Work has started on a prototype using Modgen. Modgen (Model generator) is a generic microsimulation programming language created at Statistics Canada to support the creation, maintenance and documentation of dynamic microsimulation models.

Variance estimation in longitudinal surveys with imputed responses

The performance of Generalized Estimating Equations for two methods of handling missing responses was compared: the re-weighting approach proposed by Robins, Rotnitzky and Zhao (1995) and a hot-deck imputation method. The simulation results showed that, in a variety of circumstances, the two approaches perform similarly with respect to relative bias and variance (MSE). The results were presented at the 2010 annual meeting of the SSC (Carrillo-Garcia, Kovačević and Wu, 2010).

Spatial analysis of geocoded data

The effects of spatial autocorrelation should be an important consideration when modeling regional crime distributions as this violates the assumption of statistically independent observations for regression analysis. Analysis of crime data was enhanced in 2009 with the development of the police-reported crime severity index, which measures the seriousness of crime in specific regions. Spatial lattice data, such as neighbourhood crime data that is aggregated to census tracts or dissemination areas, can be analyzed in a number of ways and has already been researched in Collins, Babyak and Moloney (2006), Collins and Singh (2008) and Collins (2009). During this fiscal year, spatial analysis techniques have been applied to geocoded crime data using the severity index instead of the traditional crime rate. Also, analyses have been performed to compare the use of different continuity structures to define the neighbouring regions. More research on this topic is planned and the results are being written up for publication. A seminar will be presented next fiscal year based on these results.

Selected topics in design-based methods for analysis of survey data

This project tackles a variety of research problems that have been identified through contact with analysts. Often, these topics are chosen based on consultations for which no satisfactory answer was immediately apparent.

1. Population risk predictions using the survey bootstrap

The research on this topic was done as part of a consultation with ICES. The research was written up (Kovačević, Roberts and Mach, 2010) and submitted to *Survey Methodology*. It has been conditionally accepted.

2. Inferring causality from survey data

A literature review was carried out on the topic of causal inferences from observational data. A technical note summarizing the findings has been written (Binder, 2011a). An invited talk on the topic is being prepared for the 2011 SSC annual meeting.

3. A model-design randomization framework for analytical methods with complex survey data

A paper on the topic is being written (Binder, 2011b), to be submitted to the *Pakistan Journal of Statistics* for a special issue dedicated to Ken Brewer (Festschrift).

4. New Developments in modelling using survey data

A discussion was prepared for the papers presented in this invited session at the 2010 SSC annual meeting (Roberts, 2010).

5. Methodological issues in the meta-analysis of observational studies

A discussion was prepared for the papers presented in this invited session at the 2010 Joint Statistical Meetings. A

write-up of the discussion was submitted to the *Proceedings of the Joint Statistical Meetings* (Binder, 2010).

6. Variance estimation when analyzing census long-form microdata

Further investigations were carried out using the 2001 long-form census microdata file available in the RDCs, which was enhanced with additional survey design information. The purpose of the investigation was to determine what aspects of the survey design could be accounted for in variance estimation in survey analysis software readily available to analysts and for which of these aspects were there major changes in the variance estimates. Several different scenarios were studied. A report on the findings will be completed soon and will likely be submitted to the RDC Technical Bulletin.

Bootstrap for model parameters

The Rao-Wu bootstrap is often used to estimate the design variance of estimators of finite population parameters. When estimating model parameters, two sources of variability must normally be taken into account for variance estimation: the sampling design and the model that is assumed to have generated the finite population. If the sampling fraction is negligible, the model variability can in principle be ignored, which is often done in practice, and the Rao-Wu bootstrap remains valid. However, this simplification is not always appropriate. The generalized bootstrap method can be used to correctly take into account both sources of variability. Our procedure may be used for any parameter defined through an estimating equation as long as the observations are assumed to be independent under the model. It is simple to apply once bootstrap weights that capture the first two design moments have been obtained (e.g., using the Rao-Wu bootstrap method). We have revised a paper (Beaumont and Charest, 2011) that was earlier submitted to *Computational Statistics and Data Analysis*.

For further information, contact:

Georgia Roberts (613-951-1471, georgia.roberts@statcan.gc.ca).

References

Collins, K. (2009). Spatial Modelling of Geocoded Crime Data. *Proceedings of the Survey Methods Section, Statistical Society of Canada*.

Collins, K., and Singh A. (2008). Use of Optimal Instrumental Variables for Spatial Autoregressive Models With Application to Crime Survey Data. *Proceedings of the Joint Statistical Meetings*.

Collins, K., Babyak, C. and Moloney, J. (2006). Treatment of Spatial Autocorrelation in Geocoded Crime Data. *Proceedings of the Joint Statistical Meetings*.

Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *JASA*, 90, 106-121.

Data collection

Optimizing and simulating CATI call scheduling

One of the main challenges for Statistics Canada is to maintain cost-effective collection strategies while continuously achieving a high level of quality. Paradata research has been useful for improving the current data collection process and practices. The research carried out with paradata suggested that collection resources are at present not always optimally allocated with respect to the assigned workload and the corresponding expected productivity.

Two approaches were tested for optimizing Computer Assisted Telephone Interview (CATI) data collection for a single survey. In the first approach, the collection process was mimicked via a micro simulation system to evaluate how different collection strategies can affect productivity and collection cost (see the project below **Microsimulation de la collecte téléphonique**). In the second approach, models were developed to predict the probability that a telephone call would result in a completed questionnaire as a function of time of day and resources over a specific collection period. The estimated parameters were inputted into a loss function that optimizes call scheduling subject to constraints.

The results suggested that cost savings could potentially be achieved for this specific case.

Future work will include an attempt to solve the problem for several surveys using CATI concurrently.

Microsimulation of telephone collection

This research project involves constructing a microsimulation model to represent a survey's telephone collection process as accurately as possible. The aim of the project is to modify collection parameters, such as time slots and interviewer distribution during collection, in a controlled environment to study the impact on collection results. We hope to gather more reliable information more quickly than we would actual field testing.

The project has two main components. First, a survey paradata analysis component determines interview result probabilities using a multinomial logistic model. The probabilities are then used as input data for the second component, which involves simulating the call scheduler (Blaise) used for telephone collection. The microsimulation model for this component is constructed in advance using Simulation Studio, SAS's new simulation software.

