EDIT AND IMPUTATION OF DEMOGRAPHIC VARIABLES ON THE CANADIAN TAXFILE

Edouard Auger, Statistics Canada

1. INTRODUCTION

The Canadian personal income tax file is annual and contains approximately 15,000,000 records. Each record represents a different tax return on which income, deductions and tax credits are recorded. In addition, geographic and a few demographic codes are included, the latter will be of interest here. The demographic variables discussed are the sex, marital status and year of birth codes.

Since the file is very large and the rates of missing and inconsistent data are low (see Table 2.1), very simple and relatively inexpensive methods had to be designed with the following two characteristics:

- the complete file is processed only once,
- each record is processed independently.

With a low error rate, the choice of an imputation method does not affect the final estimates significantly. However, it was felt desirable to preserve the distributions of the variables as a protection against a variation in the error rate or a change in the type of errors.

The probabilistic imputation methods adopted were based on empirical distributions of the variables to be imputed by categories of auxiliary variables. These methods were deemed as having the two required qualities (low cost system and preservation of distributions).

This paper will discuss the edit and imputation strategy adopted and some results based on tests runs for the 1979 New Brunswick taxfile (380,000 records).

2. EDITING

The format and content of administrative records are not generally under the control of the statistical agency as in the case of a survey. Only limited consistency checks can be done due to lack of related fields. And when inconsistencies occur, it is very difficult to decide how they should be resolved since the details of the administrative data collection and data processing methods are virtually unknown.

In this study, no consistency check could be done on the 'sex' field. The 'marital status' code could be checked with two spouse related fields. The 'year of birth' code could be checked with the 'old age security benefits' field and, in some cases, other income fields.

To decide on corrections to be made in the case of inconsistencies, the following assumptions were used:

- Income fields were deemed of better quality than demographic or other codes, because of better processing procedures by the tax agency.
- When "single" was declared as the marital status and the spouse related fields showed the presence of a spouse, it was assumed that the taxfiler was not married at the moment but had been married previously.

Table 2.1 gives the edit results from the test file. Some 1981 Census estimates are also presented for comparison purposes. These results indicate the fact that the error rates on the taxfile were very low, even lower than Census results.

3. IMPUTATION

3.1) Description of the Method

The probabilistic imputation method used consisted essentially of selecting values at random from empirical distributions obtained from valid and consistent values. These distributions took account of auxiliary variables which were best related to the variable being imputed. Different combinations of auxiliary variables were tested and the ones that appeared to provide the best predictors were selected, as described later.

The following sections describe the imputation tables for the three fields of interest including efficiency measures used to evaluate the procedures.

3.2) Year of Birth

Because of the large number of valid year of birth codes, it was separated in 6 groups in the imputation table. They represent the age groups (15-24, 25-34, 35-44, 45-54, 55-64, 65 and over). The age group (0-14) was not used in the imputation table because the proportion of that group in an imputation category was always negligible.

The imputation was done in two stages. Firstly, an age group was selected randomly from a distribution which depends on the value of the auxiliary variables described in Table 3.2.1. At the second stage, the actual year was imputed with equal probability.

To assess the efficiency of the procedure for imputation of missing/invalid codes, two measures were calculated imputing a new year of birth on every record already having a valid and consistent value.

$$\text{efficiency measure 1: } \frac{\sum_{i=1}^{5} P_{10i}\,(n_i)}{\sum_{i=1}^{5} n_i}$$

$$\text{efficiency measure 2: } \frac{\sum_{i=1}^{5} P_{15i}\,(n_i)}{\sum_{i=1}^{5} n_i}$$

where: $i=1,..,5$(categories of imputation),

P_{10i} = proportion of imputed year of birth codes within 10 years of the original value in category of imputation,

P_{15i} = proportion of imputed year of birth codes within 15 years of the original value in category of imputation,

n_i = actual number of missing/invalid codes in the category of imputation.

These gave the proportion of missing/invalid year of birth codes that should be imputed within 10 and 15 years of the true value. The results from the three trials are in Table 3.2.2.

