

Topic II: Impact of new technological developments in software, communications and computing on SDC

PROVIDING GREATER ACCESSIBILITY TO SURVEY DATA FOR ANALYSIS

Contributed paper

Submitted by Statistics Canada¹

I. INTRODUCTION

1. To ensure that its data holdings are both relevant and accessible, Statistics Canada must be able to address the evolving needs of the research community while protecting the confidentiality of its data, which remains a primary objective. Following a revision to the *Statistics Act* in 1971, the Agency started to release public use microdata files. With the advent of longitudinal household surveys in the early 1990s, the Agency began to explore approaches that would permit access by researchers to the rich databases of these surveys yet also ensure that the confidentiality requirements of the *Statistics Act* would be respected. This paper will discuss three approaches developed by the Agency to provide researchers with access to data produced by complex surveys: public use microdata files, remote access and research data centres.

II. PUBLIC USE MICRODATA FILES (PUMF)

2. Statistics Canada began producing public use microdata files following the 1971 revision to the *Statistics Act* that made possible the public release of non-confidential microdata. The release of a microdata file for a survey is authorised by the Agency only when doing so substantially enhances the analytical value of the data. Planned microdata products are submitted to a Microdata Release Committee which must be satisfied that all reasonable efforts have been made to protect the identity of respondents before it grants permission for the release of a public use microdata file. Submissions to the committee must include documentation of the survey and its data contents as well as a description of measures taken for disclosure protection. The onus is on the survey manager to take all the necessary steps to ensure that the microdata can be released and to convince the Committee that all possible measures have been taken to protect the confidentiality of survey respondents. Microdata can only be released for sample data. Since 1971, a total of 371 public use files have been reviewed and 345 have been approved by the Microdata Release Committee for public dissemination.

3. In recent years, changes in the nature and uses of survey data have led to additional considerations with regards to PUMF disclosure protection. These have to do with linkages with external files, the estimation of variances and the protection of longitudinal survey microdata. These will be treated in turn.

¹ Prepared by Jean-Louis Tambay and Pamela White.

Linkages with external files

4. Even though the PUMF license agreements stipulate that licensees must not attempt to re-identify survey units by matching records to other data files, it is necessary to remove the ability to link PUMF records with external population databases. For this reason, survey design information such as cluster or stratum membership are not disseminated, and geographical or other variables that can be used as indirect identifiers are present in limited form. A concept called ‘multiplicity’ is also used to identify for treatment units presenting a greater disclosure risk (Boudreau, 1995). Consider a situation where 10 categorical variables could be used as indirect identifiers, the multiplicity of a unit would then be defined as the number of possible three-dimensional tables involving any three of these 10 variables in which the unit sits alone in its cell. The higher the unit’s multiplicity, the greater the risk of linking it to an external database. With 10 variables the highest possible multiplicity is $10!/(3!7!)=120$.

5. The longitudinal Survey of Labour and Income Dynamics (SLID) faced a new type of linkage problem. Respondents to this household survey were given the choice between reporting income data or authorising the Agency to use their taxation data. About 75% of respondents have given permission to the Agency to link their survey response to their personal tax file in order to use their taxation data. In creating the SLID public use file, the Agency wanted to ensure that it would not be possible for users, such as the Canada Customs and Revenue Agency, to link a tax filer database to the SLID PUMF using common ‘taxation’ variables.

6. As a first attempt, ‘taxation’ variables for the entire sample, i.e., including respondents whose tax data were not used, were randomly rounded to reduce the number of unique matches with the taxation database. The taxation database had been rounded deterministically using the same rounding base. The number of unique matches between the two rounded files and the proportion of ‘true’ matches among them were examined. A second, coarser, rounding was done for some problem records to bring matching uncertainty to an acceptable level.

7. Since those intent on undertaking a linkage are quite likely to use sophisticated matching algorithms, a second match was attempted using nearest-neighbour matching, where the distance function was based on the largest univariate relative difference. The proportion of nearest-neighbour matches that were true matches was examined. Results showed a bigger risk than with the first method.

8. A third random rounding method was used to add even more noise to the data set. Rather than choosing between the closest multiples of the rounding base above and below the value, the closest n multiples in each direction were eligible for selection (Nadeau, Gagnon and Latouche, 1999). A 75% decrease in the proportion of unique matches was observed and a decrease of similar magnitude was observed in the proportion of unique matches that are correct. Further investigations using nearest-neighbour matching will be investigated by the Agency.