The simulation model has been thoroughly tested. As well, results obtained using the paradata from the Canada Survey of Giving, Volunteering and Participating shows that the model is working as expected.

We currently have a fully functioning model that consists of the following features: time slices, cap on calls, number of interviewers available throughout the collection period and per shift, and priority of each case.

The simulations outputs produced reasonable results (as expected, more interviewer efforts in the evening produces higher response rates).

A presentation was made to the Joint Statistical Meetings in August in Vancouver, and the accompanying paper was finalized.

Due to some limitations of the SAS Simulation Studio prototype, the possibility to use a different simulation software, such as ModGen, is being investigated.

Quality Indicators

Response rate is one of several measures of survey data quality. Statistics Canada program divisions, survey sponsors and data users put a lot of emphasis on the importance of response rate targets. Program managers and their survey methodologists tend to push for high overall survey response rates in order to ensure that reliable estimates for all sub-populations or geographic regions within their sample can be released. From a collection standpoint, the response rate targets set by program divisions are often not realistic, as they may not reduce the non-response bias.

The objectives of the research were to

- develop alternative non-response measures to target data collection follow-up; and
- determine response rate based thresholds that may be used to measure the reliability of released estimates.

The research investigated available statistical tools that could be used to measure non-response bias or to provide indications of the extent of the bias that may be reduced if the response rate is increased systematically by targeting certain groups of non-respondents for follow-up.

The work done during this period was an extensive literature review on statistical tools that may be used to measure or provide indications of the extent of non-response bias.

Guidelines for Statistics Canada Internet Questionnaires

New challenges are being raised by the use of multi-mode collection in Statistics Canada surveys, particularly the recent use of self administered e-questionnaires or Internet collection. To ensure that the information collected meets our quality requirements, special care must be taken in the development of the conceptual framework for e-questionnaires. Many of the design principles for self-administered questionnaires can be applied to Internet questionnaires. However, there are also a number of other factors to consider; such as the readability which can depend on the type of browser, operating system and screen resolution used by the respondent. A lot of research is currently being done in this field in the international community. As part of a previous research project, a review of the different design principles for Internet questionnaires that have been published by leading researchers was started. Experience in designing Internet questionnaires was also gained with the implementation

of the Internet questionnaire for the Census and an upcoming pilot project for the Labour Force Survey. A corporate solution for the development of e-questionnaire applications is currently being developed.

One of the goals of this research project is to summarize and document the knowledge gained so far based on international research and Statistics Canada's experience with Internet collection. This will take the form of a guide for designing Internet questionnaires for Statistics Canada surveys. Another goal is to identify specific e-questionnaire design issues that need to be further evaluated. For example, different principles may apply to our business and social surveys; they may depend on the survey content, target population and other constraints such as the common look and feel standards.

The Questionnaire Design Resources Centre of Statistics Canada (QDRC) has been directly involved in the development and implementation of end-user usability test strategies for several business surveys that are moving to the e-questionnaire environment. This has been a collaborative effort with representatives from key areas of Statistics Canada responsible for the development and design of e-questionnaires application, the management of the collection process, and dissemination activities.

Experience and knowledge regarding e-questionnaires were also acquired and shared while participating as an active member of the new e-questionnaire Standards Committee. As part of its mandate, this inter-disciplinary, inter-divisional working group is developing an *E-Questionnaire Design Guidelines and Standards* manual. A draft version of this document has been written and is currently under review.

Future work will involve summarizing and documenting the knowledge gained so far based on international research as well as Statistics Canada's experience with Internet collection.

Planning for a Collection Research Resource Centre

We are proposing the creation of a Collection Research Resource Centre that would support operational and methodological research associated with collection for business, social and agricultural surveys. A first proposal was produced that outlined the following:

- the mandate of the Centre;
- a description of the organizational structure to create, maintain and administer the centre (membership, staff management, reporting structure, planning of activities);
- a description of the types of human resources needed to carry out the planned activities;
- a description of the administrative features (such as a website) that will help create an efficient working environment; and
- short term and long term goals.

More details will be added to the proposal after further discussions take place with the different partners involved.

For further information, contact:

Hélène Bérard (613-951-1458, helene.berard@statcan.gc.ca).

References

- Schouten, B., Cobben, F. and Bethlehem, J. (2009). Indicators for representativeness of survey response. *Survey Methodology*, 35, 101-113.
- Särndal, C.-E., and Lundström, S. (2010). Design for Estimation: Identifying auxiliary vectors to reduce nonresponse bias. *Survey Methodology*, 36, 131-144.
- Bankier, M. (1988). Power Allocations: Determining Sample Sizes for Subnational Areas. *The American Statistician*, Volume 14, no. 3.
- Lavallée, P., and Hidirolou, M. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, Volume 14, no. 1, 33-43.

Disclosure control methods

Disclosure Control Resource Centre

As part of its mandate, the Disclosure Control Resource Centre (DCRC) provided advice and assistance to Statistics Canada programs on methods of disclosure risk assessment and control. It also shared information or provided advice on disclosure control practices with other departments and agencies including Industry Canada, the U.S. Census Bureau, U.S. Energy Information Administration, Statistics Sweden, the Statistics Bureau of Japan, the Children's Hospital of Eastern Ontario Research Institute and the Manitoba Centre for Health Policy. Papers were reviewed for the *Statistical Journal of the International Association for Official Statistics*, the *Statistics Survey* journal, and the Institute of Education Science, U.S. Department of Education.

Continuing support on disclosure control methods is also given to the agency's Research Data Centres (RDC) Program. Most of it is in terms of assistance for the implementation and interpretation of disclosure control rules for RDC data holdings.

Estimating Disclosure Risk for Public Use Microdata Files (PUMFs)

A research project was undertaken to apply theory proposed by Shlomo (2009) to estimate disclosure risk measures for Statistics Canada's PUMFs. Because we often work with data from complex survey designs we had to adapt the proposed method by using the pseudo-likelihood method (to take weights into account). Stratification variables also had to be included as part of the key variables. Results of the investigation were documented in an internal report (Grondin, Haddou and Laperrière, 2010).

MASSC-Basic for Statistical Disclosure Limitation of Large Administrative Databases

Work was completed on applying a simplified version of the nonsynthetic statistical disclosure method MASSC and its generalization GenMASSC (Singh, 2009) to a large administrative database, namely, the Canadian Cancer Registry (CCR). A first draft of a technical report was completed (Michaud, Singh and Tambay, 2010).