TABLE 2.1 Edit Results

Item	Taxfile		Census
Code	Missing/ Invalid	Incon- sistent	Missing/Invalid
Year of birth	.01%	.03%	1.1%
Marital Status	.25%	.22%	1.3%
Sex	.19%	-	.8%

TABLE 3.2.1 Year of Birth Imputation Table

Old Age Security Benefits ≠ 0	
Old Age Security Benefits = 0	Marital Status='missing/invalid' Marital Status='widow' Marital Status='single' Marital Status='other'
	Number of imputation categories = 5

TABLE 3.2.2 Year of Birth Efficiency Measures

Item	1st trial	2nd trial	3rd trial	Mean
Efficiency measure 1 (10 years)	.6135	.6209	.6166	.6170
Efficiency measure 2 (15 years)	.7210	.7277	.7316	.7268

This last table indicates that although the method is simple, it provides a relatively efficient predictor of broad age classes.

3.3) Marital Status

There are five valid 'marital status' codes used in the imputation table (married, widow, divorced, separated and single). The structure of the imputation table is shown in Table 3.3.1.

To evaluate this imputation scheme, the same simulation method was used as with the year of birth imputation and a similar efficiency measure was calculated. The proportion of imputed codes equal to the original code was used in the measure.

TABLE 3.3.1 Marital Status Imputation Table

Spouses SIN*=0	Spouse*=blank	Year of birth (0-14 and 80-99)**	EXEMPT* ≤ BPE EXEMPT > BPE
		Year of birth (15-24)	EXEMPT ≤ BPE EXEMPT > BPE
		Year of birth (25-34)	EXEMPT ≤ BPE EXEMPT > BPE
		Year of birth (35-44)	EXEMPT ≤ BPE EXEMPT > BPE
		Year of birth (45-54)	EXEMPT ≤ BPE EXEMPT > BPE
		Year of birth (55-64)	EXEMPT ≤ BPE EXEMPT > BPE
	Spouse ≠ blank		
Spouses SIN≠0	Spouse = blank		
	Spouse ≠ blank		
	Total number of imputation categories = 15		

* Spouses SIN = Spouse's Social Insurance Number
* Spouse = 4 first character of the spouse's name
* EXEMPT = Total personal exemptions
* BPE = Basic personal exemption.
** Year of birth (65-79) corresponds to age group (0-14) and is always imputed to "single"

The results are as follows:

TABLE 3.3.2 Marital Status Efficiency Measures

Item	1st trial	2nd trial	3rd trial	Mean
Efficiency measure	.815	.818	.818	.817

The results were very good since 4 out of 5 times the correct value was imputed. The reason for this success was that 65% of the missing 'marital status' codes were in one imputation category (Spouse's SIN=0, Spouse=blank, year of birth (55-64), EXEMPT ≤ BPE) and in this category 98.7% of the valid codes were "single".

3.4) Sex

To impute missing/invalid sex codes, an imputation table of 15 categories was used. The auxiliary variables used were the 'marital status' code and classes of 'total income'.

An efficiency measure identical to the one calculated for the imputation of 'marital status' codes was calculated. The results are Table 3.4.1

Although the results were not as good as for the imputation of marital status codes, well over half of the cases were imputed correctly. The method performed much better for high and low incomes* (which have more influence on the estimates than average incomes do).

4. IMPACT OF IMPUTATION ON INCOME ESTIMATES

4.1) Introduction

So far, the inherent qualities and the efficiency of the methods have been discussed. This section will discuss the induced variability and other effects of the imputation procedures on the final estimates. The estimates will be:

- N: Number of taxfilers showing a non-zero value,
- Mean of the N taxfilers
- Median of the N taxfilers.

For illustrative purposes two types of income were chosen: old age security benefits and employment income. The former showed a larger proportion of missing sex codes, the latter involved a large number of taxfilers and highly variable income values. These represented two extreme cases. Imputation was repeated three times on the study file and estimates were compared.

4.2) Old Age Security Benefits (OASBEN)

The estimates by sex groups for this field are in the Table 4.2.1

Although the proportion of missing sex codes was high among old age security beneficiaries, (1.18% of all beneficiaries) the three trials produced estimates of N very close to each other. The means and medians were virtually identical, but this was expected since the range and variability of this income field are very limited.

4.3) Employment Income (EMPINC)

Tables 4.3.1, 4.3.2 and 4.3.3 describe the results for this income field. Estimates by sex and marital status groups are given.