9. The third method gave satisfactory results in terms of impact of rounding on estimates of means, standard deviations and correlations, except for two tax credit items. These items often take values close to zero and as a result were eligible for a different perturbation scheme.

10. To date, two versions of the SLID public use files have been released. The first set covers the reference year 1993 and the second covers reference years 1993 and 1994. In the latter case, cross-sectional and longitudinal public use files were both produced. The Microdata Release Committee is currently evaluating the above described techniques and approaches designed to prevent linkage with other files. The issue of the release of microdata obtained from taxation sources has yet to be fully resolved by the Agency.

Estimation of variances

11. Sets of coefficient of variation (CV) lookup tables accompany PUMFs so that users can derive crude variance estimates for simple statistics such as means, proportions, ratios and differences of ratios. For categorical variables, these generated CVs are derived by applying a design effect to the variance formula for a simple random sample without replacement (e.g., N^2PQ/n). Design effects are calculated for

a set of representative variables and a conservative value, such as the 75th percentile of the design effects, is used for the tables. Sets of CV lookup tables can be provided for various levels of geography, such as Canada and for each province.

12. Increasingly, researchers would like to calculate exact estimates of variances. These analysts frequently use software designed for complex surveys, such as SUDAAN, WesVar or STATA, or they generate their own variance programs using statistical analysis packages such as SAS or SPSS. Proper variance estimation requires the use of design information such as stratum and replicate membership. This information is considered confidential and is not present on PUMFs.

13. An approach for variance estimation that does not require the release of explicit survey design information consists of generating a set of jackknife weights corresponding to a fixed number of “replicates”, for example, 30. The longitudinal National Population Health Survey (NPHS) has considered such an approach. Collapsed strata and replicates are used to form superstrata and superreplicates (Mayda, Mohl and Tambay, 1999). Strata containing a compatible number of replicates were collapsed (e.g., three-replicate strata and six-replicate strata could be combined) using a method similar to the one presented by Rust (1986). Their superreplicates contained enough original replicates that their membership could be released to users interested in exact variances. This approach has several weaknesses, however. From a confidentiality perspective, providing superreplicate information does facilitate the reconstruction of households for those surveys collecting information for all household members. Also, in situations where cross-sectional PUMFs are produced for different waves of a longitudinal survey, the use of superstrata and superreplicates can impede the data provider’s efforts to prevent the linkage of units between waves. From an analytical perspective, the method significantly reduces the number of degrees of freedom, and hence the precision of the variance estimates. Furthermore, if a researcher, for example, wished to undertake logistic regressions using a number of classification variables, it would be possible to run out of degrees of freedom.

14. An alternative approach is to suppress design information but allow for the bootstrap estimation of variance by releasing a series of bootstrap weights (say, 100 to 500). The bootstrap variance estimation method (Efron, 1982) is more versatile than the jackknife method and better suited for non-smooth statistics such as medians and percentiles. It is also well-suited for longitudinal surveys since bootstrap weights, once produced, can follow units wherever they move. This facilitates variance estimation when the longitudinal sample is also used to produce cross-sectional estimates, as is often the case. When the jackknife method is used, for example, a unit that changed its province of residence would add an extra level of complication as its stratum would have to explicitly contribute to cross-sectional variance estimates for both its old and new province.

15. An anticipated problem with the bootstrap approach concerns commonly observed patterns of zero bootstrap weights as these could allow the reconstruction of replicates (a replicate that is not in a particular bootstrap sample would have a zero weight assigned to all of its units). A mean bootstrap method was developed to address this problem. Rather than selecting a simple random sample with replacement (SRSWR) of $r-1$ replicates from the original r in a stratum, the mean bootstrap method selected 20 SRSWR of $r-1$ replicates (Yung, 1997). The process is repeated if zero bootstrap weights remain. However, a careful examination of the resulting pattern of bootstrap weights showed that the correlation between bootstrap weights for units in the same replicate was almost equal to one, and a strong negative correlation was observed for units in different replicates of the same stratum. These relationships could be used to recreate the suppressed design information (Yeo, Mantel and Liu, 1999).

