

Statistical Methodology Research and Development Program Achievements

Statistical Methodology Research and Development Program 2008-2009 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, 2009

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.



Statistical Methodology Research and Development Program 2008-2009 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 achievements during 2008-2009 in 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 utilisation 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).

Research Projects

Imputation and Robust Estimation

Progress:

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. This year, a working paper has been peer-reviewed, revised and completed (Mantel and Hidirolou, 2008). It summarizes the literature on this topic and the main issues that need to be considered when creating synthetic data.

We have also developed a simple method of variance estimation with synthetic data that only needs the creation of a single synthetic data set. This is to be contrasted with current methods that require more than one synthetic data set. We have described the theory in a preliminary report (Beaumont and Bocci, 2009a). The paper needs to be completed with empirical results.

Finally, we have just started looking into the use of these methods and creating a first synthetic data set using a subset of variables of the Survey of Labour and Income Dynamics. This subset forms the Canadian portion of the Cross-National Equivalent File (CNEF) in which five other countries are involved. We have synthesized about 15 variables so far trying to preserve as much as possible between-variables and within-households relationships using classical modeling tools such as stepwise linear and logistic regressions. Our preliminary results indicate that the synthetic data set has preserved with reasonable accuracy these relationships. This work will be presented at the next interchange with the U.S. Census Bureau in April 2009 and at the 2009 Symposium.

SEVANI and research on estimation of variance with imputation

SEVANI is the System for Estimation of Variance due to Non-response and Imputation. We programmed and tested version 2.1 of the system and programmed a preliminary version 2.2. We revised the user guide and created a complete methodology guide. We have also updated the intranet site and continued to provide support to users. Note that the New Zealand, Quebec and Australia statistical agencies have received copies of the system, which has allowed us to undertake collaborative projects with the first two.

We developed a variance estimation method that accounts for imputation with adjustment variables. For example, when using donor imputation for a given variable y , it is possible to either impute the donor's value y directly or impute the donor's y/z ratio, which is then multiply by the receiver's value z , variable z being the adjustment variable. This type of imputation is often used in business surveys. The methodology was implemented in SEVANI 2.1 and tested. This methodology was developed due to the use of donor imputation with adjustment variables to deal with missing values in tax data.

We developed a new methodology to better process imputations based on hierarchical classes. The methodology was implemented in SEVANI 2.2, but must be tested thoroughly and possibly refined. Version 2.2 is not yet available.

We continued our research on estimating variance with composite imputation. The proposed methodology is an extension of the Särndal method (1992, *Survey Methodology*) and is programmed in SEVANI 2.0. We conducted a simulation test that confirmed the method's validity. We still have to describe this simulation and submit the article for possible publication in a journal. We also showed through simulation that total variance with imputation can be smaller than the sample variance, particularly in the case of domain totals estimation.

We revised a research article on estimation of variance with donor imputation (Beaumont and Bocci, 2009b). The proposed methodology is programmed in SEVANI version 2.0.

System for outlier detection

We continued developing a Statistical Analysis System (SAS) system to detect outliers to help methodologists do exploratory data analysis for their surveys. This instrument allows them to visualize data and see the impact of choosing one method or another. In addition, we have added certain graphic functions to the system. The methods already programmed are: the Hidiroglou-Berthelot method, the sigma deviation method, the interquartile method, certain classic linear regression methods (with or without origin ordinate), including Cook's distance, and a multivariate method developed for the Wholesale and Retail Trade Survey (monthly). We are currently finalizing implementation of a robust multivariate method based on the Mahalanobis distance. We have developed a macro version to replace the graphic interface version. The Survey on Employment, Payrolls and Hours and the Unified Enterprise Survey have already used the system. We are also planning to use it on T1 data.

Research on handling extreme weights and robust estimation

We have established a link between the influence function and conditional bias in the classic context. Although extending the influence function to the survey is not trivial, the concept of conditional bias is easily transferrable. This provides an influence measure that allowed us to develop a general class of robust estimators for the presence of outliers. Our general estimator essentially boils down to the Chambers estimator (1986, *Journal of the American Statistical Association*) in the theory based on a model and the Kokic and Bell estimator (1994, *Journal of Official Statistics*) for stratified simple random plans. We wrote a preliminary report (Beaumont, Haziza and Ruiz-Gazen, 2009) that should possibly be supplemented by an empiric study.

Nonresponse weighting using paradata

A workshop on nonresponse weighting using paradata has been prepared and given at the 2008 Symposium and two other times internally. It covers standard material but also weight adjustment smoothing as well as some newer calibration methods (Deville, 1998; Kott, 2006; Chang & Kott, 2008) that handle nonresponse mechanisms that depend on the variable subject to nonresponse. The main assumption that is needed for their use is provided, which is not explicitly stated in these papers. Practical issues are also discussed such as: i) weighting when there are both unresolved (we don't know if nonrespondents are in-scope or out-of-scope) and resolved nonrespondents, ii) using or not the design weights when computing the nonresponse adjustment and iii) distinguishing between refusals and non-contacts. Finally, variance estimation is briefly discussed.

Imputation Bulletin

The *Imputation Bulletin* provides a variety of information about imputation theory, software, good practices, recent papers and conferences. Each Bulletin contains three or four short articles on topical issues and is distributed for free to those interested. Two issues of the Bulletin were produced during the year, including a special issue on the link between confidentiality and imputation.

For further information, contact

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

References

Chambers, R.L. (1986). Outlier robust finite population estimation. *Journal of the American Statistical Association*, 81, 1063-1069.

Chang, T., and Kott, P.S. (2008). Using calibration weighting to adjust for nonresponse under a plausible model. *Biometrika*, 95, 555-571.

Deville, J.-C. (1998). La correction de la non-réponse par calage ou par échantillonnage équilibré. *Proceedings of the Survey Methods Section, Statistical Society of Canada, Ottawa*, 103-110.

Kokic, P.N., and Bell, P.A. (1994). Optimal Winsorizing cutoffs for a stratified finite population estimator. *Journal of Official Statistics*, 10, 419-435.

Kott, P.S. (2006). Using calibration weighting to adjust for nonresponse and coverage errors. *Survey Methodology*, 32, 133-142.

Särndal, C.-E. (1992). Methods for estimating the precision of survey estimates when imputation has been used. *Survey Methodology*, 18, 241-252.

Sampling and Estimation

Topic Name: Sample Coordination

Project 1: Optimizing the overlap of samples for business surveys

Progress:

We completed the investigation of our method for optimal sample coordination after re-stratification. There are different SAS procedures that can solve the optimization problem (PROC LP, PROC INTPOINT and the new PROC OPTMODEL) and, while all these solutions are optimal, they are different from each other. We used the example presented in the Third International Conference on Establishment (ICES-III) paper and compared the different solutions by evaluating their conditional properties. We found that conditionally the PROC INTPOINT and PROC OPTMODEL solutions are superior to the PROC LP solution.

For the Northwest Corner Rule (NWCR) solution of the optimal sample coordination problem, we studied its impact on the variance of an estimator of change and found that this variance was minimized when samples were optimally coordinated via NWCR.

We started documenting the results of our research on methods that assure the correct first-order inclusion probabilities for all units in both surveys. These methods are less constrained than our optimal sample coordination method (discussed in the first paragraph) that requires the probability distribution of samples under each of the two survey designs to be achieved. One approach to solving the less constrained problem is using the method of *controlled selection* (first introduced by J.N.K. Rao and Nigam in their 1990 Biometrika article) which seems to be popular recently with other researchers working in the area of sample coordination.

Project 2: Sample coordination for Statistics Canada agricultural surveys

Progress:

We familiarized ourselves with the StatCan agriculture surveys 2006 redesign and the sample coordination methodology that has been used for these surveys. We documented the current coordination methodology and two recommendations in a short document "Coordination of Statistics Canada Livestock and Crop

Surveys" (Mach and Rubin-Bleuer, 2009). We plan to test the two recommendations using the agriculture surveys frame and designs.

Topic Name: Small Area Estimation in the Globalization Project

Progress:

Globalization is a major force shaping the world economies and Canadian businesses. Many policy research programs are working towards understanding how these forces influence Canadian businesses' behaviours and outcomes. However, there is a lack of basic information on the impact of globalization on the Canadian economy. There are few reliable statistics on key economic variables measuring the economic activity of Canadian or foreign controlled multinational firms (Gervais, 2006). A desirable goal would be to obtain these statistics for the major industries. For this purpose an integrated data warehouse was developed, combining the Unified Enterprise Survey (UES, 2008) database with other data sources like the exporter and importer registries produced by the Canadian International Trade Division (ITD, 2007). One of the main goals of the integrated database was to facilitate the production of new statistics such as totals and means at the level of Trade Group by province by international trade characteristics. Since the domain identifiers were not even in the survey frame at the time of the design, the sample is not large enough for many of these domains to yield direct survey estimates with an acceptable precision.

In this particular project we study the feasibility of producing Small Area Estimates (SAE) in the Wholesale Industry using administrative tax data as auxiliary information. This project started in 2007-2008. Area level models were established for each economic variable and Trade Group (based on the North American Industry Classification System). The model observations were the domain means. Last year we proposed a generalization of the Fay-Herriot (1979) model (denoted FH from now on), where the functional form of the direct survey estimate is composed by splines. We also studied the properties of the resulting "penalized" spline estimators and we fitted the data for some of the Trade Groups and economic variables required. This year we developed a theoretical formula for the Mean Squared Error (MSE) of the small area estimators when the variance components are known and a parametric bootstrap estimator of the MSE when the variance components have to be estimated from the data. We presented a paper with some results at the 2008 Statistical Society of Canada (SSC) and a proceedings conference paper has been written and submitted (Dochitoiu and Rubin-Bleuer, 2008).

