WP.8
ENGLISH ONLY

**UNITED NATIONS STATISTICAL COMMISSION and
ECONOMIC COMMISSION FOR EUROPE**

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Bonn, Germany, 25-27 September 2006)

Topic (ii): Editing data from multiple sources

EDITING A MIXTURE OF CANADIAN 2006 CENSUS AND TAX DATA

Invited Paper

Prepared by Michael Bankier, Statistics Canada[1]

I. INTRODUCTION

1. Respondents to the 2006 Census 20% sample long form had the option of giving Statistics Canada permission to link to their 2005 income tax form in lieu of answering the thirteen part income question. Early returns indicate a permission rate of 83%. The permission question (Question 51) and the income question (Question 52) are reproduced in an appendix at the end of this paper. The income question was only asked of those more than 14 years old. The tax linkage option was given to reduce response burden plus achieve a more complete set of income responses since the level of partial and total non-response to the income question was rising. In addition, the income responses in past censuses were often of the nature of an approximation while the responses on tax forms are usually very accurate.

2. In this paper, a brief review of the census/tax record linkage process is given. Next, data collection and processing performed prior to edit and imputation (E&I) is discussed. Finally, the strategy to perform E&I on a mixture of census and income tax data is outlined.

II. CENSUS/TAX RECORD LINKAGE

3. In the fall of 2006, a record linkage operation will take place to link census long forms and 2005 tax forms for those persons who gave permission in Question 51. Statistics Canada's Generalized Record Linkage System (GRLS) will perform this linkage and is based on the methodology proposed by Fellegi and Sunter (1969).

The following variables will be used in this census/tax linkage:

- Last Name
- Second Last Name
- First Given Name
- Second Given Name
- Birth Date
- Address
- Postal Code
- City
- Province
- Telephone Number
- Sex
- Marital Status
- Disability Status
- Labour Activity Status

4. Obviously not all elements will match exactly from both sources because of transcription errors, legitimate changes in responses between the two collections and conceptual differences. To increase the match rates, some tolerance is accepted; nicknames can be matched to full given names, the order of names can be reversed, small typographic errors are allowed through the use of string comparators and

[1] Michael Bankier, Social Survey Methods Division, Statistics Canada, Ottawa, Ontario, Canada, K1A 0T6. Mike.Bankier@statcan.ca. I would like to thank Eric Olson and Marcel Bureau of Statistics Canada for their extensive comments and suggestions. Updated August 30, 2006.

records are compared from across Canada because of the mobility of the population. In addition, to increase the quality of the links, more weight is given to matches based on less common names and addresses. Only very good matches are retained since incorrect matches can generate undesirable outliers. Also, no manual review will be done of all the links because of the very large volume of data[2]. Instead, the threshold chosen to accept the links will be set higher than might be acceptable with a manual review. The parameters of the linkage process will be fine tuned by running it several times and assessing the quality of the links for a sample of persons. Once a mapping between records is established, only the data required on the Census form is actually retrieved from the administrative file. Based on tests done to date with data from the 2004 Test Census, a match rate of at least 85% is expected.

III. DATA COLLECTION AND PROCESSING PRIOR TO E&I

5. In the 2001 Census, enumerators listed dwellings and dropped off a questionnaire. These questionnaires were completed and then mailed back by the respondent. In remote areas, the enumerator would interview the respondent and complete the questionnaire.

6. For the 2006 Census, dwellings in urban areas were listed in advance of the Census and then questionnaires were mailed to them. This was done for approximately two thirds of the dwellings in the country. The other one third of the dwellings was enumerated in the same way as in 2001. Respondents had the option, for the first time, of completing their questionnaires over the Internet. To use the Internet option, the respondent had to enter an Internet access code provided on their paper questionnaire. The level of encryption used with the Internet application was higher than that typically used with on-line banking transactions. Approximately 20% of the population took advantage of the Internet option.

7. Completed questionnaires were scanned and data captured at the processing centre in Ottawa using intelligent character recognition (ICR). Any responses that could not be captured were keyed from a snippet of the imaged questionnaire. In the 2001 Census, corrections were made before keying, where possible, for amounts reported in foreign currencies, amounts reported on a weekly or monthly basis or where cents were reported along with the dollar figure. For the 2006 Census, it was not feasible to do such corrections prior to keying. In the 2004 Test Census, the use of segmented boxes to facilitate ICR resulted in cents sometimes being reported as dollars. This response error was difficult for ICR or keyers to identify. This resulted in an error rate of approximately 11% for the income variables. It is expected that the situation will be similar in the 2006 Census.