Protection of Tabular Data Confidentiality in the Presence of Negative Values

We compared the effects of different methods for the treatment of real-valued data, namely data transformation and multiple decomposition approaches, on cell suppression. A more theoretical framework on the calculation of a sensitivity measure was established. Early results have allowed us to reject one of the methods considered because of undesirable effects. We are analysing results for the other methods. A presentation "Disclosure Control Strategies for Negative Values in Tabular Data" was made at the March 28-29, 2011 Statistics Canada – U.S. Census Bureau interchange.

Study on adding random noise to protect statistical confidentiality

To analyze the impact of injecting noise into the microdata over a number of cycles of the same survey, we first added noise to the data for three consecutive years of the Industrial Water Survey (independently for each year). We compared the resulting estimates and trends with the original estimates and trends. We then adopted a different strategy that involved controlling the noise added from one year to the next. In short, the aim is to introduce noise by applying the same percentages to the units each year. We generated the estimates and trends again and then compared them with the previous ones. The project itself is now complete; all that remains is to document the results.

Creation of synthetic data

There is an emerging literature on methods for the creation of synthetic or simulated data. These methods attempt to preserve as much as possible the relationships in the original data while keeping the risk of divulging confidential information at a low level. The current methods involve modelling the multivariate relationships in the collected data so as to reproduce these observed relationships in the synthetic data.

We have continued synthesizing data from the Cross National Equivalent File (CNEF) in which five other countries are involved. The Canadian portion of the CNEF is a subset of about 40 variables from the Survey of Labour and Income Dynamics. This year, we have synthesized the second year of the six-year panel. Synthesizing changes in the household structure, such as household splits and deaths, was the most difficult issue that we faced and resolved through a careful modeling effort. We plan to complete synthesizing the data for all six years at the end of the next fiscal year. We have written a report that describes our methodology and shows some empirical results (Beaumont and Bocci, 2011). As expected, synthetic data lead to conclusions that are similar to those obtained with real data but there are sometimes some discrepancies. These discrepancies could possibly be removed by refining the models used to generate synthetic data. We plan to update our paper and submit it to a peer-reviewed journal.

An empirical analysis of disclosure control methods

This research project completed the literature review into methods to validate the quality and utility of masked or perturbed data at different levels of aggregation; from micro-data consistency to the quality of aggregated data. An empirical study has been done to compare different methods (suppression, perturbation and hybrid methods) of disclosure control and their effect on data quality, confidentiality, flexibility and coherence using income tax data. The data quality was defined as the proximity to the original data. Formally we looked at the number of suppressed cells, the mean absolute error, the maximum absolute error, the mean relative error and the maximum relative error. Coherence and flexibility concepts are discussed qualitatively.

The results show that Tambay's hybrid method (Tambay, 2009) achieves slightly less suppression than that of the G-Confid cell suppression program but introduces noise in the whole table. The balanced method introduces much less noise and does not suppress. The amount of noise produced by the hybrid is however usually less than the amount generated by the purely random noise method. The table produced by the hybrid method might provide higher protection since the margins do not add up.

For further information, contact:

Jean-Louis Tambay (613 951-6959, jean-louis.tambay@statcan.gc.ca).

References

Shlomo, N. (2009). Releasing Microdata: Disclosure Risk Estimation, Data Masking and Assessing Utility. Working paper.

Tambay, J.-L. (2009). A Hybrid Method for the Disclosure Control of Tables of Magnitude for Administrative Data. Internal note, September 2, 2009.

Statistics Canada

Support Activities

Time series

The projects can be split into seven sub-topics:

- Consultation in Time Series (including course development)
- Time Series Processing and Seasonal Adjustment
- Benchmarking (including development and support to G-Series)
- Reconciliation (including development and support to G-Series)
- Calendarization
- Trend Estimation
- Other R&D Time Series Projects

Consultation in Time Series (including course development)

As part of the Time Series Research and Analysis Centre (TSRAC) mandate, consultation was offered as requested to various clients. Following one of this request, an investigation of treatments for abrupt changes in seasonality (such as seasonal outliers from X-13 and change of regime) was performed and shared within TSRAC.

Some papers were reviewed for official Statistics Canada publications or refereed for external journals.

To support the publication of Wyman (2010), a working paper on gain and phase shift was produced (Quenneville, 2010).

Various courses and introductory workshops were given. The new course on raking was completed and successfully presented as a pilot course. The course will be incorporated in the regular training offering.

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 (for analysis and development or production) or SAS Proc X12 (for production). TSRAC had various exchanges with the SAS Institute and ONS on the use of PROC X12.

The latest versions of X-12-ARIMA, Win X-12 and X-12 Graph packages were downloaded and evaluated as needed.

A beta version of the new Time Series Processing System now includes a benchmarking module (from the G-series/Forillon software). The input and output requirements were also documented (TSRAC, 2010).

Following a suggestion from both the Policy and Methods Committee and the Standards Committee, a horizontal review on seasonal adjustment was carried out within the agency (Julien and Fortier, 2011).

Benchmarking (including development and support to G-series/Forillon)

Benchmarking refers to techniques used to ensure coherence between time series data of the same target variable measured at different frequencies, for example, sub-annually and annually.

Support for PROC BENCHMARKING was provided as needed. Many information requests were received from international agencies and national banks. Draft methodological specifications to include alterability coefficients and to facilitate the use of implicit forecasted benchmarks were prepared.

The use of benchmarking in the context of seasonal adjustment was summarized in a review paper (Quenneville and Fortier, 2010).

Development of compatibility tests continued and was documented. The technique was first presented at the Advisory Committee on Statistical Methods and then updated and submitted for publication (Quenneville and Gagné, 2010).

Benchmarking techniques from an error-model point of view continue to be explored and documented (Chen and Wu, 2010a and 2010b).

Raking (including development and support to G-series/Forillon)

Reconciliation is a method for imposing contemporaneous aggregation constraints on tables of time series, so that the "cell" series add up to the appropriate "marginal total" series.

Support for PROC TSRAKING was provided as needed. The procedure was publicized with a poster presentation at the SAS Global Forum (Bérubé and Fortier, 2010).

Time series reconciliation or raking is most often used on seasonally adjusted data to resolve the inconsistencies between a direct and an indirect adjustment. This use is presented in Mazzi, Fortier and Quenneville (2010).