Topic Name: Bootstrap Estimation of the Mean-Squared Error in Small Area Level Models for Business Data

Progress:

We considered small area estimation using area level models with random area effects and penalized splines in order to accommodate departures from linearity. We developed estimators of the unknown variance components by the method of moments. We adapted several parametric and non-parametric bootstrap methods to estimate the mean squared error of the small area estimators under a Penalized Splines model. We developed a simulation algorithm to compare the various methods and to study their robustness against non-normality of the model errors. We plan to finish writing the program, analyze the output and present this work at the SAE 2009 Satellite International Statistical Institute (ISI) Conference in June 2009 (Rubin-Bleuer and Dochitoiu, 2009).

Topic Name: Robust Small Area Estimation for Business Surveys

Progress:

We propose to study unit level models for business surveys using data from the globalization project. The data contains many outliers and influential points and we rely on robust estimation methodology to tackle this problem. We model the survey observation, e.g., Labour Costs at the establishment level, as a function of the establishment Labour Costs obtained from administrative tax data.

We modeled only the values of the simple-single units and some other units (there are very few) whose weight is larger than one. A preliminary analysis of the data revealed mostly linear relationships between the survey observations and corresponding tax data entries. The outliers we detected were mainly due to inaccuracies in the auxiliary variables that could not be identified from the frame. This could pose a problem for the prediction of the values not in the sample.

We are in the process of adapting the methodology for robust estimation developed by Sinha and Rao (2008). We developed a different Newton-Raphson algorithm to estimate the parameters of the model and we are writing the necessary programs and continuing with the analysis.

Topic Name: Small Area Estimation for the Survey of Employment, Payrolls and Hours (SEPH)

Progress:

The purpose of this project is to develop small area estimators of Average Weekly Earnings (AWE). SEPH is planning to use a composite estimator for large domains that incorporates the previous month estimate. We developed a time series and cross-sectional area level model where the vector of composite estimators at time t and time $t - 1$ is a function of the Average Monthly Earnings (cross-sectional data) and both the sampling errors and the model errors are correlated over time. We plan to analyse the results in terms of design-based average relative bias and average relative root mean squared error. We plan to present this work at an invited session of the Statistical Society of Canada Conference in May 2009 (Yung and Rubin-Bleuer, 2009).

Topic Name: Evaluation and comparison of small area models with spatial correlation

Progress:

This is a project for health statistics small area estimation. Cross-sectional area level models are commonly used in practice to obtain reliable model-based estimates for small areas when direct survey estimates and area level auxiliary variables are available. The classical FH model consists of a linear sampling model for the direct survey estimates with the sampling variance assumed known and a linear linking model that includes a random error to account for the small area effects. We consider here an extension of the FH model, a spatial area level model where the linking model includes spatially correlated random area effects. Estimates are obtained by a fully Hierarchical Bayes (You and Rao, 2002) approach. We will use the FH model as a baseline model in our study. We propose to evaluate several possible models using Bayesian model selection methods such as Deviance Information Criterion (DIC) of Spiegelhalter *et al.* (2002) and other model checking methodology. We highlight the new methodology with an application to the Canadian Community Health Survey (CCHS) data. Last year we developed the spatially correlated models, applied them to health data from the CCHS and wrote a Branch working paper (Zhou and You, 2007). This was the basis of a presentation at the 2008 SSC Meeting this year. A conference proceedings paper has been finished and submitted.

We evaluated the proposed spatial models using the model selection method DIC and the model checking method developed in another small area project, based on the posterior predictive p-values at the overall and individual levels. We submitted research paper based on this work for possible publication in a peer reviewed journal (You and Zhou, 2008).

Topic Name: General area level models for small area estimation with applications and model checking

Progress:

In small area estimation, area level models including the Fay-Herriot model are widely used to obtain efficient model-based estimates for small areas. You and Rao (2002) studied a general linking function to create unmatched sampling and linking models as an extension of the Fay-Herriot model. For the problem of sampling variance, treating the sampling variance known in the model ignores the variability associated with the estimation of the sampling variance. This may in turn lead to underestimation of the errors associated with the small area estimators. You and Chapman (2006) proposed a complete Hierarchical Bayes (HB) modeling approach under the estimated sampling variances and obtained efficient HB small area estimators. In this project, we various area level modeling problems including: (1) generalized area level models with general linking functions and random effects variations for small areas; (2) methods of model evaluation and posterior predictive model checking based on different test quantities for area level models; (3) sampling variance modeling and its sensitivity analysis on model fit and estimates; and (4) estimation of variance components in area level models. Generalized area level models will be developed under a fully hierarchical Bayes approach to include general linking functions for small area parameters and various sampling models for the estimated sampling variances. The proposed model and methodology can be applied generally to many small area estimation problems. The method of benchmarking HB estimators proposed by You, Rao

and Dick (2004) to satisfy higher level constraints will also be studied and included in the model. We have reviewed and studied various area level models for general area level modeling structure and also for the purpose of model checking using posterior predictive p-values. In particular we studied several test quantities through simulation study using the Fay-Herriot model as a studying model to evaluate the effectiveness of model checking. A presentation was made at the 2008 SSC annual meeting in Ottawa. A conference paper (You, 2008) has been finished and published at the SSC website.

Topic Name: Small area disability estimation for the Participation and Activity Limitation Survey (PALS)

Progress:

This project studies model-based methods to improve the small area disability rate and count estimates for the PALS. PALS is a post-censal survey which used the 2006 Census as a sampling frame to identify its population. The 2006 Census questionnaire included two general questions on activity limitations. The PALS respondents were selected through the use of the census information on age, geography and the responses to these two general questions. The PALS interview began with the census disability filter questions identical to the Census questions followed by a series of detailed screening questions on activity limitations. If respondents answered NO to all of the filter questions and screening questions, the interview ended and the respondent was not considered to be a person with a disability according to PALS. If respondents answered YES to any of the filter questions or screening questions, i.e. if they are considered disabled, the interview continued to collect information on the impact of that disability on their everyday activities and other aspects of their life. The PALS sample was 48,000, consisting of approximately 39,000 adults and 9,000 children. The estimates at national and provincial level are reliable. But the survey estimates at the sub-provincial level such as the Census Metropolitan Area (CMA)/Census Agglomeration (CA) level and health region level are not satisfactory due to relatively large **Coefficient of Variations** (CVs), particularly for the estimates of smaller Census Agglomerations (CAs) and/or health regions and for the estimates of children. Model-based estimation methods are to be considered to improve the direct survey estimates. We have studied area level models including the Fay-Herriot model and unmatched log-linear model. In particular, we used weighted Generalized Least Squares (GLS) method to select auxiliary variables in the model. We also used log transformation method and Generalized Variance Functions (GVF) to smooth the sampling variances. Finally we produced improved area level disability estimates for both adults and children at CMA/CA level and health region level. The results have been delivered to the client division. A detailed methodology working paper has been finished (Bizier, You, Veilleux and Grondin, 2008). Some results will be presented at the 2009 Joint Statistical Meetings as a topical contributed paper.

Topic Name: Flexible small area estimation system

Progress:

Small area procedures can be programmed by different users using SAS. However, the development of such custom made programs requires both the knowledge of small area procedures as well as programming. We are therefore developing a Flexible Small Area Estimation (FSAE) system that will start to address the needs of producing small area estimates in production at Statistics Canada. A similar development took place in the early nineties with the Generalized Estimation System (Estevao, Hidirolou, and Särndal 1995). Computer programs developed by the EURAREA project (Heady and Ralphs, 2004) project present an important source of documentation and code. However, this set of programs does not represent the latest advances in small area estimation, and are no longer in development mode. Statistics Netherlands is currently developing a flexible small area system. Some of the basic requirements of a flexible small area system include:

1. Determining the type of procedure used (area level or unit level based),
2. Incorporating the sampling design in the estimation of primary statistics (estimating the parameters of interest) and secondary statistics (mean squared error),
3. Smoothing within and between variance components that are part of the small area procedures,
4. Ensuring that the estimates add up to reliable higher level estimates (i.e. totals), and
5. Developing diagnostic tools to test the adequacy of models.

The statistical theory of model-based SAE is rather complex and the available software at National Statistical Offices is often not suitable to conduct the required calculations in a straightforward manner in a production environment. It is therefore expected that the development of such a system will be both beneficial as a production system, as well as learning tool for SAE procedures. For the FSAE system, we have used the Fay-Herriot model as starting model and implemented different estimation methods such as Restricted Maximum Likelihood Estimation (RMLE), Fay-Herriot Iterative (FHI) and Adjusted Maximum Density (AMD) Method for variance component estimation. We

also consider missing values of small areas as well as different benchmarking methods (Wang, Fuller and Qu, 2008) for model-based estimates. The models and methods are under evaluation using different data sets and a preliminary SAS program has been developed for the FSAE system based on a detailed SAE implementation internal notes (Hidioglou, You, Patak and Estevao, 2009 and Rubin-Bleuer, 2008).