8. Non-respondents or partial respondents with non-response to a large number of questions were contacted by phone or visited if necessary by local staff.

9. Various coverage edits were applied at the processing centre and occasionally persons were added or subtracted from households based on a manual review.

10. Other edits were used to flag those persons with one or more income responses outside specified limits. These persons were reviewed manually by comparing their income responses to income related characteristics and observing the questionnaire image when available for additional information. The responses were manually modified if necessary. These erroneous responses might have been the result of reporting errors (deliberate or otherwise), errors caused by the character recognition system or because of keying errors. As of mid-June, the main problems observed have been decimals not being recognized during capture or not being provided by the respondent, confusion between income sources, monthly amounts being reported and occasionally erroneous amounts being entered as a prank. Responses from tax forms were excluded from this process because the tax linkage process is completed later and the subject matter experts consider tax data mostly error free for the important fields.

11. The Dwelling Classification Survey (DCS), after field follow-up, picked a sample of enumeration areas. The DCS revisited households for which a questionnaire was not received because their occupancy status was listed by field operations as

- unoccupied and not part of the housing stock,

[2] It is expected to attempt to match 4 million consenting Census respondents to 24 million tax filers.

- unoccupied or
- occupied but could not be contacted or refused to respond.

The occupancy status was double checked by the DCS as well as how many people lived there on census day. These results were compared to those from the census. Estimates of undercoverage and overcoverage that are the result of errors made regarding occupancy status will be produced based on the DCS sample and will be used to adjust the census data base in October 2006.

12. The Reverse Record Check (RRC) Study measures undercoverage and overcoverage from all sources and is used to adjust the provincial population counts. The census data base, however, will not be adjusted using the RRC estimates because of timing issues and because the estimates are not sufficiently reliable at the small area level.

IV. EDIT AND IMPUTATION OF THE INCOME QUESTION

13. In February 2007, with the completion of the linkage activities, both income data from the tax source and the census source will be available on the census database. Flags will be used to indicate the source of the income data both before and after imputation. The automated E&I of the income responses will be described in this section.

14. The **Can**adian **E**dit and **I**mputation **S**ystem (CANCEIS) (see Bankier 1999, 2003) will be used to perform deterministic imputation, donor imputation plus derive new variables for all census variables including income. It is assumed that the income data given by most respondents is correct and hence every attempt will be made to change as few responses as possible. Some income fields will be imputed deterministically. Donor imputation will be used to resolve non-response for many income fields. There will also be balance edits to make sure that the income components sum to within 10% of the total income after imputation for non-response. Then, in a later step, the total income is adjusted to ensure perfect agreement.

15. A series of modules (mostly using CANCEIS but with one using SAS) will be run to process the income data. The first three modules will merge the tax and census income data together, calculate the average employment income by occupation and geography for later use as a matching variable, define strata to be used in later modules and determine the response status for each income field (income with amount reported, income indicated but not reported, loss indicated but not reported, no income, non-response). Modules 4 to 6 will impute missing income responses using donors while ensuring that the total is within 10% of the sum of the components. Module 4 will impute partial respondents who have provided total income. Module 5 will impute partial respondents who have not provided total income but have provided all components of employment income. Module 6 will impute the remaining partial respondents and total non-respondents to the income questions. Because in the past, there was a problem with the under-reporting of benefits from the Canada and Quebec pension plans, Modules 7 and 8 will select a sample of respondents with no pension benefits and will impute positive amounts through donor imputation. Modules 9 and 10 will do something similar but for Employment Insurance Benefits. Module 11 will derive other government benefits such as the Old Age Security Pension. Module 12 will use donor imputation to resolve non-response for the income tax field. Module 13 will derive total income after tax. Other modules aggregate incomes to the Census Family, Economic Family and household levels plus derive two low income flags.

16. During donor imputation, income data from tax records will be generally treated the same as income data from the census forms. The following income sources, however, will not be available from the tax records chosen for the linkage and must be deterministically imputed:

- the provincial tax portion of income tax for residents of Quebec,
- Child Benefits and
- Good and Services Tax (GST) Credits.

Note that Child Benefits will also be generated deterministically for census forms as well since the responses to this question have been found to be of poor quality. Also since many persons on the census form erroneously do not report the GST Credit, this has to be derived for them as well.

