WP.35
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE
CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (iv): New and emerging methods, including automation through machine learning, imputation, evaluation of methods

EDIT AND IMPUTATION FOR THE 2006 CANADIAN CENSUS

Supporting Paper

Submitted by Statistics Canada[1]

1. Many minimum change imputation systems are based on the approach proposed by Fellegi and Holt (1976). In the 1996 Canadian Census of Population, a somewhat different approach was used successfully to impute for both non-response and inconsistencies for the demographic variables of all persons in a household simultaneously. The method used is called the Nearest-neighbour Imputation Methodology (NIM). This implementation of the NIM allowed, for the first time, the simultaneous minimum change donor imputation of qualitative and quantitative variables for large E&I problems. In Bankier (1999), an overview of the NIM algorithm is provided.

2. The main difference between the NIM and the Fellegi/Holt imputation methodology is that the NIM first finds donors and then determines the minimum number of variables to impute based on these donors. The Fellegi/Holt methodology determines the minimum number of variables to impute first, and then attempts to find donors. Reversing the order of these operations confers significant computational advantages to implementations of the NIM while still meeting the well-accepted Fellegi/Holt objectives of minimum change and preserving sub-population distributions. The NIM, however, in its present form, can only be used to carry out imputation using donors while the Fellegi/Holt can be used with any imputation methodology.

3. For the 2001 Census, a more generic implementation of the NIM was developed. It is called the CANadian Census Edit and Imputation System (CANCEIS). Besides the demographic variables, it was used in 2001 to perform minimum change donor imputation for the labour, mobility, place of work and mode of transportation variables. This corresponds to 40% of all variables on the 2001 Census questionnaire. The System for Processing Instructions from Directly Entered Requirements (SPIDER), which has been used since 1981, did all the other donor imputation plus all the deterministic imputation. For the 2006 Canadian Census, CANCEIS will process all census variables and will be extended to do deterministic imputation. CANCEIS has also been used by the annual Canadian Survey of Household Spending.

4. Prior to the development of CANCEIS, enhancements to the NIM were implemented in prototype software. This software was used to process variables in the 2000 Brazilian and Swiss Censuses. CANCEIS has also been provided to the national statistical offices of the United Kingdom, Italy, Brazil, USA, New Zealand and Australia, among others, for their evaluation. Studies by Italy and

[1] Prepared by Michael Bankier, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6. Mike.Bankier@statcan.ca.

USA have shown that CANCEIS does a good job in preserving both univariate and multivariate distributions.

5. In the 2001 Census, the SPIDER modules were implemented using approximately five thousand machine-readable decision logic tables (DLTs). Because of budget and time constraints, the conversion from SPIDER to CANCEIS for 2006 must be done as efficiently as possible. It is planned to initially convert SPIDER modules to CANCEIS with few changes. After all modules are converted, some enhancements may be done. A few modules, however, will have to be rewritten because of major changes to Census questions.

6. The syntax used in CANCEIS DLTs allows the user to define a single generic rule comparing responses for two or more persons. CANCEIS then replicates this rule for all pairs of persons in the household. In addition, the data dictionary allows the use of labels such as “Married” rather than a numeric code. This makes the DLTs easier to read.

7. CANCEIS processed the 2001 Census variables on personal computers (PCs) under DOS with no Windows interface. Processing these on the mainframe computer would have cost almost four hundred thousand Canadian dollars. For 2006, Windows interfaces for CANCEIS have been created to simplify its use. The primary interface helps the user to submit jobs plus specify most of the input files required by CANCEIS. A second interface makes it easier to specify the DLTs by allowing access to lists of variables and possible responses from the data dictionary.

8. Besides the Windows interfaces, a number of other improvements are being made to CANCEIS. In 2006, CANCEIS will be able to process alphanumeric variables as well as discrete, continuous, and coded variables. In 2006, the introduction of the Lp norm will allow more extensive use of continuous variables. Other extensions to the CANCEIS methodology will be made, as required, to ensure that SPIDER modules are successfully ported to CANCEIS.

9. The enhancements to CANCEIS for 2006 have been developed in an iterative and collaborative fashion by the methodologists, systems analysts and subject matter experts. The CANCEIS syntax has been made as similar as possible to the SPIDER syntax to simplify the conversion of DLTs. At the same time, the implementation of new features have been prioritized to ensure that the software, with the essential features, is ready on time.

REFERENCES

Bankier, M. (1999), "Experience with the New Imputation Methodology used in the 1996 Canadian Census with Extensions for Future Censuses", Proceedings of the UN/ECE Work Session on Statistical Data Editing, Italy (Rome). (http://www.unece.org/stats/documents/1999.06.sde.htm)

Fellegi, I.P. and Holt, D. (1976), "A Systematic Approach to Automatic Edit and Imputation", Journal of the American Statistical Association", March 1976, Volume 71, No. 353, 17-35.