Subject: Use of the Business Register's Contact Archive and Tracking System (CATS) in sample selection

Progress:

Part A: Response probability modeling

From the Business Register (BR) response burden file, we created a historical response file. For each collection entity, various binary variables summarizing the results of contacts at the collection entity and enterprise levels in 2005 and 2006 (number of contacts, response status, non-response type, *etc.*) were derived. Those binary variables are used as auxiliary variables to model a collection entity's response probability.

We then used the UES collection entity response codes file for reference year 2006 to create a response variable with a corresponding value of 1, if the entity was considered respondent, or 0, if otherwise. We matched that file to the historical response file. Other binary auxiliary variables were also added based on the information available at the start of the 2006 collection cycle (collection mode and priority, regional office, *etc.*). We tried to estimate the relationship between the response variable and the binary auxiliary variables using a logistical model. The first modeled estimates show encouraging potential results.

We still have to complete the modeling exercise, check the results and describe the main non-response factors according to historical information available.

Part B: Inclusion of response probability in the sample plan

In parallel with the modeling work, we started from an already existing algorithm to create a synthetic population from given assumptions, draw independent random samples, produce estimates and calculate accuracy measures using the Monte Carlo method (bias, standard error, mean square error). We adapted that algorithm to be able to generate response probabilities by various types of non-response (none, negligible and non-negligible) and to adjust for non-response by the inverse of the estimated response probability.

We are going to continue changing the algorithm to consider response probability in the sampling method. Then, for each type of non-response, we are going to produce different sampling and estimation scenarios and compare them in terms of bias and mean square error on estimates of interest.

Topic Name: New Applications of Estimating Functions for Point, Variance, and interval estimation for moderate samples

Progress:

In modeling outcome prevalence such as the healthy life-style behavior outcomes from the Canadian Community Health Survey, logit models are commonly used. For this purpose, the quasi-likelihood Estimating Function (EF) appropriately weighted for survey data can be used for estimating model parameters. For interval estimation of prevalence, asymmetric Confidence Interval (CI) is generally preferable due to marked skewness in the distribution of estimates of low or high prevalence outcomes. This can be obtained by using the logit-Wald in which the normal CI is first constructed for the logit-linear predictor and then transformed back to the original proportion scale. Although this method is commonly used in practice, it is known to have instability in the variance estimate for moderate sample sizes and/or for low or high prevalence outcomes. Consequently, the confidence limits may have poor coverage properties. In this project, we consider an alternative based on the Randomly Re-centered Estimating Equations (RREE) approach of Singh (2007) to address the above concerns. In the RREE method, we construct an approximately normal pivotal based on the EF. Replicate parameter estimates are created by equating the standardized EF to random values drawn from the standard normal distribution. Those replicates are used to compute new variance and interval estimates. Unlike the previous applications of RREE reported in Singh (2007), the EF-based pivotal for this

application is a vector and so the empirical distribution of the parameter estimator is multivariate. This may pose computational difficulties for obtaining replicate estimates. A simulation study was completed to evaluate the performance of the proposed method in comparison to logit-Wald in terms of bias and variance of the variance estimator, and average length and coverage probability of the CI. The performance of the RREE method was found to be quite favorable as expected from earlier studies. The results were presented by Claude Nadeau at the Joint Statistical Meetings 2008; the title of the presentation was "An Alternative to the Logit-Wald Method for Inference with Models for Proportions". A paper for the JSM proceedings was written (Singh and Nadeau, 2008).

Topic Name: Optimal instrument variables for estimating spatial modeling parameters

Progress:

The problem of using instrument variables in estimating regression and correlation parameters arises in the context of spatial modeling under a quasi-likelihood framework. The reason for this is that in the joint modeling of spatial dependence and regression, the dependent variables also appear as regressors. Existing solutions (such as two stage least squares) to this problem do not use optimal Instrument Variables-IVs. The main purpose of this project is to study how optimal instruments can be defined using estimating functions (Singh and Rao, 1997) and consider efficiency gains. Applications of the methodology to data from the Statistics Canada crime survey will be considered. A simulation study based on the violent crime data was completed to compare the optimal IV method with a non-optimal IV method Two Stage Least Squares (2SLS), and maximum likelihood under assumed normality. The optimal IV method was found to perform very well in terms of reducing bias in the estimated spatial lag correlation parameter. Results were presented by Krista Collins at the JSM '08. A paper was written for JSM proceedings (Collings and Singh, 2008).

Topic Name: Single phase variance estimation approach to two phase designs

Progress:

Two phase designs are useful for increasing estimation efficiency for reasons including deep stratification, improved non-response adjustment and calibration controls, and for integrating survey designs by using the first phase as a master sample. Despite these benefits, such designs are not in wide use due to difficulty in variance estimation. This is in contrast to single phase multi-stage designs where the With Replacement Primary Sampling Unit (wrpsu) assumption is often invoked for simplified variance estimation. In this project, we consider a simple modification of two phase designs such that the second phase sampling within each first phase cluster/Primary Sampling Unit (psu) is performed Probability Proportional to Size (pps) independently across psu's and conditional on the first phase sample. It turns out that this modification leads to the usual simplified variance estimation. Here the size measures for the pps at the second phase correspond to the sampling rates determined for the second phase stratification. The main theoretical results were obtained and were presented at the JSM '08 by Harold Mantel. The paper was well received. A draft of the paper was written for JSM '08 proceedings (Singh, 2008).

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

References

Estevao, V., Hidirolou, M.A. and Särndal, C.-E. (1995). Methodological principles for a generalized estimation system at Statistics Canada. *Journal of Official Statistics*, Vol. 11, No. 2, 181-204.

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.

Gervais, Y. (2006). *Proposal Summary: Globalization Project (A-530-01)* (Internal Document, Business and Trade Statistics (BTS), Statistics Canada).

Heady, P., and Ralphs, M. (2004). Some findings of the Eurarea Project – and their implications for statistical policy. *Statistics in Transition*, April, Vol. 6, No. 5, 641-653.

ITD (2007). http://icn-ci.statcan.ca/fpt/fpt02/fpt02_cbs_200710min-eng.htm.

Rao, J.N.K., and Nigam, A.K. (1990). Optimal controlled sampling design. *Biometrika*, 77, 807-814.

Singh, A.C. (2007). A new application of estimating functions to point, variance, and interval estimation for simple and complex surveys. *Federal Committee on Statistical methods Conference proceedings*, Washington, D.C. (www.fcsm.gov).

Singh, A.C., and Rao, R.P. (1997). Optimal Instrumental Variable Estimation for Linear Models with Stochastic Regressors Using Estimating Functions. In Basawa, I.V., V.P.

Sinha, S., and Rao, J.N.K. (2008). Robust small area estimation. Carleton University Internal Report.

Spiegelhalter, D.J., Best, N., Carlin, B.P. and van de Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society*, B, 64, 583-639.

Unified Enterprise Survey (2008). www.statcan.ca/bsolc/english/bsolc?catno=68F001X.

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). Small area estimation using unmatched sampling and linking models. *The Canadian Journal of Statistics*, 30, 3-15.

You, Y., Rao, J.N.K. and Dick, P. (2004). Benchmarking Hierarchical Bayes small area estimators in the Canadian census undercoverage estimation. *Statistics in Transition*, April, Vol. 6, No. 5, 631-640.

Zhou, Q.M., and You, Y. (2007). Hierarchical Bayes small area estimation for the Canadian Community Health Survey. Methodology Branch Working Paper, HSMD-2007-007E, Statistics Canada.

Estimates for 3 Digit Occupations for the Labour Force Survey (LFS)

Topic Name: 3-digit Occupation Estimates for LFS

Progress:

Statistics Canada is working closely with Prof. Fuller and Prof. Rao on the theory and application of SAE techniques to the LFS. To produce reliable SAE estimates, one needs to find reliable auxiliary variables available on a population basis. In the case of LFS, the Census was identified as the source of auxiliary information at the population level providing reliable proportions of 3-digit occupation detail within the corresponding 2-digit counterpart. The Census proportions are used to model 2-digit LFS estimates down to the 3-digit level.

Initially, a model was defined for each 2-digit occupation code by province. To improve stability of the model variance, provincial estimators were pooled and averaged over time. Each step led to more precise estimates with increased stability.

Another way to increase the precision of SAE estimates is to improve the quality of the auxiliary information. This was done by taking the Census occupation proportions at the 3-digit level and fitting them proportionally in an iterative manner using the 2-digit level LFS estimates. This technique is called Structure PREServing Estimation, or SPREE. The fitted values were then used as input into the SAE model. This led to yet another improvement in precision.

The SAE modelling was done within month and across time using both 2001 and 2006 Censuses, and occupations that remained relatively stable, decreased or increased during the intercensal period. It was important to contrast the results under a number of scenarios that were of interest to the client. The initial test study was confined to 3-digit occupations by province. Estimates of proportions of 3-digit occupations within their 2-digit counterparts along with the corresponding CVs were produced.

In collaboration with Prof. Fuller a variance estimator for SPREE was obtained for estimated counts – this is a new result. When compared to the direct LFS estimates the new theory leads to a significant increase in efficiency.

Data Analysis Research (DAR)

Topic Name: Methods for Data Analysis

The Data Analysis Research Group was conducting research on methodological problems currently identified by analysts and methodologists related to statistical inference about finite population parameters and on issues arising from the modelling of survey data. The Group was also involved in transfer and exchange of knowledge and experience through reviewing and publishing technical papers, and by giving the seminars, presentations, and courses.