The last few years of development of the methods and tools for benchmarking and reconciliation were summarized in a review paper (Quenneville and Fortier, 2011).

Calendarization

This sub-topic deals with the support and development of calendarization methods both from a benchmarking point of view and regarding the more recent and extremely promising spline interpolation techniques.

Spline interpolation techniques for calendarization were developed and presented at the Joint Statistical Meeting (Vancouver, August 2010). The paper was also submitted for publication in a refereed journal (Quenneville, Picard and Fortier, 2010).

An update of the calendarization methods and their applications in business surveys was also prepared for the Advisory Committee on Statistical methods (Fortier, Quenneville and Picard, 2010).

Trend estimation

Methods used for long-term trend estimation in environmental data were reviewed. The use of a trend-cycle approach was presented in Bemrose, Meszaros, Quenneville, Henry, Kemp and Soulard (2010) and in Bemrose, Meszaros and Quenneville (2011). This paper won Statistics Canada's Tom-Symons Award. The work continued to further develop the method and to incorporate the statistical properties of the linear trend estimate computed on smoothed data (Quenneville, Benmore and Meszaros, 2011).

Methods for short-term trend analysis (also referred to as "real-time" analysis) were studied, especially those related to reproducing kernel Henderson filters. (Bianconcini and Quenneville, 2010).

Other R&D time series projects

Modeling is another area of time series research, and unit-root testing is a core part of identifying ARIMA models. In collaboration with professors from the Chinese University of Hong Kong and the Hangzhou Normal University, work continued on the theoretical distribution – exact, approximate, and asymptotic – of the sample variogram to obtain a test for unit root (Chen, Wu and Wang, 2011).

For further information, contact:

Susie Fortier (613-951-4751, susie.fortier@statcan.gc.ca).

Reference

Wyman, D. (2010). Seasonal adjustment and identifying economic trends. Statistics Canada: Canadian Economic Observer, March 2010, Catalogue no. 11-010-X.



Top of Page

Record linkage resource centre

The objectives of the Record Linkage Resource Centre (RLRC) are the following:

- To provide consulting services to both internal and external users of record linkage methods, including recommendations on software and methodology and collaborative work on record linkage applications.
- To evaluate alternative record linkage methods and develop improved methods.
- To evaluate software packages for record linkage and, where necessary, develop prototype versions of software incorporating methods not available in existing packages.
- To assist in the dissemination of information concerning record linkage methods, software and applications to interested persons both within and outside Statistics Canada.

Below is a list of our activities in 2010-2011

- Continued to provide support work to the development team of G-Link, including participating in Record Linkage User Group meetings.
- Provided various consulting services to internal users of record linkage methods.
- Consultation services (steps to access data for record linkage, provision of documentation and information on RL policy), services associated with the probabilities calculated using G-Link to select better rules and thresholds.
- Sharing pre-processing macros for cleaning up much of the data (treatment of dates, standardization and parsing names and surnames, validity of identifiers (SIN, HIN)).
- Presentation of a seminar on the Record Linkage Resource Centre (RLRC) and all the resources available to methodologists and external users in Methodology Branch's Learning Week.
- Preparations for the presentation of an article by Winkler (2009) entitled "Record linkage" for Methodology Branch's article-of-the-month program. The focus will be on the following topics: the Fellegi-Sunter theoretical approach, the generalized EM algorithm approach, string comparators (Jaro-Winkler), the importance of pre-processing steps, estimation of the error rate, and adjustment of the linear regression analysis using matched data.
- An update of Methodology Branch's list of probabilistic linkage projects.
- A study of the linkage error rate due to false positives and false negatives under the usual probabilistic matching assumptions to automatically generate the thresholds for charts of observed error rates. For approaches using conditional weights, certain tools must first be developed in G-Link when frequency weights are used. A project is under way to create that functionality.

For further information, contact:

Abdelnasser Saidi (613-951-0328, abdelnasser.saidi@statcan.gc.ca).

Data analysis resource centre (DARC) support

The Data Analysis Resource Centre (DARC) is a centre of data analysis expertise within Statistics Canada whose purpose is to encourage, suggest and provide good analytical methods and tools for use with Statistics Canada data. While DARC gets funding for its support activities from a number of sources, this report will concentrate mainly on work charged to the Methodology Block Fund. DARC also uses a portion of its Methodology Block Fund support activities for doing research. Those research activities are also partially funded by Data Analysis Research (1919) and are described in the report for that project.

Consultations

In 2010/11, DARC did a great number of consultations on statistical methods with analysts from many areas within Statistics Canada. As well as giving advice on a methodological approach, many consultations also included support for determining and using suitable software for an analysis. Many of the problems were related to the more traditional analytical approaches with survey data, such as estimation and inference involving descriptive statistics and fitting of linear and logistic models to survey data. However, other consultations involved extensions of the traditional methods and newer, more complex issues, such as accounting for the correlation of repeated measures, the use of hierarchical models with survey data and the integration of data from more than one survey into a single analysis. Some consultations were not based on survey data, such as the estimation of confidence intervals for ratios of standardized rates estimated from administrative sources when the counts are small.

Again this year, there was a considerable consultation with other methodologists who were either doing analysis themselves or were supporting subject-matter analysts of the survey data for which they were responsible. One example of this was the support to Canadian Health Measures Survey (CHMS) methodologists in determining how to modify analytical methods to account for the limited degrees of freedom for variance estimation, and how joint research might be done with CHMS and the American survey NHANES. Another example is the support given for an analysis of data from the 2007 Survey of Drinking Water Plants; this resulted in a coauthorship on a paper in *Environmetrics* (See Nelligan, Wirth, De Cuyper and Mach 2011).

Several consultations also took place with people external to both Statistics Canada and the Research Data Centre (RDC) network. One consultation was with researchers from the Institute for Clinical Evaluative Sciences, wishing assistance with centering and standardization and with calculating survey bootstrap variance estimates. Another consultation involved assistance to a student doing research with a survey PUMF file.

The consultation function also included technical reviewing and refereeing of articles for both internal and external journals, including *Environmetrics*, *Canadian Social Trends*, and the proceedings of the Survey Methods Section of the Statistics Society of Canada (SSC) and of the Joint Statistical Meetings.