16. For the NPHS, there is a preference for the bootstrap method to estimate variances; however, these bootstrap weights are not released to outside users. Those who wish to calculate variances must submit variance calculation programs using remote access. The survey area continues to investigate other more practical solutions. For example, one option being considered is to combine the bootstrap method (for its simplicity) and the collapsing of strata and replicates (for the protection it offers). A method that swapped units between replicates was also investigated, but it failed to give satisfactory variance estimates.

Protection of longitudinal survey data

17. In response to the need for a more comprehensive understanding of Canadian society by policy analysts and researchers and the need for information that could be used to investigate outcomes, the Agency initiated in the 1990s several longitudinal household surveys. These include SLID, NPHS and the National Longitudinal Survey of Children and Youth (NLSCY). Microdata from longitudinal surveys present additional disclosure risks because of their rich data content and the unpredictable evolution of their panels over time. The release of longitudinal and cross-sectional public use sample files from these surveys present considerable challenges to the Agency.

18. The team involved with SLID, a household survey that follows panel respondents for six consecutive years, studied the evolution of characteristics such as marital status. This indicator was used as a way to decide on an appropriate level of detail to include in the PUMF for its first wave. It was intended for SLID to produce longitudinal PUMFs for each year of the lifetime of the panel and therefore it was important not to jeopardise future releases: once information is released, it cannot be withdrawn. A precursor longitudinal survey whose respondents were surveyed on three occasions, the Labour Market Activity Survey (LMAS), was used to estimate change patterns. Certain categories were combined, and geographical detail was minimal. A longitudinal PUMF from SLID waves one and two was released and the transitions in values were also subjected to disclosure analysis (Grondin, 1995). However, alternatives to PUMFs such as remote access and research data centres may be more appropriate venues given the considerable confidentiality challenges that need to be addressed. The other longitudinal surveys that started at the same time have not released longitudinal PUMFs. The LMAS was the only other Statistics Canada survey to release a longitudinal PUMF, and it did so only after its third, and final, wave. The decision not to release longitudinal PUMFs was easier to make for surveys that had data sharing agreements under section 12 of the *Statistics Act*. For example, over 95% of NPHS respondents give permission to share their survey information with Health Canada and provincial ministries of health.

19. Even without the release of longitudinal PUMFs, many challenges remain. Although the surveys are longitudinal, there is a strong demand for cross-sectional microdata. For the NPHS, for example, provinces at different times have purchased large additional samples to allow for reliable estimates at the level of their Health Regions. Furthermore, in addition to the core content data items, each wave collects information on a specific topic such as mental health or the use of health services. Such additional samples and topics generate interest in data for individual waves.

20. In releasing cross-sectional PUMFs, the Agency must prevent the linking of PUMF units between waves, even though the PUMF license agreement forbids such an activity. The choice of data to put on successive PUMFs is thus influenced by the ability of users to use these variables for linking records. Prior to public release, Agency staff attempt to undertake linkages between waves using common but stable variables. In doing these in-house diagnostic linkages, both one-to-one linkages and the percentage of correct one-to-one linkages are examined. Data variables are dropped or regrouped until these parameters are below acceptable levels. This is typically done in consultation with subject matter experts, as they have preferences as to which variables should be dropped first. Certain units representing particular problem cases are also examined and dropped or modified (leaving observed inconsistencies in the data can also help to reduce the ability to link).

21. It is noted that, even if there was interest in releasing even a very limited longitudinal PUMF, the previous release of cross-sectional PUMFs severely precludes that eventuality. In addition to verifying that the data contents of the longitudinal PUMF are safe from a confidentiality perspective, one would have to ensure that the PUMF could not be used to link the content-rich cross-sectional PUMFs (Béland, 1999). In consideration of these difficulties, it has been decided that the third cycle release of a public use file from the NPHS would be the last to be submitted to the Microdata Release Committee. This is due to the increased risk of linking records of respondents to the previous cycles and possibly identifying unique individuals. The preferred strategy for expanding the analytical richness and value of the longitudinal survey files is to consider proposals from researchers and to provide on-premise access or access in the research data centres as “deemed employees” or to use remote access. Remote access is also an option for users who have signed section 12 agreements and where the “sharing” portion of the longitudinal file experiences attrition over time. As respondents have not given permission for Statistics Canada to share

their information with the data sharing partner, access to the master file at the research data centre or on-premises is not possible, thus remote access is the most viable option for the data sharing partner.