Progress:

The data analysis research conducted in 2008/2009 involved a number of independent projects. In this annual report we present only those that made progress during the reporting period.

Selected Topics in Design-based Methods for Analysis of Survey Data

1) The estimation of individual and population-average risks of having or developing a disease for individuals who do not belong to the surveyed population is often done by health researchers. While the estimation of these risks follows from fitting an appropriate model to the available data, the design-based standard error estimation of so-obtained estimates is not that obvious. We proposed a bootstrap method for estimation of the standard errors of predicted individual and aggregate risks. When a model estimated from one sample is to be used for estimating (validating) risks with another sample, we proposed a two-part variance estimate that captures both variabilities, one coming from the estimated model parameters, and another from the variability of the validation sample. The method is illustrated using data from the Canadian Community Health Survey. The results were presented at the JSM 2008 conference. The proceeding paper was written (Kovačević, Mach and Roberts, 2008). An extended version of the paper will be submitted to a journal.

2) An overview of the role of survey weighting in analysis of complex survey data. This research is a joint work with Prof. J.N.K. Rao, M. Hidiroglou and W. Yung. The idea is to discuss the use of survey weights and other components of sample design when fitting models to survey data. We summarize issues related to the design-based inference about model parameters distinguishing between the aggregated and disaggregated models. A paper "The role of survey weighting in analysis of complex survey data" (Kovačević and Rao, 2009) was drafted.

Multi-level Modelling of Survey Data

1) This year we studied how to approximate survey weights for units at different levels of sample hierarchy. We also studied different effects of weight approximations on the estimation of model parameters. In a typical two level model we need to weight at both levels, unless we can prove that the sampling is not informative at both levels. An additional problem may occur when the higher level units specified by the model are not the units defined by the sample design. In practice, some naïve methods for weight approximation are occasionally used. We empirically compared several of these methods including a new method developed for weight estimation based on ideas of "retrospective" sampling (Kovačević and Rai, 2003, *Communications in Statistics, Theory and Methods* (CS-TM)). A first draft of a technical report "Approximation of survey weights for multi-level modeling" (Kovačević and Pierre, 2009) is being written.

2) Standard error estimation for the hierarchical models fitted to survey data. We studied and extended applications of the bootstrap method to situations where the weights for higher level units were estimated. A SAS macro for calling the HLM6 software in order to do the bootstrapping of the hierarchical model has been improved. We also started developing a SAS macro that can call the Mplus software, so that the estimation of parameters of a variety of structural equation models and hierarchical models can be done using the Mplus while the bootstrap variance estimation is done using the bootstrap weights and the SAS macro. A paper "Implementing Resampling Methods for Design-based Variance Estimation in Multilevel Models: Using HLM6 and SAS Together" was written for the JSM Proceedings. Another paper "BHLMSAS_V0: Nouveau "logiciel" pour l'estimation de la variance par les methods de rééchantillonnage dans les modèles multiniveaux appliqués aux données d'enquêtes" has been submitted to the Research Data Centres (RDC's) Information and Technical Bulletin.

3) Material on multi-level modeling for the SSC2008 workshop on "Selected topics in analysis of survey data" containing examples and illustrations from the National Longitudinal Survey of Children and Youth (NLSCY) was put together.

Regression Analysis of Linked Survey Data

Combining data from different files using techniques of record linkage may introduce linkage errors due to a possible mismatching of individuals from different data sets. Some authors, Scheuren and Winkler (1993, 1997, *Survey Methodology Journal*) and more recently Lahiri and Larsen (2005, *Journal of the American Statistical Association*), studied impact of linkage error on estimates of linear regression model parameter, pointing out that the resulting estimates can be subject to bias and additional variability in the presence of linkage errors. The issues of joint accounting for the linkage errors and the complex sample design when analyzing the linked data sets haven't been studied so far. We extended the results of Lahiri and Larsen into a complex sample design situation and provided appropriate modifications to the methods for analysis of survey data when applied to the linked data files. A draft document on Regression analysis of linked survey files was written, Kovačević (2008). A new estimator for the linear regression coefficient is proposed. This estimator is design consistent for the finite population regression coefficient and model unbiased for the model parameter when two linked data files are of the same size. These properties extend to the case of linkage a file with survey data to a much larger file of, say, the size of the population. For instance, such situation occurred when the CCHS file was linked to the Health Person-Oriented Information (HPOI) database, or when a postcensal survey is linked to the census data. The proposed estimator is design-consistent and model unbiased. However, it is less efficient than the estimators that ignore the sample design proposed by Lahiri and Larsen (2005) and Chambers (2008). The theoretical development of the estimator for the logistic regression case has been done. An application using the real linked data set will be added to the SSC 2009 presentation.

Research in Methodology of Microsimulation Models

The PopSim model for demographic projections is being specifically designed to project changes in population composition taking into account differentials in fertility, mortality, internal and international migrations, marital status and intergenerational transfers. The PopSim model was conceived, conceptualized, and designed by the Demographic Analysis Section of the Demography Division. The contributions of methodologists relate to the development of the base population used in PopSim, as well as to the studies of identifiability and validity of considered models, efficiency, consistency, and unbiasedness of estimation procedures, *etc.* PopSim employs a host of models such as logistic regressions, log-log regressions and proportional hazards. The models are specified, estimated and validated using the base population or some other data sources including 2001 Census 2B sample, 2006 General Social Survey, *etc.* The variables selection for these models was based on a mix of statistical techniques and subject-matter importance of variables. The validation of these models is a big challenge given that these models are prediction models. The validation of "prediction" quality of some of them was done retrospectively using the 2001 Census. A certain amount of research was done especially with respect to the estimation and modeling of probabilities (or risks) of a variety of event. The estimated probabilities are used to advance individuals along a life course. Quantifying the uncertainty of projections based on microsimulation modeling is a big challenge too. We did the extensive research on these and other issues related to statistical methods used in demographic projections using the microsimulation modeling. A paper has been prepared for the Advisory Committee on Statistical Methods (see Kovačević *et al.*, 2009).

Spatial Analysis of Geocoded Crime Data

The focus of this research was on data that have been aggregated to regions and on the simultaneous autoregressive approach for modelling this type of data. Within this analytic framework we considered two commonly accepted methods that can account for the spatial effects in the model, namely the 'spatial lag model' and the 'spatial error model'. It was also important to develop diagnostics for determining the form of spatial dependence present in a given dataset so as to properly model the data. A User's Guide was developed for the SAS program to enable our client division to directly implement the spatial analysis of their data with minimal methodological intervention. During the period, further research was conducted on the use of optimal instrument variables and estimating functions to estimate the parameters in the spatial lag model. This work involved creating formulas for the variance and writing SAS code to estimate the model parameters and the variance. Data was also simulated based on the observed crime rate in Montreal neighbourhoods with both normal and exponentially distributed error terms. Parameter estimates of the

simulated data were then obtained using least squares, maximum likelihood, two-stage least squares and optimal instrument variables to compare the bias and relative efficiency of the different methods. This work was presented at the JSM in Denver in August and was followed by a paper on the use of estimating functions and optimal instrument variables for spatial models published in the proceedings (see Collins and Singh, 2008).

Collaboration with University Researchers (Mathematics of Information Technology and Complex Systems (MITACS), National Program for Complex Data Structures (NPCDS))

As a part of our participation in the NPCDS we provide guidance to up to three Ph.D. students per year during their internships in Methodology Branch. We also collaborate with these students and their thesis supervisors in the research on topics within a broad area of analysis of complex sample data. The submissions were received, and the decision was made. The results of the collaborative research will be written up in documents that would be made available to researchers within Statistics Canada. The results of the collaborative research would have important application in analyses of Statistics Canada longitudinal surveys. Three selected students proposed the following topics "Analysis of survey data with presence of measurement error", "Survival and Duration Analysis in Longitudinal Survey Data" and "Growth trajectories of youth mental health by the multiple indicator latent growth models." Appropriate data, methods, approaches, and software for these projects were discussed and made available. Two interdivisional seminars were given and one final report was written.

Some research was also done on several other subprojects "Modelling of Incomplete Longitudinal Survey Data", "Covariance Modelling Based on Longitudinal Survey Data", "Using Copula Theory in Multivariate Quantities Distribution Estimation" and "Growth Curves Estimation."

For further information, contact
Milorad Kovačević (613 951-9891, kovamil@statcan.gc.ca).

References

Chambers, R. (2008). Regression analysis of probability-linked data. *Official Statistics Research Series*, Vol. 4, Statistics of New Zealand.

Kovačević, M., and Rai, S. (2003). A pseudo maximum likelihood approach to multilevel modelling of survey data. *Communications in Statistics, Theory and Methods*, Vol. 32, No. 1, 103–121.

Lahiri, P., and Larsen, M.D. (2005). Regression analysis with linked data. *Journal of the American Statistical Association*, 100, 222-230.

Scheuren, F., and Winkler, R. (1993). Regression analysis of data files that are computer matched. *Survey Methodology*, 19, 39-58.

Scheuren, F., and Winkler, R. (1997). Regression analysis of data files that are computer matched. *Survey Methodology, Part II*, 23, 157-165.