Provision of training

During 2010/11, DARC was responsible for presenting Course 0438, Statistical Analysis of Survey Data. Both parts (0438A and 0438B) were presented in both English and French. Additional material was also developed. Some methodologists outside of DARC were involved in the teaching.

DARC members were involved in the brainstorming sessions for each session of the Data Interpretation Workshop, in the project reviews and in presenting a seminar on the analysis of survey data. There has been follow-up with some of the participants who wanted further assistance with their analysis projects.

DARC presented a seminar in the seminar series aimed at new recruits on the topic of analysis of survey data.

Several "brownbag" telephone seminars were provided to RDC analysts on analytical topics of interest to them. An overview seminar was also given at the COOL RDC. While most of the preparation time for these seminars was covered by budget from the RDCs, a small amount is charged to Methodology Research and Development Program (MRDC) since the resulting material is useful outside of the RDCs.

Software evaluation, development and promotion

During 2010/11, DARC continued to examine various commercial software packages with respect to their suitability for analyzing Statistics Canada survey data. One activity was keeping abreast of new features in software that can make use of survey bootstrap weights for design-based analysis (e.g., SUDAAN 10, Stata 11 and SAS 9.2). Feedback was provided to Stata Corp on the new bootstrapping features they were releasing. As well, a "how-to" document was prepared to introduce new researchers to the implementation of weighted estimation and bootstrap variance estimation in a number of different software packages (See Gagné, Roberts and Keown 2011). Such a document was requested by RDC analysts (and its preparation was mainly covered by

DARC support resources), but it has been made available for evaluation, in draft form, to a number of other researchers.

A seminar called "Introduction to SAS-callable SUDAAN" was presented to a division in Statistics Canada that had newly gained access to SUDAAN.

Work was done on updating our knowledge of diagnostics available in survey analysis software related to logistic regression of survey data. These findings will be incorporated into our training and written material.

DARC spent a little time on supporting the use of Stata and SUDAAN at Statistics Canada.

Other activities

DARC resources were used to support one of the PhD students at Statistics Canada on NICDS/MITACS internships.

DARC members were involved in a variety of organizing functions for the program of the Survey Methods Section of the SSC in 2010 and 2011 and for the Methodology Interchange of Statistics Canada and the U.S. Census Bureau held in March 2011.

Lenka Mach and Karla Fox organized a case study entitled "Gender Gap in Earnings among Young People" for the 2011 annual meeting of the SSC. This involved finding appropriate data, developing research questions and writing background material for the students. Groups of students registered for the case study will compete for an award.

Georgia Roberts sat on the editorial boards of *Health Reports* and *Canadian Social Trends*.

For further information, contact:

Georgia Roberts (613-951-1471, georgia.roberts@statcan.gc.ca).

Generalized systems

G-Confid

G-Confid (formerly known as Confid2) is the generalized software used to suppress sensitive cells in a data release. It preserves the confidentiality of data providers while masking the minimum amount of information. Version 1.0 of Confid2 was released in May 2009, and version 1.02 was released in January 2010. Version 1.03 will be released in April 2011. Enhancements in version 1.03 include the following:

Re-engineering

- Despite being a relatively new system, advances in SAS itself and in programming techniques precipitated a re-engineering of G-Confid resulting in faster run-time.
- Users now have the option to scale the cost coefficient used in cell suppression, another strategy for reducing run-time.

Greater flexibility

- Users can specify aggregates on which the confidentiality will be assessed.
- An additional macro was created to allow users to create tables from the suppression patterns produced by G-Confid.
- An additional macro was created to giving users more information about the aggregates.
- Users can perform G-Confid tasks using SAS Enterprise Guide 4.3.

Banff

The Banff system for edit and imputation was developed to satisfy most of the edit and imputation requirements of economic surveys at Statistics Canada. Banff can be called with programs in SAS, tasks in SAS Enterprise Guide or Metadata with the Banff Processor. The Banff Processor allows users to enter their edits in a spreadsheet format which is then read and parsed into SAS code. The Banff Processor was released internally in August 2009

and user support was provided. The production version of Banff 2.04 (including the Processor) was expected to be released in March 2011. Additional enhancements have been investigated and will be included in the next version of Banff, expected to be released in December 2012. These include:

- Sigma-gap outlier detection method;
- Identification of top contributors as an alternative outlier detection method;
- The ability to keep track of the fields that have already been imputed, so that the user can prioritize these for imputation again if the record still does not pass edits;
- Allowing the user the option to limit the number of times a record is used as a donor.

StatMx

StatMx is a collection of SAS macros that provide additional functions and capabilities beyond those currently available in either the Generalized Sampling System (GSAM) or the Generalized Estimation System (GES).

In April, StatMx 3.6 was released. This version includes updated documentation and support for SAS 9.2.

A literature review on jackknife and bootstrap replication methods for estimating sampling variance was undertaken. Some design issues are still to be resolved, after which a prototype will be built.

Documentation describing the macro parameters and the methodology was completed and made available for the calibration estimation and calibration weights functions. The system documentation for the stratification and allocation functions was also completed.

For further information, contact:

Laurie Reedman (613-951-7301, laurie.reedman@statcan.gc.ca).

Quality assurance

Quality Control Data Analysis System (QCDAS)

QCDAS is SAS based with a menu-driven user interface. It is used to analyze, estimate and tabulate summary statistics for quality control programs. The system is also used to generate reports to reflect the analysis. The system is performing well in production, allowing the team to discover more functionality to add. These include the addition of several reports allowing the users to interpret data capture quality at the operator or the survey level.

Generic intelligent character recognition (ICR)

ICR is the technology used for data capture, which makes use of a combination of automated machine capture (using optical character, mark and image recognition), along with the manual heads-up capture of data by operators. The software used is called "AnyDoc". BSMD has developed and implemented a generic Quality Control for this ICR software for the document preparation, scanning, automated data entry (ADE) and manual key-from-image (KFI) stages of data processing. A simulation study was designed and executed to use the software to not only detect the presence of data on a questionnaire but to capture that data as well. The study involved manually capturing data from a set of test questionnaires, then allowing the software to capture that same data, using different error tolerance levels that can be set as parameters in the software itself. The results from the first type of questionnaire studied were promising. Next, more tests will be performed on different types of questionnaires. If successful, this will improve the efficiency of the capture process by reducing the amount of data that has to be captured manually by an operator, while maintaining the high accuracy rate of the captured data.