III. REMOTE ACCESS

22. Remote access allows researchers access to a richer base of survey microdata without compromising the confidentiality of the data. Two types of remote access are used at Statistics Canada. One approach involves users e-mailing programs to Statistics Canada where they are submitted on confidential survey master data files residing inside the Agency's firewall. Outputs are vetted for confidentiality before being e-mailed to the users. This type of remote access was used in a pilot study for the LMAS and is now available to NPHS and NLSCY users. The other type of remote access has been developed to facilitate access to the Census tabulation system for five federal departments that provided funding for the 1996 Census. Both types of remote access are described.

Remote access for survey data

23. At Statistics Canada remote access for survey data is provided for research purposes only. The aim is to increase the analytical scope of the data and to simplify procedures for custom tabulations; for example, users can write their own programs that they submit to the Agency. Setting up a remote access program entails certain considerations that are presented in this paper for the National Population Health Survey (NPHS). This is the Agency's most successful application to date (Mantel and Nadon, 1999).

24. A remote access program's success is contingent on the availability of good survey documentation, the creation of synthetic (dummy) files for program testing, the ability to run a variety of software and a relatively fast turn-around time. The NPHS had all these. The thorough PUMF survey documentation was supplemented with information on the non-PUMF variables. The dummy files closely resembled the master files. For wave 2, there were three sets of each, one set containing basic cross-sectional information on each household member, a second set containing detailed cross-sectional information for the selected panel member and a third set containing longitudinal information for the selected panel member. The dummy files had the same record layout as the master files but had much fewer records (5% for the first file and 10% for the other two). A reasonably large number of records was used to allow proper functioning of regressions involving many variables. Records on the dummy files were created by taking blocks of variables from different randomly selected donor records from the master files. In order to produce internally coherent dummy records, records from the master file were divided into donor classes based on similar pathways through the questionnaire. Dummy records were created from within donor classes thus allowing a more realistic testing scenario. The dummy file data has no analytical value: imputed values come from several records, different donors being used for different parts of the questionnaire.

25. Although most survey data analysis at Statistics Canada is done using SAS, the NPHS remote access program also accepted programs written in other languages including SPSS and STATA. To encourage the successful analysis of the NPHS, analysis workshops were given to users across the country. These covered the use of analysis techniques for surveys having a complex design as well as aspects specifically related to the analysis of the NPHS. SAS macros were provided to assist the calculation of variances that incorporate the survey design for totals, ratios, differences of ratios, regressions, logistic regressions and general linear models using the bootstrap weights. Just as with the master file, dummy versions of the bootstrap weights files were created for testing. For users who wanted to use specialised survey analysis packages such as STATA or to write their own variance estimation programs, the design information was provided (simulated for the dummy files).

26. As was hoped, the analysis workshops generated considerable interest in the analysis of NPHS data. In the calendar month following two particularly important workshops remote access requests doubled to a high of 190.

27. NPHS remote access is free for PUMF licensees (this includes all Canadian universities) and holders of a National Health Research & Development Program grant. Potential users send a research

outline to Statistics Canada. If approval is given, the users are sent a User Guide, the dummy master files and dummy bootstrap weights files. They e-mail programs to the Agency where rules including a thirty-respondent minimum are applied for confidentiality and reliability purposes. The vetting of outputs is done manually and results are usually e-mailed back within two working days. Users making elaborate requests are informed that the production and verification of their results may take more time. Currently, there are no plans to automate the system as the workload is manageable. About one half of a person-year is devoted to remote access and level of resource might decrease with the introduction of research data centres.

Remote access for the Census

28. The Census Products and Services System (PASS) is a collection of decentralized tools for the specification, compilation and production of Census information products. A major component of PASS is the Computer Assisted Product Specification System (CAPSS), which includes GUI for product specification (including tabular product preview), automated code generation and job processing, automated implementation of Census concepts and business rules, including rules for disclosure control and data quality. CAPSS supports user-defined areas or variables and a variety of output media, such as disquettes, CD-ROMs, and paper printouts, and a variety of formats: including Beyond 20/20, ASCII, C91, CSV. Outputs are manually reviewed for validity and for residual disclosure before being sent to users.

29. In order to provide more timely access to Census data for five federal departments that provided funding for the 1996 Census, the PASS system was decentralized. Prior to decentralization, these external users had to pre-specify their requirements several months in advance and their tables were then specified internally by Statistics Canada employees. With decentralization, clients from these departments now have designated workstations at their site where they can specify requirements using CAPSS and get results much sooner as the minimum turnaround for these users is 2-3 days. The decentralized system is illustrated in Figure 1.