Data Collection

Project title: Mixed Mode Research Project including Web Based Survey

Description:

A literature review was undertaken within the Framework of the Harmonized Content Modules Project for the Household Survey Strategy Initiative in order to identify the best practices in questionnaire design for mixed-mode surveys that is, when different methods of collection are used to collect the same information in a survey. It was found that much research has been done on the effect of mixed modes on the survey data and the results show that mixing modes affects the quality of the data; in fact, respondents may give different answers to the same question depending on whether the question was asked during an interview or in a self-administered questionnaire.

This raises concerns when data collected through different modes are compared. The measurement differences can confound time and mode effects in the case of surveys repeated over time and can impact on

the comparability of national and international data. However, using multiple methods of collection in a survey have practical advantages and can reduced cost, coverage and non-response errors.

It was also identified that no guidelines are available at Statistics Canada on how to best design collection instruments.

It will be beneficial that more research work on this topic be done, more specifically;

- review research work that has identified how the survey should be introduced to the respondent in view of improving response rates and standardized introductions.
- perform a literature review of books and articles related to best practices with regard to methodology, types of questions, appropriate survey length, design, layout, and collection and summarize the findings for each mode.
- find out what has been done and is being done in the area of questionnaire design in mixed mode surveys at Statistics Canada.
- establish contacts both within Statistics Canada and with others who work in similar capacities at international statistical agencies, with the idea of putting together a body of knowledge including best practices that represent the most up to date information on questionnaire design in mixed mode surveys.

Progress:

- Directors from other divisions have been contacted and positive feed-back has been received from most of them to support the project.
- Discussion has taken place regarding a future empirical study to measure mode effect in personal interview and self-administered collection modes.
- The workplan for the review of questionnaires of all social surveys has been identified. In January 2009. A COOP student was hired to create a database of questions used by Statistics Canada since 1976 for the collection of Health, Education and Employment data. The purpose of this exercise is to identify differences in the wordings in the questionnaires for the same items, difference in meanings between French and English versions and the evolution of the collection tools for these data items in order to determine if the survey data collected are comparable and what effect the question wording had on responses.
- Some guidelines for best practices in questionnaire design for web surveys has been drafted.
- Literature research on mixed mode methods and on the effect of mixed mode on the quality of survey data is on-going.

For further information, contact
Paul Kelly (613 951-6998, paul.kelly@statcan.gc.ca).

Disclosure Control Methods

Progress:

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 provided advice to or shared information on disclosure control practices with other departments and agencies including l'Institut National de la Statistique et des Études Économiques (INSEE), the Service Public Fédéral (SPF) Économie (Bruxelles) (Bruxelles), the Organisation for Economic Co-operation and Development (OECD), the Australian Bureau of Statistics, Statistics Singapore, the Bank of Korea, the Canadian Institute for Health Information, l'Institut de la Statistique du Québec, the Children's Hospital of Eastern Ontario (CHEO) Research Institute, the Institute for Clinical Evaluative Sciences, the University of Southampton, l'Institut national de la recherche scientifique (Université du Québec) and Carleton University. Continuing support on disclosure control methods is also given to the agency's RDC Program. Most of it is in terms of assistance for the implementation and interpretation of disclosure control rules for RDC data holdings.

Other resource centre activities included participation as a discussant at the 2008 Joint Statistical Meetings invited session "Reducing Disclosure Risks While Reaching Out to the Data Needs of the Public" and as a panelist on privacy and confidentiality at the "Privacy, Security and Technology: Affirming Our Rights" one-day seminar in Ottawa March 31, 2008. Lectures on Statistical Disclosure Control were given as part of the in-house course on Confidentiality, Privacy & Security. The section on Disclosure Control was revised for the

5th edition of the Statistics Canada Quality Guidelines (*to appear*). The DCRC also undertakes and coordinates research into disclosure control methods suitable for use by Statistics Canada's programs.

This was the second year of a two-year project on the redevelopment of the CONFID system for tabular data protection (called Confid2). Systems Development Division developed Alpha and Beta versions of Confid2 and drafted a preliminary users' guide. The Beta version is available for internal testing by program areas. Methodology is involved in the analysis and testing of Confid2. Information sessions on the transition to Confid2 were given to program areas and a preliminary transition plan was prepared.

Problems with the agency's Duffett Rules, which are used in several program areas to determine the sensitivity of cells in tables of magnitude, were presented to the Advisory Committee on Statistical Methods and to some subject matter areas. After an exploration of the problem an alternative rule was proposed but has not been approved yet (Tambay, Fillion and Frolova, 2008).

In business surveys, Statistics Canada uses cell suppression to protect published estimates. Another approach, developed in the U.S. Census Bureau, is to add random noise to the underlying microdata. An investigation into the effects of the random noise approach on data protection and on the quality of estimates was completed (Pursey, 2008).

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

Support Activities

For more information on the program as a whole, contact Mike Hidirolou (613-951-0251, mike.hidirolou@statcan.gc.ca).

Time Series

Progress:

Courses:

Course 433 titled 'ARIMA Modelling and Forecasting' has been revised. Courses 434 titled 'Seasonal Adjustment with X-12-ARIMA' has been reduced to 2 days and revised to use the new interface, called WinX12, from the United States Bureau of the Census (USBC).

Conversion to X12ARIMA:

Work started to update the prototype system for time series processing to use SAS Proc X12 (V9.2). Thorough testing of Proc X12 took place with extensive communication with the SAS Institute. The updated prototype will permit to make the transition to X-12-ARIMA for programs running on UNIX servers such as those in the Labour Division. A topic contributed poster session is being prepared for Joint Statistical Meetings (JSM) 2009 by Ferland and Fortier.

Benchmarking:

A draft version of the working paper "Benchmarking with Default and Fitted Survey Error Models", by Chen and Wu (2008) was released and presented at the Statistical Society of Canada (SSC) meeting in Ottawa.

The paper "A Nonparametric Iterative Smoothing Method for Benchmarking and Temporal Distribution" by Quenneville, Fortier and Gagné (2009) has been accepted for publication in Computational Statistics and Data Analysis. Further work by Quenneville, Fortier and Gagné on the convergence of the method will be presented as a topic contributed paper at JSM 2009.

Work progressed on developing compatibility tests for benchmarking. Preliminary results will be presented at JSM 2009 as a topic contributed paper by Gagné.

A new version of Proc Benchmarking with the by-statement has been released. This version is going to be used for the calendarization of the Goods and Services Tax (GST) data. Testing continued on fixing a possible

problem with a matrix inversion algorithm.

Work has been initiated toward using interpolation methods for benchmarking flow variables by cumulating the flows and the indicator series. This method proves to be useful for temporal distribution and in particular for the calendarization of the GST data.

Raking and reconciliation of time series:

Certification of Proc TSRAking was started and progressed. A topic contributed poster presentation by Bérubé and Fortier is being prepared for JSM 2009. Work started to incorporate Proc TSRAking in the updated time series processing system.

Work has been done on validating the 2-step approach to benchmarking and reconciliation and comparing it with a simultaneous approach. Results will be presented at a JSM 2009 Invited session by Fortier and Quenneville.

Trend estimation:

Work progressed on adapting and implementing Dagum (1996) non-linear filter with X-12-ARIMA. A topic contributed paper for JSM 2009 is being prepared on this subject by Martin-Romero. One of the trend-estimation methods has been applied in Hidiroglou and Patak (2008) to smooth small area estimates of occupational data. An invited presentation for JSM 2009 is being prepared by Dagum on Recent Developments in Short-Term Trend Prediction for Real Time Data Analysis.

Quality guidelines:

Work progressed on creating 'Quality Guidelines for Benchmarking' and 'Technical Guidelines on Benchmarking and Related Techniques' for the Statistics Canada Methods and Standard Committee. Similarly, work progressed on updating Statistics Canada (2000) Methods and Standard Committee Guidelines on Seasonal Adjustment and Trend-cycle Estimation.

Additional presentations:

Various presentations have been given reporting on research work that took place at the Time Series Research and Analysis Centre; see Fortier and Quenneville (2008), Quenneville (2008a) and Quenneville (2008b).

For further information, contact
Benoit Quenneville (613 951-1605, benoit.quenneville@statcan.gc.ca).

References

Dagum, E.B. (1996). A new method to reduce unwanted ripples and revisions in trend-cycle estimates from X-11-ARIMA. *Survey Methodology*, 22, 77-83.

Statistics Canada (2000). Methods and Standards Committee Guidelines for Seasonal Adjustment and Trend-Cycle Estimation.

Record Linkage

Progress:

- Continued to provide support work to Systems Development Division (SDD) for the Generalized Record Linkage System (GRLS) and Fundy, including participating in GRLS User Group meetings.
- Provided consulting services to external users of record linkage methods (Canadian Post-M.D. Education Registry).
- Provided various consulting services to internal users of record linkage methods.
- Met Jill Charker from the Australian Bureau of Statistics to discuss one of their record linkage projects.
- A committee for the maintenance of the "nickname table" in GRLS was created. It is chaired by Claude Nadeau. The "nickname table" tells GRLS that "Michael" and "Mike" (for instance) are to be seen as

equivalent when comparing given names. The committee has begun its activities. As a first step, the committee is going through a substantial backlog of nickname suggestions made over the past few years.

- Gave seminar on record linkage in series of seminars on introductory methodology topics, directed primarily to methodology new recruits (April 2008).
- Gave methodology portion of Course 0419 (Introduction to Record Linkage) in November 2008, in English. Paid for by other PRACAS codes but mentioned here anyway.