Automated coding, parsing and scoring methodology (G-Code)

G-Code (formerly known as ACTR) is used to automatically assign predefined codes to responses to open-ended questions. This is done in two steps. In the first step, both the input and reference text are parsed by a user-defined parsing strategy in order to reduce the text to a standard form. Parsing deals with problems such as common spelling variations, abbreviations, etc. The parsing strategy plays a strong role in determining the success rate of the coding process. The second step is to match the parsed input text to a list of parsed descriptions in a reference file and assign the associated code when a match is successful. Direct or indirect

matching can be done. For indirect matching, a weight is assigned to each matching word in the input phrase and a score for this phrase is computed based on weights and the number of words in common between the input and the reference descriptions.

Research to improve indirect matching was continued using the Levenshtein algorithm. Simulation studies demonstrated that the new strategy is complementary to the one currently in the G-Code software. Users can expect to achieve more matches if they use one strategy followed by the other rather than using only one of the two strategies. The new strategy will be included in a new version of the G-Code software, and the interface is being enhanced to allow the users more flexibility and additional functionality.

Methodology acted as a consultant to the Standards Division regarding which parsing strategy to implement in the new generalized interactive coding tool to be launched in April 2011.

For further information, contact:

Laurie Reedman (613-951-7301, laurie.reedman@statcan.gc.ca).

Statistical training

The Statistical Training Committee (CFS) coordinates the development and delivery of 24 regularly scheduled courses in survey methods, sampling theory and practice, questionnaire design, time series methods and statistical methods for data analysis. During the year, a record number of 31 scheduled sessions (104 days of training) were given in either English or French.

The suite of courses continues to expand as 4 new courses were given for the first time in 2010-2011

0437: Theory and Application of Raking

0448: Calibration and balanced sampling in surveys

0488: Variance estimation: Theory and Practice

0495: Using Paradata in Applied Operational and Methodological Research for Social Surveys

In addition, three other new courses are currently being developed:

0494: Introduction to Longitudinal Survey Data Analysis

0460: Methodological Issues for New Data Collection Methods

0492: Automated Coding.

For more information, contact:

François Gagnon (613-951-1463, francois.gagnon@statcan.gc.ca).

Conferences

2009 Symposium:

The 25th International Methodology Symposium was held on October 27 to 30, 2009 at the Palais des congrès in Gatineau under the theme of "Longitudinal Surveys: From Design to Analysis". The individual papers for the 2009 Symposium proceedings were translated and verified. Final formatting, approval by the Dissemination Division, loading of the papers onto the Statistics Canada website and creation of a CD-ROM are upcoming.

2010 Symposium:

The 26th International Methodology Symposium was held on October 26 to 29, 2010 at the Crowne Plaza Hotel under the theme of "Social Statistics: The Interplay among Censuses, Surveys and Administrative Data". Close to 530 people were in attendance. **Jelke Bethlehem** was the keynote speaker, and Ivan P. Fellegi was the Waksberg Award speaker. Three workshops were also given on the first day of the symposium:

- Karla Fox and Lori Stratychuk: "*Record Linkage Methods*";

- André Cyr, Julien Bérard-Chagnon, Éric Caron Malenfant, and Dominic Grenier: *"From Traditional Demographic Calculations to Projections by Microsimulations"*;
- Fritz Scheuren and Young Chun: *"Using Administrative/Operating Systems to Strengthen Statistical Survey/Census Systems"*.

In the months leading up to the Symposium, the organizing committee put together a program made up of over 87 presentations from 25 different countries. It also took care of the details surrounding registration, operations and facilities. The proceedings should be disseminated by late 2011.

2011 Symposium:

The preparations for the 27th International Methodology Symposium are well under way. The Symposium will take place from November 1 to 4 at the Ottawa Convention Centre in Ottawa and the theme will be "Strategies for Standardization of Methods and Tools - How to get there". The keynote speaker will be Susan Linacre and Danny Pfefferman will be the Waksberg Award speaker. The Symposium will be preceded by a day of workshops.

For further information, contact:

Colin Babyak (613-951-2045, colin.babyak@statcan.gc.ca).

Survey methodology

Survey Methodology (SM) is an international journal 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. It is one of just two major journals in the world dealing with methodology for official statistics.

The June 2010 issue, SM 36-1, was released on June 29th, 2010. The issue contains 10 papers.

The December 2010 issue, SM 36-2, was released on December 21, 2010. It contains 11 papers including the tenth in the annual invited paper series in honour of Joseph Waksberg. The recipient of the 2010 Waksberg Award is Ivan Fellegi.

In 2010, the journal received 57 submissions of papers by various authors.

The available on-line electronic versions of SM now range from 36-2 (December 2010) to 25-1 (June 1999). Electronic index and abstracts are also available up to 22-2 (December 1996). The conversion of back issues to an electronic version continues.

During the review period, the editorial board and the assistant editors worked to standardize, streamline and document the editorial process including the various contacts with the authors and the Associate editors.

For more information, contact:

Susie Fortier (613-951-4751, susie.fortier@statcan.gc.ca).

Statistics Canada

Research papers and presentations sponsored by MRDC

Beaumont, J.-F. (2011). How to evaluate the impact of imputation on data quality? Paper written in the SIMP II - Economic Statistics Document Series, Document #48.

Beaumont, J.-F., and Bissonnette, J. (2011). Variance estimation under composite imputation: The methodology behind SEVANI. *Survey Methodology* (to appear).

Beaumont, J.-F., and Bocci, C. (2011). Some Theory on Synthetic Data Generation and its Application to the Canadian Portion of the Cross National Equivalent File. Internal report (draft).

Beaumont, J.-F., and Charest, A.-S. (2011). Bootstrap variance estimation with survey data when estimating model parameters. Paper under revision in *Computational Statistics and Data Analysis*.

Beaumont, J.-F., and Patak, Z. (2010). Generalized Bootstrap in Prices Surveys. Paper presented at the Advisory Committee on Statistical Methods, April 2010.

Beaumont, J.-F., and Patak, Z. (2011). On the generalized bootstrap for sample surveys with special attention to Poisson sampling. Paper under revision in the *International Statistical Review*.

Beaumont, J.-F., Haziza, D. and Bocci, C. (2011). On variance estimation under auxiliary value imputation in sample surveys. *Statistica Sinica*, 21 (to appear).