17. When adjusting for under-reporting of the Old Age Security Pension, Employment Insurance Benefits and the Canada or Quebec Pensions, responses from tax records are not subjected to this adjustment because of the general policy not to modify them. When imputing for the income tax field from census forms, the donors will probably be restricted to data from tax forms because of the expected poor quality of the responses for the income tax field from the census forms.

18. Within a donor income module, the number of strata in the 2006 Census will be reduced dramatically compared to the 2001 Census (where an E&I software called SPIDER was used). This is because a number of variables used for stratification by SPIDER, such as Age, will now be used in the CANCEIS distance measure to identify nearest neighbour donors. In addition, only exact matches were allowed within a stratum by SPIDER when searching for donors while CANCEIS will allow "near" matches (e.g. an age difference of 3 years). Because SPIDER insisted on exact matches, occasionally no donor was found and default imputation was used. With CANCEIS, a donor will always be found but the donor may occasionally not resemble the failed record that closely. Note that CANCEIS allows the user great flexibility in the distance measures used so that the matching of a 14 year old to a 17 year old might not be allowed while the matching of a 75 year old to a 87 year old might be considered completely acceptable.

19. Reducing the number of strata in CANCEIS will simplify processing and reduce "boundary effects". With SPIDER, for example, persons might have been stratified as under 35 years of age, 35 to 60 years of age, 61 to 65 years of age etc. Thus if a 35 year old required imputation, we could not have used a 34 year old as a donor since they would have belonged to a different stratum.

20. Donor Selection edits are also used extensively in the CANCEIS income edits. They specify that certain records, while they pass the edits, cannot be used as donors because certain income fields have unusually large or small values when cross-classified with other variables. By excluding these records from being used as donors, the number of outliers and inconsistent records generated by imputation is reduced.

21. Note that in the search for donors by CANCEIS, the distance measure used applies larger weights to income fields considered more reliable or important such as total income. For some income components, such as net farm income, the numeric amount can be missing but boxes can be checked indicating that there is net income and that it is a loss and hence negative. The distance measure of CANCEIS can be configured such that the donor imputation can be almost guaranteed to impute a negative quantity in such cases.

22. On an experimental basis, a certain number of income responses were blanked out and then a CANCEIS donor module was used to impute these blanks. It was found that CANCEIS was quite effective at replicating the responses and preserving distributions when the other variables used in the selection of donors were correlated with the variables being imputed.

23. Eventually it is hoped that tax data will permit us to reduce the number of deterministic modules by obtaining Child Benefits, for example, from other sources. Using CANCEIS, it may be possible to reduce the number of modules used in later censuses and possibly improve consistency with the labour and education variables.

24. A certain number of persons provide both permission to link to tax data and then proceed to answer the income questions on the Census form. It will be interesting to compare their census form responses to those on their tax forms. In the 2004 Test Census, a study of these respondents showed that they tended to provide responses on the Census form rounded to the nearest thousand or five thousand dollars. The new Census collection procedure with multiple channels for forms (mail, Internet, respond over the telephone during follow-up) permits duplication (some of which is usually caught during the coverage edits). These duplicates will be extremely interesting for the evaluation of responses for proxy or mode effects.

V. CONCLUSIONS

25. Many aspects of the processing of income data have changed for the 2006 Canadian Census including the use of tax data, the new questionnaire layout to allow the scanning and intelligent character recognition of the data and the use of a new E&I system. These many changes will require careful monitoring during production and may require some fine-tuning of the processing techniques used. Given the generally high quality of tax data, its availability should prove very beneficial.