For further information, contact

Claude Nadeau (613 951-5510, claudio.nadeau@statcan.gc.ca).

Data Analysis Resource Centre (DARC) Support

Progress:

The 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. Since DARC gets funding for its support activities from a number of sources as well as the Methodology Block Fund, all of its support activities are mentioned in this report. DARC also uses a portion of its Methodology Block Fund support activities for doing research. Those research activities that are also partially funded by Data Analysis Research (1919) are described in the report for that project.

Consultations

In 2008/09 DARC did a great number of consultations on statistical methods with analysts from many areas within Statistics Canada. Some problems related to more traditional analytical approaches, such as estimation and inference related to descriptive statistics and fitting of generalized linear models to data. But other consultations involved newer more complex issues, such as the use of hierarchical models with survey data and integrating of data from more than one survey into a single analysis. As well as giving advice on a methodological approach, many consultations also included support on determining and using suitable software for an analysis.

Again this year, there was a considerable amount of consultation with other methodologists, some of whom were doing analysis themselves, others of whom were supporting subject-matter analysts of the survey data for which they were responsible. A couple of questions studied were the estimation of a confidence interval for a proportion when the estimated proportion is 0 or 1, and the use of fixed-effect regression models with survey data.

As well, the consultative support given to the Research Data Centres stayed high since the number and variety of projects underway in centres across the country continues to increase. The majority of consultations involved questions regarding appropriate methods and tools for complex survey data. Many of these consultations were with the Statistics Canada analysts who support the Research Data Centres (RDC) researchers, but some consultations were with the researchers directly. One consultation led to a submission to the RDC Information and Technical Bulletin (Roberts, 2008).

Several consultations also took place with people external to both Statistics Canada and the RDC network. As an example, an initial enquiry about bootstrap variance estimation for the Heart Health Surveys has turned into a small contract to provide suitable bootstrap weights. Another consultation was with researchers from Department of National Defence (DND) who are analyzing data from a periodic survey carried out in Afghanistan.

As well as providing advice on statistical methods for a particular problem, the consultation function also included technical reviewing and refereeing of articles for both internal and external journals, including Health Reports, Canadian Social Trends, *Journal of Official Statistics*, *Canadian Journal of Statistics (CJS)*, *Biometrika*, the RDC Information and Technical Bulletin, and the *Proceedings of the Joint Statistical Meetings* and of the *Statistical Society of Canada*.

Provision of Training

During 2008/09 DARC took over the responsibility of 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.

A one-day workshop, "Selected topics in analysis of survey data", was given at the joint meetings of the Statistical Society of Canada and the Société Française de Statistique in May 2008. There were three presenters but all DARC members contributed to the content. (Binder *et al.*, 2008)

A seminar was given at the RDC analysts' conference on some aspects of analyzing the census microdata files that are being placed in the RDC's. This included the status of our research on a variance estimation approach.

DARC members have been involved in the brainstorming sessions for 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 wished further assistance with their analysis projects. DARC also presented a seminar aimed at new recruits on analysis of survey data.

A short document, "Statistical Testing Using Survey Estimates", was written for the business survey project with the National Bureau of Statistics of China. (Mach and Roberts, 2008)

Software Evaluation, Development and Promotion

During 2008/09, DARC continued to examine various commercial software packages with respect to their suitability for analyzing Statistics Canada survey data. As well as keeping abreast of new features in software that can make use of survey bootstrap weights for design-based analysis (e.g., SUDAAN, Stata 10 and SAS 9.2), work was done in identifying what advances are being made in allowing for design information in software for multilevel modeling and for fixed effects regression. Some corrections were also made to the Statistical Package for the Social Sciences (SPSS) version of Statistics Canada's Bootvar software.

DARC was extensively involved in the process of extending access to software packages SUDAAN and Stata to analysts from a large number of divisions outside of Methodology. Time was also spent on supporting the use of these software packages.

Research

DARC support resources also go toward research-related activities. Most of these activities arise from problems identified through consultations.

There has been a great increase in the use of the census long-form microdata files as a source of data for analysis, particularly as these data are being made available in the RDC's. Because of many enquiries on how to take account of the survey design when analyzing these data, a research project was initiated to examine the topic of variance estimation. A first-year statistics student from the University of Ottawa participated in the research during the summer of 2008. The objective of the research is to determine an approach (or approaches) to variance estimation that yield estimates of adequate quality which can be obtained using software generally employed by researchers in their analysis of other sources of survey data. Linearization techniques are the first area of investigation. Different possibilities are being investigated through the examination of a variety of realistic analytical examples. An overview of the project has been submitted to the RDC Information and Technical Bulletin (Pierre *et al.*, 2008)

A chapter on "Using Marginal Mean Models for Data from Longitudinal Surveys with a Complex Design: Some Advances in Methods" was published this year in *Methodology of Longitudinal Surveys*, edited by Peter Lynn. (Roberts *et al.*, 2009).

Research progressed on how survey bootstrap variance estimation might be carried out for Structural Equation Modeling (SEM) and Multilevel Modeling (MLM) models making use of the Mplus software. As well, there was investigation of how survey bootstrap variance estimation might be carried out for MLM models making use of the gllamm procedure of Stata. Currently, no commercial software exists that can make use of the survey bootstrap weights for variance estimation for MLM or SEM. Thus, it is useful to investigate possibilities of combining commercial MLM and SEM software with custom-written programs in order to provide tools for analysts.

David Binder and Geoff Rowe of Socio-Economic Analysis and Modeling Division jointly prepared and presented a Topic Contributed Joint Statistical Meetings paper titled "Can Survey Bootstrap Replicates be used for Cross-validation?" The paper was written up for publication in the Joint Statistical Meetings

Proceedings. (Rowe and Binder, 2008) A shortened version of the presentation was given at the 2008 Fall Seminar of the Statistical Society of Ottawa.

A chapter on "Design and Model Based Inference for Model Parameters" was finalized for the Handbook of Statistics, Vol. 29 on *Sample Surveys: Theory Methods and Inference* being edited by D. Pfeffermann and C.R. Rao. Galley proofs have been reviewed. (Binder and Roberts, 2009)

David Binder prepared and presented paper on "Answers to criticisms of using design-based methods for inference" as a special Survey Methods Section session at the 2008 annual meeting of the SSC. (Binder, 2008b) This paper was also presented as a Social Survey methods Division (SSMD) seminar, at the Quebec Inter-university Centre for Social Statistics (QICSS) in Montreal and at Carnegie-Mellon University.

David Binder worked part-time in DARC, as part of the alumnus program. He was involved in various consultation and research activities described above. As well, he presented an invited paper at the Sample Surveys and Bayesian Statistics Conference held at Southampton Statistical Sciences Research Institute, Southampton - 27-29 August, 2008. The paper is titled "Bayesians and Frequentists under Informative and Non-informative Survey Sampling". (Binder, 2008d) The talk will be repeated at Harvard in April 2009. He also had an entry on design-based estimation in *The Encyclopedia of Survey Research Methods*, published in this fiscal year (Binder, 2008c), as was his discussion in *Journal of Official Statistics* on Sedransk's paper entitled "Assessing the Value of Bayesian Methods for Inference About Finite Population Quantities" (Binder, 2008a).

Other activities

DARC resources were used for involvement with the National Institute on Complex Data Structures (NICDS). DARC members organized the competition for the (NICDS)/Mathematics of Information Technology and Complex Systems (MITACS) internships of PhD students at Statistics Canada and two of the three students completed their internships during the reporting period under the guidance of Milorad Kovačević, Georgia Roberts and Abdelnasser Saidi. Milorad Kovačević was also a member of the program committee of the NICDS.

DARC members were involved in a variety of organization functions for the program of the Survey Methods Section of the SSC in 2008 and 2009.

Georgia Roberts sat on the editorial boards of Health Reports and Canadian Social Trends and is on the Analysis Steering Committee.

Georgia Roberts and David Binder gave a guest lecture at the course on Advanced Epidemiology at the University of Ottawa.

Milorad Kovačević is a member of the Appeal Accreditation Committee of the SSC.

David Binder reviewed and was an external examiner of a PhD thesis on weighting in mixed linear effects models for survey data (Carnegie Mellon University). David also met with Office for National Statistics (ONS) methodologists in Titchfield and Newport, U.K., to discuss a variety of topics.

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

Generalized Systems

Progress:

Research work focussed on three topics during the year – the redesign of the Confid system, Banff and Statistical Macro Extensions (StatMx).

Confid:

The specifications for the Confid2 tabular data confidentiality system were completed and provided to the programming team in SDD. Several preliminary versions of the software were developed and sent to methodology for testing. Tests were performed to compare the results with those coming from the old Confid system and a prototype of the new system to ensure that the results matched or that differences could be

explained. Volume testing was also performed. Some problems in performance were identified and resolved either by adjusting the methodology or by alternative means of programming. A beta version was provided to a limited number of users for testing and the system is scheduled to be put into production April 30, 2009.

Banff:

A new version of the Banff edit and imputation system (version 2.03) was released which resolves some of the problems that users had experienced when edits contained either very large or very small coefficients. This was fixed by means of a normalization of the parameters which decreases the chance of extreme values being generated in the linear programming step. The new version is also compatible with SAS Enterprise Guide 4, and Microsoft Vista. A generic processing module which allows Banff to be run via metadata tables was developed by the programmers and tested by methodology. Feedback and suggested improvements were provided and implemented. This processor will be available in a beta version in April 2009. A review of potential improvements for both Banff and the processor that had been identified by the development team or users took place. Priorities for improvements were established.