30. For data security, Statistics Canada uses two internal networks: a closed network designed to hold and protect confidential information (Network A) and a network accessible by public means of communication such as telephone lines (Network B). There is no direct access to Network A from the outside. PASS decentralization works by placing a copy of the Census Metadata on a server on Network B, which is accessible to a federal client workstation for table specification purposes. Through the use of "Replication Server" software, a synchronized copy of the table specifications created on Network B are moved to a switchable server on Network A via a robotic switch during a specified period each day. The robotic A/B switch ensures that no access can be gained to the internal Network A. Once copied into the Network A environment, the client specifications are queued and executed against the Census microdatabase. Resulting tabular outputs are then verified, packaged and delivered to the client. Clients can continue to specify other requests in the Network B environment which will queue for replication the next day. Apart from the required manual review and delivery of products, the process is fully automated.

IV. RESEARCH DATA CENTRES

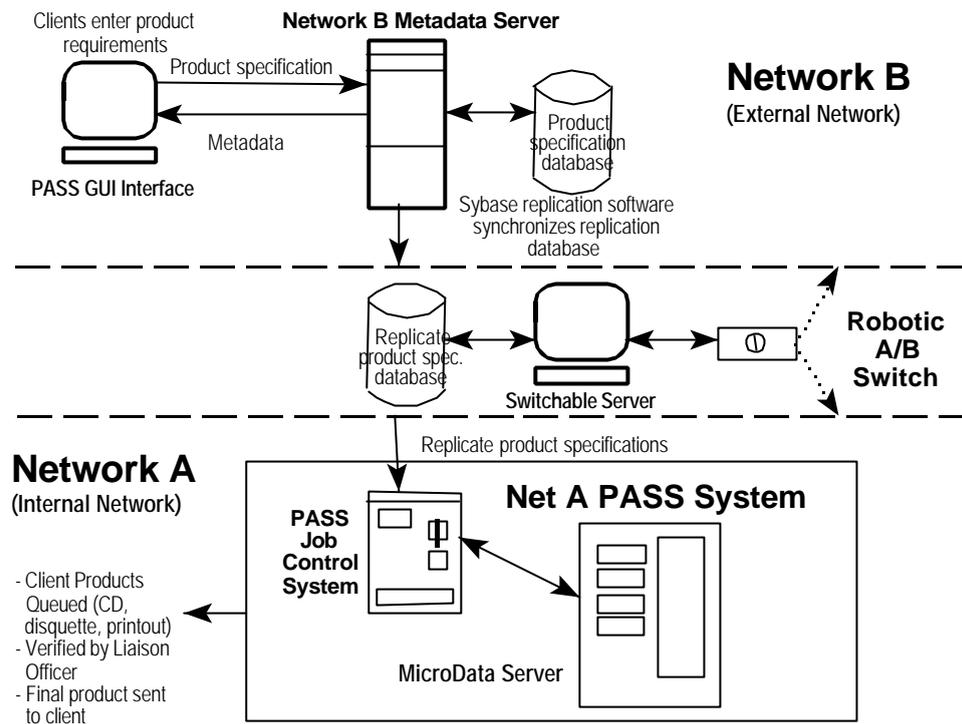


Figure 1: Strategy for Allowing Access to the Census PASS Specifications Tools by Other Government Departments

31. Nine research data centres (RDC) are being established in Canadian universities to improve the accessibility to longitudinal microdata by accredited researchers conducting projects approved by a joint committee of Statistics Canada and the Social Sciences and Humanities Research Council (SSHRC). The proposal submission process is administered by SSHRC and once the work has been approved, the researchers will take the Oath of Secrecy under the *Statistics Act* and undergo an enhanced reliability security check before being provided access to the RDC. Only the files specified in the research proposal will be supplied to these “deemed employees”. All Statistics Canada microdata files used in the RDC will be stripped of name, address and any identifiers. As deemed employees under the *Statistics Act*, these researchers are subject to the same restrictions and penalties as all other Statistics Canada employees. These researchers must also participate in a confidentiality training course offered by the RDC before they are permitted to use the facility.

32. The first centre was officially opened in December, 2000, with others scheduled to come into operation over the next year. Each centre has been constructed according to Statistics Canada physical security requirements with each operating a self-contained internal computer network with no links to any outside computing centre or to the Internet. All researchers work at diskless workstations.