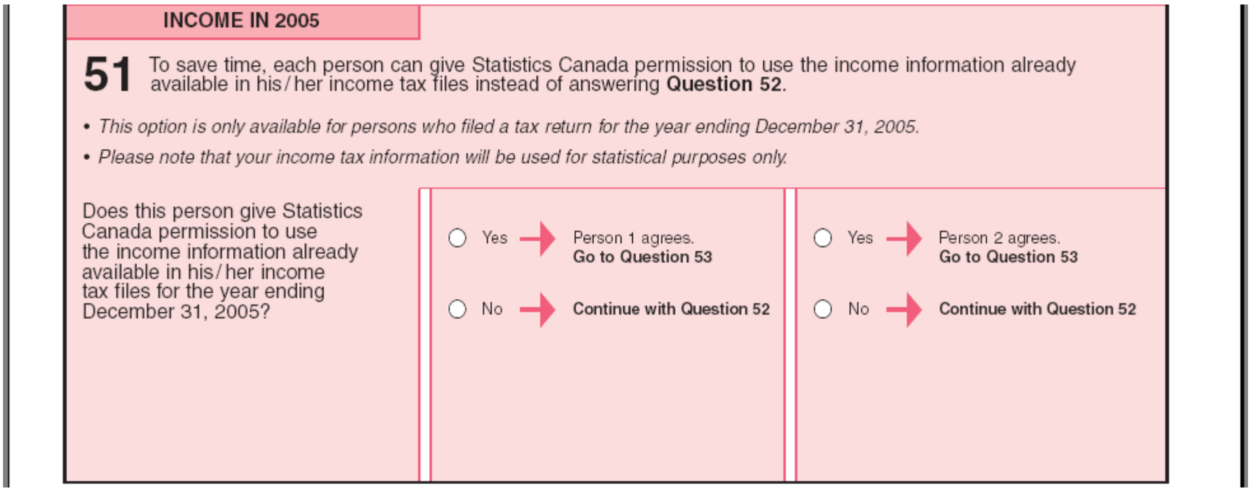
References

Bankier, M. (1999), "Experience with the New Imputation Methodology Used in the 1996 Canadian Census with Extensions for Future Censuses", Working Paper 24, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Italy (Rome).
(http://www.unece.org/stats/documents/1999.06.sde.htm)

Bankier, M. (2003), "Current and Future Applications of CANCEIS at Statistics Canada", Working Paper 18, Proceedings of the UN/ECE Work Session on Statistical Data Editing, Spain (Madrid).
(http://www.unece.org/stats/documents/2003.10.sde.htm)

Fellegi, I. P. and Sunter, A. B. (1969), “A Theory for Record Linkage”, Journal of the American Statistical Association, 64, 1183-1210.