A study was carried out on the effectiveness of Banff for imputing categorical Labour Force Survey data using both ordered and unordered categorical data. The results confirmed that in its current format it may be feasible to use it for ordered data, but not unordered data. This was the first study as part of a project to potentially make changes to Banff which would allow it to process qualitative data.

StatMx:

A new version of the calibration estimation function of the StatMx was released to users. This version included functionality for two-phase estimation and variance calculations beyond those that are available in the Generalized Estimation System (GES) and complements the two-phase weighting which was released earlier. Minor upgrades were made to other previously released modules in order to extend its functionality to a wider range of users

Other:

A new generalized systems website was developed for the Statistics Canada intranet in conjunction with SDD (<http://gensys>). The Generalized Estimation System and Banff sites were completed and published. Others will be completed in the future

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

Quality Assurance

Progress:

Quality Control Data Analysis System (QCDAS)

The QCDAS system is a SAS system 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. Work continued on the development of the QCDAS system. Regular meetings were held between methodology and SDD to address any issues or confusion about the system requirements. A beta version of the system was provided to methodology and thoroughly tested. Comments were provided to the programmers for improvements and changes were made which met methodology's requirements. Following the initial testing, a final test version was used in a parallel run with the old system. Results from the two systems were compared. Following the completion of this step, the software was certified by methodology and put into production.

Automated Coding Parsing and Scoring Methodology

Research continued on the current method used for word scoring in the Automated Coding and Text Recognition (ACTR) system. A literature search for alternative methods was conducted and a number of possible alternatives were identified. Some of these alternatives were programmed into the ACTR system for

testing while others are being tested in a standalone system. To this point no method has been found which is vastly superior to the quantity and quality of results coming from the current method. However the research will continue. Some initial ideas for a generic quality control module to assess the quality of both automated and manual coding were proposed.

Methodology acted as consultants to Standards Division regarding a new generalized coding tool that is being developed. They identified numerous issues that should be considered and explained how changes to the word weighting methodology could impact this new system.

Generic Intelligent Character Recognition (ICR)

A study to examine the thresholds being used to reject batches of questionnaires which are captured using the generic ICR process took place. The initial thresholds were set prior to the development of the ICR system. This new study used actual data to determine the success of each kind of capture (numeric fields, alpha-numeric fields, check boxes etc.) Based on the results some modifications were made to the thresholds to better align themselves with the observed quality. This Quality Control (QC) methodology is used on all data captured through ICR by Operations and Integration Division (OID). The Census of Population and Census of Agriculture were not captured by OID.

Methodology has also been in contact with the US Census Bureau to discuss our respective ICR systems and to exchange ideas on ways to improve them. A delegation will visit Statistics Canada in April 2009.

Statistical Training in Quality Control:

The support team continued to offer courses in quality control to both methodologists and non-methodologists.

Consultation: Training and advice on measures related to quality control and assurance were provided to individuals both internal and external to Statistics Canada. Internal areas in which assistance was provided included Standards Division, the Census of Population, the Census of Agriculture and Collection and Regional Services Branch.

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

Training and Consultation

Progress:

The Statistical Training Committee co-ordinates the development and delivery of twenty 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 total of 31 scheduled sessions were given in either English or French.

The suite of courses continues to expand. The course 0451, Corrected Exercises on Sampling, was given for the first time in both English and French in the winter of 2009. In addition, a new course on Reconciliation of Time Series is in preparation.

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

Conferences

Symposium 2008, 2009

Progress:

The 24th International Methodology Symposium took place from October 29th-31st, 2008 in Gatineau Quebec. Over 500 people took part. The theme this year was "Data Collection: Challenges, Achievements and New Directions". The keynote address was given by Robert Groves and the Waksberg address by Mary Thompson. Three workshops were held the day prior to the symposium.

Edith D. de Leeuw: **Mixed-Mode surveys: Innovation and Challenge**
Roger Tourangeau: **The Psychology of Survey Response**
Jean-François Beaumont: **Using Paradata to Weight for Non-response**

In the months leading up to the symposium the 2008 organizing committee put together the program consisting of presentations from thirteen different countries. A poster session was also held. The team looked after all of the issues related to registration and the operations of the event. Following the symposium they collected the papers which will be included in the official proceedings of the symposium.

Preparations for the XXV International Methodology Symposium are also well underway. This Symposium will be held from October 28 to 30, 2009 at the Palais des congrès in Gatineau with the theme "Longitudinal Surveys: From Design to Analysis". The keynote speech will be given by Peter Lynn and Graham Kalton will present the Waksberg Award Address. A day of workshops will precede the Symposium and an exhibition dedicated to the 25th anniversary will complete the program.

For more information, contact
Christian Nadeau (613 951-4555, christian.nadeau@statcan.gc.ca)
Julie Trépanier (613 951-9437, julie.trepanier@statcan.gc.ca).

Survey Methodology

Progress:

The journal *Survey Methodology* (SM) is published twice a year in both official languages and features articles on various statistical development topics that are of interest to statistical organizations. Emphasis is placed on the development and evaluation of methodologies applicable to data collection and the data itself.

- The June 2008 issue (Volume 34-1) was released on June 26, 2008. It contains 11 articles on various topics.
- The December 2008 issue (Volume 34-2) was released on December 23, 2008. It contains 11 articles, including the eighth of a series of annual solicited articles in honor of Joseph Waksberg. This year, the author was Mary E. Thompson.
- *Survey Methodology* has been available in electronic form since the June 2000 issue (Volume 26-1). We are continuing putting past editions on line.
- The SM program's daily activities include corresponding with authors, reviewers, assistant editors and subscribers, examining reviewers' comments and authors' revisions, reformatting manuscripts, revising manuscripts, liaising with translation and printing services and updating the database on articles submitted for publication.

For more information, contact
Eric Rancourt (613 951-5046, eric.rancourt@statcan.gc.ca).

Date Modified: 2009-06-08

Research Papers Sponsored by MRDC

Beaumont, J.-F., and Patak, Z. (2009). Generalized bootstrap for prices surveys. To appear in the *Proceedings of the International Statistical Institute*, Durban.

Beaumont, J.-F., and Bocci, C. (2009a). Synthetic data for the Canadian portion of the Cross-National Equivalent File: Some theory. Preliminary report (in progress).

Beaumont, J.-F., and Bocci, C. (2009b). Variance estimation when donor imputation is used to fill in missing values. To appear in the *Canadian Journal of Statistics*.

Beaumont, J.-F., and Bocci, C. (2009c). A practical bootstrap method for testing hypotheses from survey data. *Survey Methodology*, 35, 25-35.

Beaumont, J.-F., Haziza, D. and Ruiz-Gazen, A. (2009). A unified approach to robust estimation in finite population sampling. Report to be presented at the International Statistical Institute (ISI) conference in Durban.

Beaumont, J.-F., and Rivest, L.-P. (2009). Dealing with outliers in survey data. To appear in the Handbook of Statistics, *Sample Surveys: Design, Methods and Applications*, Vol. 29A.

Bérubé, J., and Fortier, S. (2009). PROC TSRAKING: A user-defined SAS® procedure for balancing time series. *Proceedings of the Business and Economic Statistics Section (B&E) of the American Statistical Association (ASA)*, Washington, to appear.

Binder, D.A. (2008a). Discussion on "Assessing the value of Bayesian methods for inference about finite population quantities" by Joseph Sedransk. *Journal of Official Statistics*, 24, 513-516.

Binder, D.A. (2008b) Answers to criticisms of using design-based methods for inference. Special Presentation at 2008 Annual Meeting of the Statistical Society of Canada, Survey Methods Section. Ottawa.
www.ssc.ca/survey/documents/SSC2008_D_Binder.pdf.

Binder, D.A. (2008c). Design-based Estimation. Entry in *Encyclopedia of Survey Research Methods*, (Ed., P.J. Lavrakas). Sage: Los Angeles.

Binder, D.A. (2008d). Bayesians and Frequentists under Informative and Non-informative Survey Sampling. Sample Surveys and Bayesian Statistics Conference. Southampton, UK. August 26-29, 2008.
www.s3ri.soton.ac.uk/ssbs08/papers/12%20David%20Binder.pdf.

Binder, D.A., Kovačević, M. and Roberts, G. (2008). Selected Topics in the Analysis of Survey Data. Survey Methods Section Workshop at the 2008 Annual Meeting of the Statistical Society of Canada.

Binder, D.A., and Roberts, G. (2009). Design and model based inference for model parameters. In *Sample Surveys: Inference and Analysis*, (Eds., D. Pfeffermann and C.R. Rao). Elsevier: Amsterdam, Vol. 29, B, Chapter 24.

Binder, D., Roberts, G. and Kovačević, M. (2008). Selected topics in analysis of survey data. Workshop presented at the 2008 Annual Conference of the Statistical Society of Canada (SSC), Ottawa.