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2011). A unified approach to robust estimation in finite population sampling. Paper submitted to a refereed journal.

Bélanger, Y., Couture, K. and Neusy, E. (2011). An application of SAS[®] Simulation Studio: The Microsimulation of a Computer Assisted Telephone Interviewing System.

Bemrose, R., Meszaros, P. and Quenneville, B. (2011). Trend-cycle approach to estimate changes in Southern Canada's water yield. *Estudios de Economía Aplicada*, 28, 3, 595-606.

Bemrose, R., Meszaros, P., Quenneville, B., Henry, M., Kemp, L. and Soulard, F. (2010). Using a trend-cycle approach to estimate changes in Southern Canada's water yield from 1971 to 2004. Statistics Canada, Environments Accounts and Statistics Division, Catalogue no. 16-001-M, no. 14. Won Statistics Canada Tom-Symons Award.

Bérubé, J., and Fortier, S. (2010). PROC TSRAKING: An in-house SAS[®] procedure for balancing time series. *Proceedings of the SAS Global Forum*, April 2010.

Bianconcini, S., and Quenneville, B. (2010). Real time analysis based on reproducing kernel Henderson filters. *Estudios de Economía Aplicada*, Volume 28, No. 3, 533-574.

Binder, D. (2010). New developments in modelling using survey data- a discussion. *Proceedings of the Joint Statistical Meetings*.

Binder, D. (2011a). Inferring Causality from Survey Data: Summary of Findings. Unpublished Manuscript.

- Binder, D. (2011b). Estimating model parameters from a complex survey under a model-design randomization framework. To be submitted to *Pakistan Journal of Statistics*.
- Carpenter, R. (2011). Methodology Software Library StatMx System Documentation version 1.0 Description of Macro Parameters. Internal document.
- Carrillo-Garcia, I., Kovačević, M. and Wu, C. (2010). A Comparison of Reweighting and Imputation Approaches for Missing Responses in Longitudinal Surveys. Presentation in a Contributed Session at the Annual Meeting of the SSC.
- Chen, Z.G., and Wu, K.H. (2010). Benchmark Forecast and Error Modelling. Presented at the 4th Annual International Conference on Mathematics & Statistics, Athens, Greece, June 2010.
- Chen, Z.G., and Wu, K.H. (2010a). Benchmarking with working error-model. *Advances and Applications in Statistics*, Volume 19, No. 2, 113-148.
- Chen, Z.G., and Wu, K.H. (2010b). Benchmark forecast and error modeling. Presented at 4th Annual International Conference in Mathematics and Statistics, in Athens, Greece (also submitted to a referred publication).
- Chen, Z.G., and Wu, K.H. (2011). Benchmarking using working error-model. *Advances and Applications in Statistics*, 19, 2, 113-148.
- Chen, Z.G., Wu, K.H. and Wang, W. (2011). On the variogram for unit-root testing (working title). Working paper (in progress).
- Choudhry, G.H., Hidirolou, M.A. and Laflamme, F. (2011). Optimizing CATI call scheduling to minimize data collection costs. To be presented at Joint Statistical Meetings 2011.
- Choudhry, G.H., Rao, J.N.K. and Hidirolou, M.A. (2010). On sample size allocation for planned small domains. Submitted to *Survey Methodology*.
- Couture, K., Bélanger, Y. and Neusy, E. (2010). Modelling and Simulation of Survey Collection Using Paradata.
- Davis, K. (2010). Automated Data Entry (ADE) Recognition Confidence Threshold Study (2010). Internal presentation.
- Demers, G., et Godbout, S. (2011). Généralisation de la méthode de linéarisation de Taylor pour l'estimation de la variance totale. Study report.
- Dongmo Jiongo, V., Haziza, D. and Duchesne, P. (2010). Robust Small Area Estimation. Presentation in a Contributed Paper Session at the 2010 SSC Annual Meeting.
- Estevao, V. (2010). Methodology Software Library StatMx Calibration Estimation Function version 2.1 Description of Function Parameters. Internal document.
- Estevao, V. (2010). Methodology Software Library StatMx Calibration Weights Function version 2.1 Description of Function Parameters. Internal document.
- Estevao, V. (2011). Methodology Software Library StatMx Calibration Weights and Calibration Estimation version 2.1 Description of Methodology (2011). Internal document.
- Fortier, S., Quenneville, B. and Picard, F. (2010). Calendarization method and applications. Presented to the ACSM, October 2010.
- Fox, K. (2010). A framework for the meta-analysis of survey data. *Joint Statistical Meetings* Vancouver, invited presentation and proceedings paper.
- Fox, K. (2011). The superpopulation, the finite population and meta-analysis. Unpublished working manuscript.
- Gagné, C. (2011). *An empirical analysis of disclosure control methods*. Unpublished technical paper.
- Gagné, C., Roberts, G. and Keown, L.A. (2011). Weighted Estimation and Bootstrap Variance Estimation for Analysing Survey Data: How to Implement in Selected Software. Draft manuscript.