VII. APPENDIX – QUESTIONS 51 AND 52 FROM 2006 CANADIAN CENSUS LONG FORM

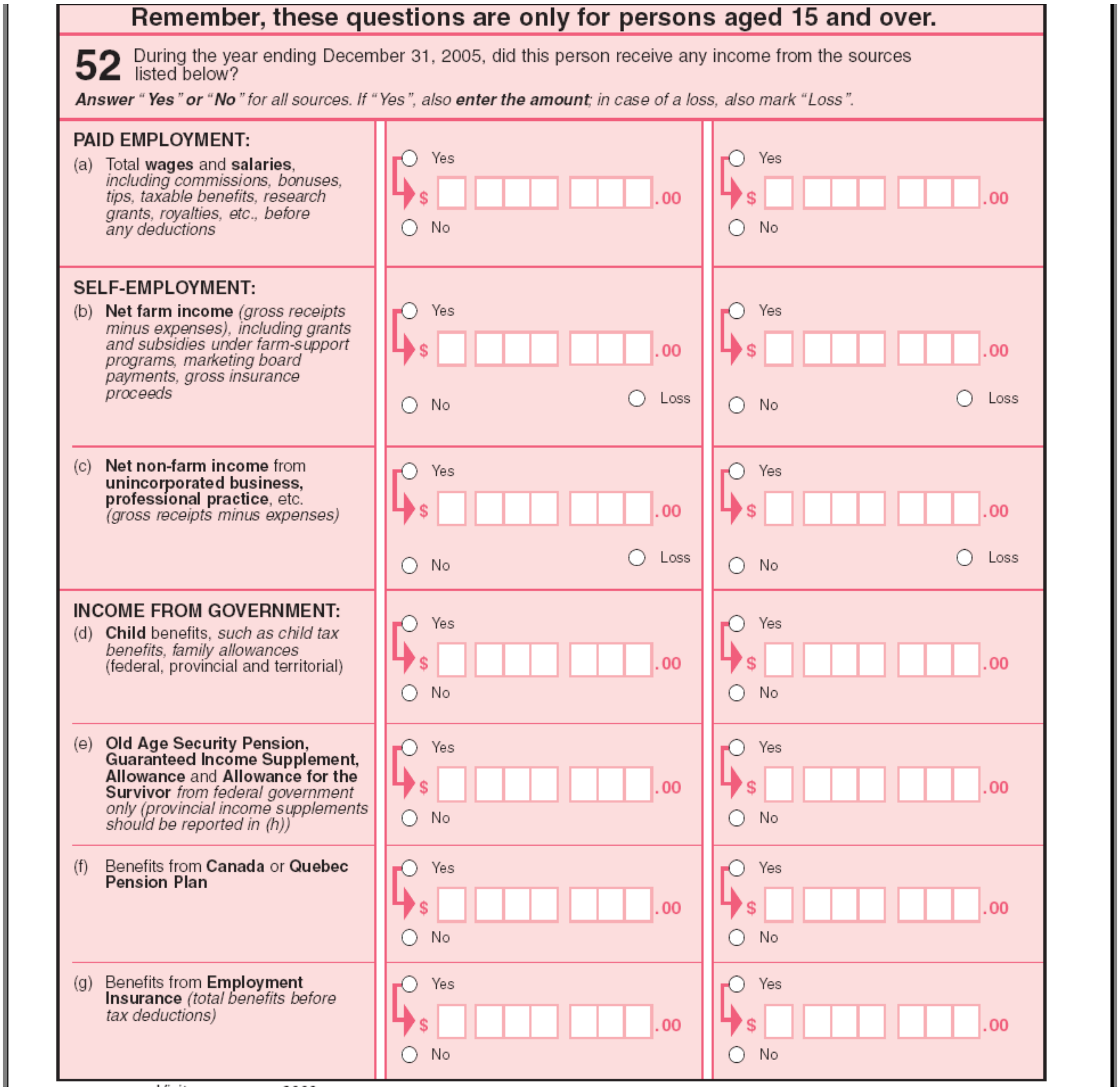


INCOME IN 2005

51 To save time, each person can give Statistics Canada permission to use the income information already available in his / her income tax files instead of answering **Question 52**.

- *This option is only available for persons who filed a tax return for the year ending December 31, 2005.*
- *Please note that your income tax information will be used for statistical purposes only.*

Does this person give Statistics Canada permission to use the income information already available in his / her income tax files for the year ending December 31, 2005?	○ Yes → Person 1 agrees. **Go to Question 53** ○ No → **Continue with Question 52**	○ Yes → Person 2 agrees. **Go to Question 53** ○ No → **Continue with Question 52**



Remember, these questions are only for persons aged 15 and over.

52 During the year ending December 31, 2005, did this person receive any income from the sources listed below?

Answer* " *Yes* " *or* " *No* "** *for all sources. If " Yes ", also* ***enter the amount*; in case of a loss, also mark " Loss ".*

PAID EMPLOYMENT: (a) Total **wages** and **salaries**, *including commissions, bonuses, tips, taxable benefits, research grants, royalties, etc., before any deductions*	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No
SELF-EMPLOYMENT: (b) **Net farm income** *(gross receipts minus expenses), including grants and subsidies under farm-support programs, marketing board payments, gross insurance proceeds*	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No ○ Loss	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No ○ Loss
(c) **Net non-farm income** from **unincorporated business, professional practice**, etc. *(gross receipts minus expenses)*	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No ○ Loss	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No ○ Loss
INCOME FROM GOVERNMENT: (d) **Child** benefits, *such as child tax benefits, family allowances* (federal, provincial and territorial)	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No
(e) **Old Age Security Pension, Guaranteed Income Supplement, Allowance** and **Allowance for the Survivor** *from federal government only (provincial income supplements should be reported in (h))*	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No
(f) Benefits from **Canada** or **Quebec Pension Plan**	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No
(g) Benefits from **Employment Insurance** *(total benefits before tax deductions)*	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No

(h) Other income from government sources, *such as provincial income supplements and grants, the GST/QST/HST credit, provincial tax credits, workers' compensation, veterans' pensions, welfare payments*	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No
OTHER INCOME: (i) Dividends, interest on bonds, deposits and savings certificates, and other investment income, *such as net rents from real estate, interest from mortgages. Do not include capital gains/losses.*	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No ○ Loss	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No ○ Loss
(j) Retirement pensions, superannuation and annuities, *including those from RRSPs and RRIFs. Do not include withdrawals from a pension plan or RRSP.*	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No
(k) Other money income, *such as alimony, child support, scholarships*	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No
TOTAL INCOME in 2005 from all sources	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No ○ Loss	○ Yes $ ☐ ☐☐☐ ☐☐☐ .00 ○ No ○ Loss
INCOME TAX PAID on 2005 income (federal, provincial and territorial)	$ ☐ ☐☐☐ ☐☐☐ .00	$ ☐ ☐☐☐ ☐☐☐ .00