- Bizier, V., You, Y., Veilleux, L. and Grondin, C. (2008). Model-based Approach to Small Area Estimation of Disability Count and Rate Using Data from the 2006 Participation and Activity Limitation Survey. Methodology branch working paper, SSMD-2008-007E/F, Statistics Canada, Ottawa, Canada.
- Chen, Z.G., and Wu, K.H. (2008). Benchmarking with default and fitted survey error models. Presented at the SSC meeting, Ottawa, Canada.
- Collins, K.J., and Singh, A.C. (2008). Use of optimal instrumental variables for spatial autoregressive models with application to crime survey data. In *Joint Statistical Meetings, Proceedings Business and Economics Statistics Section*. Alexandria VA: American Statistical Association.
- Dagum, E.B. (2009). Recent developments in short-term trend prediction for real time data analysis. *Proceedings of the Business and Economic Statistics Section (B&E) of the American Statistical Association (ASA)*, Washington, to appear.
- Dochitoiu, C., and Rubin-Bleuer, S. (2008). Small area estimation to study the impact of globalization. *Proceedings of the Survey Methods Section*, SSC Annual Meeting, May 2008.
- Estevao, V., Hidirolou, M.A., You, Y. and Patak, Z. (2009). Implementation notes for small area estimation. Business Survey Methods Division (BSMD)/Statistical Research and Innovation Division (SRID) technical paper.
- Ferland, M., and Fortier, S. (2009). Recent developments in Statistics Canada's time series processing system - Transition to SAS PROC X12. *Proceedings of the Business and Economic Statistics Section (B&E) of the American Statistical Association (ASA)*, Washington, to appear.
- Fortier, S., and Quenneville, B. (2008). Assisted Forecasting for Linking Non Overlapping Time Series. Presented at the 28th International Symposium on Forecasting, Nice, France.
- Fortier, S., and Quenneville, B. (2009). Reconciliation and balancing of accounts and time series, from concepts to a SAS procedure. *Proceedings of the Business and Economic Statistics Section (B&E) of the American Statistical Association (ASA)*, Washington, to appear.
- Gagné, C. (2009) Compatibility tests for benchmarking. *Proceedings of the Business and Economic Statistics Section (B&E) of the American Statistical Association (ASA)*, Washington, to appear.
- Gambino, J., and Hidirolou, M.A. (2008). Conducting experiments for surveys at Statistics Canada. Paper presented at the November 2008 Advisory Committee on Statistical Methods.
- Hidirolou, M.A. (2008). Basic area-level small area procedure. Preliminary specifications for the SAE system. Internal SRID document.
- Hidirolou, M.A. (2008). Benchmarking small area totals. Preliminary specifications for the SAE system. Internal SRID document.
- Hidirolou, M.A., and Lavallée, P. (2009). Sampling and estimation in business surveys. To appear in the Handbook of Statistics, *Sample Surveys: Design, Methods and Applications*. Vol. 29A.
- Hidirolou, M.A., and Patak, Z. (2009). Estimation for occupations counts using the Labour Force Survey. Invited paper to be presented at the SSC meetings to be held in Vancouver.
- Hidirolou, M.A., and Patak, Z. (2009). Small area estimation feasibility study for human and skills development Canada. SRID Internal working document.
- Hidirolou, M.A., Rao, J.N.K. and Haziza, D. (2008) Variance estimation in two-phase sampling. To appear in *Australian New Zealand Journal of Statistics*.
- Hidirolou, M., You, Y., Patak, Z. and Estevao, V. (2009). Implementation notes for small area estimation. Statistics Canada Internal Documentation.
- Kovačević, M. (2008). Regression Analysis of Linked Survey Files. An internal technical manuscript.
- Kovačević, M., and Pierre, F. (2009). Approximation of survey weights for multi-level modeling. An internal technical manuscript.

Kovačević, M., Grondin, Ch., Titus, R.-C. and Carrillo-Garcia, I. (2009). Challenges in methodology of microsimulation models: Examples from projections of the diversity of the Canadian population, 2006 - 2031. A paper written for the *Advisory Committee on Statistical Methods (ACSM)*, Session 48.

Kovačević, M., Mach, L. and Roberts, G. (2008). Bootstrap variance estimation for predicted individual and population-Average risks. *2008 Proceedings of the Survey Research Methods Section*, American Statistical Association. Alexandria VA: American Statistical Association.

Kovačević, M., and Rao, J.N.K. (2009). The role of survey weighting in analysis of complex survey data: Overview. A draft of an invited paper for the *International Statistical Review (ISI)* conference.

Mach, L., and Roberts, G. (2008). Statistical Testing Using Survey Estimates. *SIMP II - Economic Statistics Document Series, Document # 32*.

Mach, L., and Rubin-Bleuer, S. (2009). Coordination of Statistics Canada Livestock and Crop Surveys. Reference Period: December 2007 – November 2012. Internal report. *SRID*.

Mantel, H., and Hidiroglou, M. (2008). Synthetic data for confidentiality. Methodology Branch Working Paper, Statistics Canada.

Martin-Romero, B. (2009). An investigation of trend. *Proceedings of the Business and Economic Statistics Section (B&E) of the American Statistical Association (ASA)*, Washington, to appear.

Nugyen, L., Davis, K., Oddy, C., St-Jean, H. and Melki, L. (2009). Generalized Quality Control for Optical Data Capture within Statistics Canada. Proceedings of the Statistics Canada's 2009 Methodology Symposium

Patak, Z., and Beaumont, J.-F. (2009). Generalized bootstrap with an application to prices surveys. Report to be presented at the ISI conference in Durban.

Pierre, F., and Saidi, A. (2008). Implementing resampling methods for design-based variance estimation in multilevel models: Using HLM6 and SAS together. *2008 Proceedings of the American Statistical Association*. Alexandria VA: American Statistical Association.

Pierre, F., and Saidi, A. (2008). BHLMSAS_V0: Nouveau « logiciel » pour l'estimation de la variance par les méthodes de rééchantillonnage dans les modèles multiniveaux appliqués aux données d'enquêtes. Has been submitted to the *Information and Technical Bulletin of Research Data Centres (RDCs)*.

Pierre, F., Villeneuve, A., Saïdi, A. et Roberts, G. (2008). Plans de sondage et méthodes d'estimation alternatifs pour l'analyse des données contenues dans le fichier de microdonnées du Recensement de 2001: Un aperçu. Submitted for publication in the Research Data Centres (RDC) Information and Technical Bulletin (ITB).

Pursey, S. (2008). Disclosure Control: The Application of Random Noise to the Microdata of a Business Survey to Protect the Confidentiality of Respondents' Data. Internal Business Survey Methods Division (BSMD) paper, August 1, 2008.

Quenneville, B. (2008a). Applied Time Series at Statistics Canada, presented at the Spring Symposium of the Statistical Society of Canada, Ottawa, Canada.

Quenneville, B. (2008b). Time Series Methods for Official Statistics, presented at the workshop "Application of statistical methods to national accounts", headquarters of CEPAL, the Economic Commission of the United Nation for the Latin American and Caribbean countries, Santiago, Chile.

Quenneville, B., Fortier, S. and Gagné, C. (2009). A Nonparametric Iterative Smoothing Method for Benchmarking and Temporal Distribution, Computational Statistics and Data Analysis, doi:10.1016/j.csda.2009.02.021.

Roberts, G. (2008). Analyzing Census microdata in an RDC: What weight to use? Submitted for publication in *The Research Data Centres Information and Technical Bulletin*.

Roberts, G., Ren, Q. and Rao, J.N.K. (2009). Using Marginal Mean Models for Data from Longitudinal Surveys with a Complex Design: Some Advances in Methods. In *Methodology of Longitudinal Surveys*. (Ed., P. Lynn), Wiley, U.K., Chapter 20, 351-366.

Rowe, G., and Binder, D. (2008). Can Survey Bootstrap Replicates be Used for Cross-Validation? *2008 Joint Statistical Meetings – Section on Survey Research Methods*.

Rubin-Bleuer, S. (2008). IGLS to estimate variance components for the Fay-Herriot area level model. Specifications for variance estimation in SAE system. Internal document, SRID.

Rubin-Bleuer, S., and Dochitoiu, C. (2009). Bootstrap estimation of mean squared error of small area estimators under area level models for business data. Internal SRID Working paper. Will be part of a paper co-written with J.N.K. Rao.

Rubin-Bleuer, S., and Yeung, A. (2009). Specs on robust estimation of SAE for unit level models.

Singh, A.C., and Nadeau, C.J.J. (2008). An Alternative to the Logit-Wald Method for Inference under Models for Proportions. Contributed paper presented at *American Statistical Association (ASA)* in Denver.

Singh, A.C. (2008). Single Phase Simplified Variance Estimation Approach to Two Phase- Stage Hybrid Designs Contributed paper presented at ASA in Denver.

Tambay, J.-L., Fillion, J.-M. and Frolova, O. (2008). Revising Confidentiality Rules for the Release of Business Data. Technical Report Presented at Statistics Canada's Advisory Committee on Statistical Methods, October 27-28, 2008.

You, Y. (2008). Small area estimation using area level models with model checking and applications. *Statistical Society of Canada 2008 Proceedings of the Survey Methods Section*.

You, Y., and Zhou, Q.M. (2008). Hierarchical Bayes Small Area Estimation under a Spatial Model with Application to Health Survey Data. Internal Document, HSMD.

Yung, A., and Rubin-Bleuer, S. (2009). Small Area Estimation: A Business Survey Application. Internal Working Document. Will be presented at an invited talk at the SSC Conference in 2009.