33. To ensure that no confidential statistical information is disclosed by researchers, all computer outputs such as aggregate tables or results of modelling or data analysis must be screened for confidentiality by the Statistics Canada RDC manager, who is present at all times that the RDC is open. The results of these approved research projects will be first published in the RDC Paper Series by Statistics Canada and must not contain any content of a program policy or political nature. Following the initial publication by Statistics Canada that will be peer reviewed by SSHRC and Statistics Canada, researchers can then publish secondary analyses, however, this work must be based on the research findings first released by Statistics Canada.

34. To facilitate the analysis of complex business survey data, the Agency is considering the possibility of opening a microeconomic research data centre on-premises in 2001. This centre would operate under the procedures similar to those of university located RDCs with a joint SSHRC/Statistics Canada approval of the proposed research. All researchers would take the Oath of Secrecy and undergo an enhanced reliability check. The results of the research would first be published by Statistics Canada. The proposal is at the exploratory stage and a decision on whether and if so how to proceed will be made sometime in 2001.

V. CONCLUSION

35. Statistics Canada is not alone in trying to provide greater accessibility to its data for analysis (Horm, 1999). This paper presents three approaches to the problem, each at a different stage of evolution. Public use microdata files have been produced by the Agency for almost two decades, but demands for better analytical products that allow exact variances, and the use of rich data sources such as longitudinal or taxation data have led to new sets of problems for which solutions are being sought. At Statistics Canada, the development of a remote access capability and of research data centres are still in their early stages, though researchers undertaking approved research for Agency have been accessing microdata on-premises (Headquarters and in the Regional Offices) for several years. These approaches address many of the more complicated issues related to PUMFs, but raise new ones in their turn, for example, protection of confidentiality when PUMFs and remote access are both available and requests for access to the master file by those signatory to a section 12 data sharing agreement. In conclusion, the three approaches are in many ways complementary and will coexist for some time. Developing ways to provide access to these rich data sources and also ensuring the confidentiality of individual respondents and respecting their privacy rights as it pertains to the sharing and linking of their personal information remains a formidable challenge to Statistics Canada.

References

- Béland, Y. (1999). Release of Public Use Microdata Files for NPHS? Mission ... partially accomplished! *Proceedings of the Survey Research Methods Section, ASA*.
- Boudreau, J.R. (1995). Assessment and Reduction of Disclosure Risk in Microdata Files Containing Discrete Data. *Proceedings of Statistics Canada Symposium 95*. Statistics Canada, No. 11-522-XPE.
- Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans*. Philadelphia: Society for Industrial and Applied Mathematics.
- Grondin, C. (1995). Confidentialité du fichier de microdonnées de l'Enquête sur la dynamique du travail et du revenu (EDTR). *Actes du colloque sur les méthodes et applications de la statistique 1995*, Bureau de la statistique du Québec.
- Horm, J. (1999). National Center for Health Statistics Approaches to Protection and Release of Microdata. *Proceedings of the Joint Eurostat/UN-ECE Work Session on Statistical Data Confidentiality*, Thessaloniki, March 1999.
- Mantel, H. and Nadon, S. (1999). Dummy File Creation for the Remote Access Program of the National Population Health Survey. *Proceedings of the Survey Methods Section, SSC Annual Meeting*, June 1999.
- Mayda, J., Mohl, C. and Tambay, J.L. (1996). Variance Estimation and Confidentiality: They Are Related! *Proceedings of the Survey Methods Section, SSC Annual Meeting*, June 1996.
- Nadeau, C., Gagnon, É. and Latouche, M. (1999). Disclosure Control Strategy for the Release of Microdata in the Canadian Survey of Labour and Income Dynamics. (1999). *Proceedings of the Survey Research Methods Section, ASA*.
- Rust, K. (1986). Efficient replicated variance estimation. *Proceedings of the Survey Research Methods Section, ASA*.
- Statistics Act, R.S.C., 1985, C. S-19.*
- Yeo, D., Mantel, H. and Liu, T.P. (1999). Bootstrap Variance Estimation for the National Population Health Survey. *Proceedings of the Survey Research Methods Section, ASA*.
- Yung, W. (1997). Variance Estimation for Public Use Microdata Files. *Proceedings of Statistics Canada Symposium 97 - New Directions in Surveys and Censuses*, Statistics Canada, No. 11-522-XPE.