Once again, although the employment income values are more variable, the three trial estimates were very stable for N and the means and medians were again virtually identical.

From the table of missing codes, (Table 4.3.1) it can be seen that records with missing codes usually had smaller than average employment income values. This means that the imputation had the effect of diminishing slightly the means and medians although the differences observed were small as shown in Table 4.3.4.

5.0 CONCLUSION

The imputation methods described in this paper are based on empirical distributions of the variable to be imputed by categories of auxiliary variables.

These methods were chosen because of their simplicity, low cost and the fact that they preserve distributions. Those criteria take into account the small rates of error, huge size of the file and possible unforeseen changes in the error occurrence.

The efficiency of the methods was also discussed and it was concluded that the method performed satisfactorily and had minor effects on the final estimates. This method is currently being implemented.

BIBLIOGRAPHY

Fellegi, I.P. and Holt, D.A., "A SYSTEMATIC APPROACH TO AUTOMATIC EDIT AND IMPUTATION". Journal of the American Statistical Association, Vol.71, March 1976, pp.17-35.

Sande, I.G., "IMPUTATION IN SURVEYS: COPING WITH REALITY". Survey Methodology/Techniques d'enquete, Vol.7, June 1981, pp.21-43.

FOOTNOTE

* High incomes were found among a proportionally larger group of males and the contrary was true for low income.

TABLE 3.4.1 Sex Efficiency Measures

Item	1st trial	2nd trial	3rd trial	Mean
Efficiency measure	.641	.644	.644	.643

TABLE 4.2.1 Old Age Security Benefits Estimates

OASBEN by Sex		Before Imputation	After Imputation		
			Trial 1	Trial 2	Trial 3
Male	N	16,253	16,330	16,328	16,326
	Mean	1,958.43	1,958.26	1,958.21	1,958.58
	Median	2,074	2,074	2,074	2,074
Female	N	11,325	11,577	11,579	11,581
	Mean	1,982.5	1,982.85	1,982.91	1,982.38
	Median	2,074	2,074	2,074	2,074
Missing	N	329	-	-	-
	Mean	1,980.47	-	-	-
	Median	2,074	-	-	-

TABLE 4.3.1 Missing Codes for Employment Income

	N	% of EMPINC+	*	Mean	Median
Sex="Missing"	286	.09%	.19%	5,977.67	2,581.5
Marital Status= "Missing"	700	.23%	.25%	2,481.89	1,517.5
Year of Birth= "Missing"	13	.004%	.01%	1,414.31	642

*Percentage of missing codes on the study file.
+Percentage of taxfilers with employment income with a missing value.

TABLE 4.3.2 Employment Income By Sex

		Sex	
		Male	Female
Trial 1	N	188,798	114,990
	Mean	12,130.7	6,367.65
	Median	11,077	5,043
Trial 2	N	188,814	114,974
	Mean	12,130	6,368
	Median	11,076.5	5,044
Trial 3	N	188,805	114,983
	Mean	12,130.3	6,367.92
	Median	11,077	5,043

TABLE 4.3.3 Employment Income by Marital Status

		Marital Status		
		Married	Single	Other
Trial 1	N	193,792	88,253	21,743
	Mean	11,739.9	6,225.84	9,102.13
	Median	10,507	4,394	7,909
Trial 2	N	193,792	88,259	21,737
	Mean	11,739.9	6,226.1	9,102.21
	Median	10,507	4,395	7,909
Trial 3	N	193,798	88,257	21,733
	Mean	11,739.7	6,226.62	9,100.95
	Median	10,507	4,395	7,908

TABLE 4.3.4 Comparison of Estimates from Unimputed and Imputed Data for Employment Income

Item		Unimputed Data		Imputed Data*	
		Mean	Median	Mean	Median
Sex	Male	12,141.2	11,091	12,130.3	11,076.8
	Female	6,373.16	5,052	6,367.86	5,043.7
Marital Status	Married	11,748.1	10,517.5	11,739.83	10,507
	Single	6,269.18	4,449	6,226.19	4,394.7
	Other	9,158.02	7,974	9,101.63	7,908.7

* Mean of 3 trials