- Grondin, C., Haddou, M. and Laperrière, C. (2010). Research project on Estimating Disclosure Risk for Public Use Microdata Files (PUMF). SSMD Technical Report, April, 2010.
- Hidiroglou, M.A. (2010). Current Developments in Small Area Estimation at Statistics Canada. Invited Paper Presented at *Joint Statistical Meetings 2010*, Meetings held In Vancouver.
- Jamroz, E. (2010). Report on Small Area Estimation Using a Linear Mixed Effects Model. Internal SRID report.
- Julien, C., and Fortier, S (2011). Seasonal Adjustment Horizontal Review. Internal report (in progress).
- Kovačević, M., Roberts, G. and Mach, L. (2010). Bootstrap variance estimation for population risk predictions using complex survey samples. Submitted to *Survey Methodology*.
- Lavallée, P., et Labelle-Blanchet, S. (2011). *Indirect Sampling Applied to Skewed Populations*. Article in preparation.
- León, C.A. (2011). *Optimally controlled sampling designs: to guard you against bad samples*. Presentation for the SSC's annual meeting.
- Mach, M., Beaumont, J.-F., Bosa, K. and Lafortune, Y. (2010). Mean Bootstrap Variance Estimation for Survey Data. Presentation at the 2010 meeting of the Statistical Society of Canada, Québec, June 2010.
- Mazzi, G.L., Fortier, S. and Quenneville, B. (2010). Multivariate Benchmarking - The Direct Versus Indirect Problem. Paper written for a chapter in a Handbook on seasonal adjustment to be published by Eurostat.
- Michaud, I., Singh, A. and Tambay, J.-L. (2010). MASSC-Basic for Statistical Disclosure Limitation of Large Administrative Databases. DRAFT - Technical Report, SSMD.
- Nelligan, T., Wirth, S., De Cuypere, C. and Mach, L. (2011). **Operation and Maintenance Costs of Drinking Water Plants. Envirostats, Spring 2011**. Cat.No. 16-002-X. /pub/16-002-x/2011001/part-partie3-eng.htm.
- Oyarzun, J., and Nambeu, C.O. (2011). Calcul de la variance due à l'imputation avec l'utilisation de méthodes de rééchantillonnage. Internal report, Statistics Canada.
- Quenneville, B. (2010). Gain and Phase Shift for the Annual and Monthly Differences. Working paper.
- Quenneville, B., and Fortier, S. (2010). Benchmarking and Temporal Consistency. Paper written for a chapter in a Handbook on seasonal adjustment to be published by Eurostat.
- Quenneville, B., and Fortier, S. (2011). Restoring Accounting Constraints in Time Series - Methods and Software for a Statistical Agency. Sent for possible publication in an edited volume in honour of David Findley to be published by Chapman & Hall/CRC.
- Quenneville, B., and Gagné, C. (2010). Testing Time Series Data Compatibility for Benchmarking. Presented at ACSM April 2010; presented at the 7th International Institute of Forecasters' Workshop, Swiss, January 2011; sent for possible publication.
- Quenneville, B., and Gagné, C. (2011). Testing Time Series Data Compatibility for Benchmarking. Presented at ACSM April 2010; presented at the 7th International Institute of Forecasters' Workshop, Swiss, January 2011. Sent for possible publication in a special issue of the International Journal of Forecasting.
- Quenneville, B., Benmore, R. and Meszaros, P. (2011). Presenting and Estimating Global and Local Trend in Environmental Data. Working paper (in progress).
- Quenneville, B., Picard, F. and Fortier, S. (2010). Interpolation, benchmarking and temporal distribution with natural splines. *ASA Proceedings of the Business and Economic Section*, American Statistical Association, and also submitted for possible publication.
- Rao, J.N.K., Hidiroglou, M.A., Kovačević, M. and Yung, W. (2010). Role of weights in descriptive and analytical inferences from survey data: An overview. *Journal of Indian Society of Agricultural Statistics*, 64, 2, 129-135.
- Rao, J.N.K., Verret, F. and Hidiroglou, M.A. (2010). A Weighted Estimating Equations Approach to Inference for Two-Level Models from Survey Data. SSC Annual Meeting, *Proceedings of the Survey Methods Section*, May 2010.

- Rao, J.N.K., Verret, F. and Hidirolou, M.A. (2010). A weighted estimating equations approach to inference in two-level models from survey data. *Proceedings of the Survey methods section, Statistical Society of Canada*.
- Roberts, G. (2010). New Developments in Modelling using Survey Data - A Discussion. Discussion of papers presented at the 2010 *SSC* annual meeting.
- Rubin-Bleuer, S. (2010). The proportional hazards model for survey data from independent and clustered super-populations (matériel additionnel). SRID-2010-004-E.
- Rubin-Bleuer, S. (2011). The proportional hazards model for survey data from independent and clustered super-populations. *Journal of Multivariate Analysis*, 102, 884-895.
- Rubin-Bleuer, S., Yung, W. and Landry, S. (2010). Adjusted maximum likelihood method for a small area model accounting for time and area effects. SRID-2010-005-E.
- Schiopu-Kratina, I., Fillion, J.-M., Mach, L. and Reiss, Ph.T. (2011). *Maximizing the conditional overlap in business surveys*. Paper in preparation.
- Théberge, A. (2011). Calage et estimation de la variance. Article in preparation.
- Trépanier, J., and Beaumont, J.-F. (2010). Strategic Thinking Process of the Methodology Branch. Paper presented at the Advisory Committee on Statistical Methods, October 2010.
- TSRAC (2010). ISB - Time Series Processing System: Input/output specifications. Technical working paper.
- Verret, F., and Kevins, C. (2010). Calage aux marges des poids d'enquêtes à poids complexe pour le refus à l'appariement. *Proceedings of the Survey Methods Section, Statistical Society of Canada*.
- Verret, F., Hidirolou, M.A. and Rao, J.N.K. (2010). Small area estimation under informative sampling. *Proceedings of the Survey methods section, Statistical Society of Canada*.
- Verret, F., Hidirolou, M.A. and Rao, J.N.K. (2010). Small area estimation under informative sampling. *SSC Annual Meeting, Proceedings of the Survey Methods Section, Statistical Society of Canada*, May 2010.
- Xu, C., Chen, J. and Mantel, H. (2010). **Penalized Likelihood Methods for Variable Selection in Analysis of Survey Data. Proceedings of the Survey Methods Section, Statistical Society of Canada.**
- Xu, C., Chen, J. and Mantel, H. (2010). Smoothly Clipped Absolute Deviation in Analysis of Survey Data. Presented at 2010 *SSC* meetings, published in the *proceedings of the Survey Methods Section of the SSC*.
- Yang, X., and Pelletier, C. (2010). Improving the Quality of Automated Coding at Statistics Canada. *Proceedings of the Survey Methods Section, Statistical Society of Canada*.
- You, Y. (2010). Small area estimation under the Fay-Herriot model using different model variance estimation methods and different input sampling variances. Methodology branch working paper, SRID-2010-003E, Statistics Canada, Ottawa, Canada.
- You, Y. (2011). Hierarchical Bayes sampling variance modeling for small area estimation based on area level models. Internal Manuscript.
- You, Y., and Zhou, Q.M. (2011). Hierarchical Bayes small area estimation under a spatial model with application to health survey data. *Survey Methodology*, 37, 1, 25-37.
- You, Y., Rao, J.N.K. and Hidirolou, M.A. (2011). Benchmarking small area estimators under the Fay-Herriot model. Draft paper.
- Yung, W., Rubin-Bleuer, S. and Landry, S. (2010). Small Area Estimation for Business Surveys. *Proceedings of the Survey Research Methods Section, American Statistical Association, Joint Statistical Meetings*.