THE POTENTIAL OF THE CANADIAN PERSONAL INCOME TAX FILE AS A SOURCE OF FAMILY DATA

Édouard Auger, Statistics Canada

1. INTRODUCTION

In recent years, there has been a persistent demand among the user community for more frequent small area family income statistics. The decennial Census is currently the only source that provides family income statistics for small areas. The Survey of Consumer Finances provides annual family income data but only at the provincial level.

The Canadian Personal Income Tax file, even if it is individually based, can be viewed as a possible source for these kinds of statistics. In general, the information available on the records on spouse and also on dependent children is key to the development of family income data. This means that a family-based file will be created from an individually-based administrative file. This is viewed as a very interesting and challenging methodological problem.

This paper presents a brief description of the theoretical potential and of the overall goals of the project and a few preliminary results. The overall objective is to generate discussion on this particular kind of work.

The first goal of the project is to develop a tax-based series close to the census family concept (see Section 3) and then to model the series into a population-based series. Results from a preliminary study on the 1980 tax file of Prince Edward Island are presented and compared to results from the 1981 Census of Population (which collected 1980 Income).

These comparisons show that over 85% of husband-wife families are covered by the tax records (over 90% for families with husbands between 15-64 years of age). While these results are very encouraging there are still some gaps. For example, families with couples over 65 years of age show much lower coverage. High income families are also poorly covered, although this may be due to differences between census and tax file reporting.

These and other results are expanded in this report and a discussion of the future outlook of the project is also presented.

2. DATA CONTENTS

There are numerous fields on the tax records that pertain to spouses and dependent children. A description of these fields and their possible uses and drawbacks will indicate what kind of family units can be reconstructed from the Income Tax file.

The cornerstone of this project is the presence of a field on a married taxfiler's record containing the SIN (Social Insurance Number) of her(his) spouse. This field permits the direct matching of spouses. There are two possible links (the SIN on the husband's record to the spouse's SIN on the wife's record and vice versa). A potential problem is that the reporting of the spouse's SIN is not compulsory. There are other fields that can be used to assess the validity of the SIN-spouse's SIN links or used to match married filers that did not report their spouse's SIN:

Surname: They do not have to be the same within a couple.

"First name of spouse" field: This field can be used in connection with the first name of the filer.

Exemption for married: This field is related to the net income of the spouse. It can be used for one of two purposes, to assess the validity of matched couples or to impute non-reporting dependent spouses (with no income).

Postal address (including postal code): In general this field contains the home address. However addresses may be different for married spouses living together, if non-residential addresses are used in filing a tax return.

There is no direct information on children on a parent's tax return and in the case of children who file their own returns there is no exact link. However, a number of fields are of use in identifying families and also in matching children to parents:

Surname: This field can be compared to the surnames of the two parents.

Postal Address: This field can also be compared to the postal addresses of the two parents.

Age difference with mother.

Exemption For Children: This field is again related to the net income of the dependent. It can then be used for validity evaluation of children matched to parents (children under 21 years of age only) and also for the imputation of non-reporting dependent children (with no income).

Child Tax Credit: The value of the credit is determined by the number of children under 18 years of age and the total net income of the two parents. It can then be used for similar purposes as the exemption for children.

Family Allowance, Child Care Expenses: These fields can also help detect the presence of dependent children.

The matching procedure for children will probably use the surname and postal address information. It will probably be less efficient than the matching of spouses.

Another very useful field in this study is the marital status code,which can be used to define couples as married or previously married or define single young filers as potential children. The main drawback of this code is that it is as of December 31st and thus there can be a marital change between that day and the filing date (up to April 30).

3. FAMILY CONCEPTS AND PROJECT GOALS

Since the Income Tax records do not represent the entire Canadian population, there are two kinds of series that can be developed:

A) A hybrid tax-based family income indicator used only to be modelled into a population-based series (similar to Census or Survey of Consumer Finances series).

B) A new family income indicator developed solely from the Income Tax records.

The modelled series (approach A) will be attempted first for reasons of consistency and comparability. To pursue this, a tax-based series highly correlated to a population-based series must first be developed.

There are two widely-used family concepts at Statistics Canada:

Census Family: A couple or parent(s) and never-married children, if any, either own or adopted, who live in the same household.

Economic Family: All relatives by birth, marriage, or adoption, who live in the same household.

Since the modelled series (approach A) has been chosen, one of these family concepts has to be used. The Census family concept was chosen because of the available information on spouses and of the information on dependent children present on the Tax file. Furthermore, the Census family concept is included in the Economic family concept. However, the Census family concept is not served perfectly by the Tax file since there is no exact link between children and their parents.

To summarize, the first goal of this project is to create a tax-based series highly correlated to a population-based series using the Census family concept. The second goal is to model this tax-based series into the population-based series. The availability of the postal code will allow the development of these series for small geographic areas.

4. PRELIMINARY STUDY

Since the matching of spouses using the SIN-spouse's SIN link is the essence of the project, a preliminary study was done to assess this matching process and, more generally, the potential of the Tax file as a source of family income statistics. The study file was the 1980 Prince Edward Island (PEI) Tax file. The 1980 PEI file was chosen because of its small size (60,000 observations) and because it could be compared to the 1981 Census. The two matching possibilities--SIN (male) to spouse's SIN (female) and vice versa--were used to get as many records matched as possible. The four parts to this study included:

A study of the performance of the matching procedure;

A study of the matched couples;

A study of the remaining unmatched records; and

Comparisons with the 1981 Census.

1) **Performance of the Matching Procedure**

Results showed that 83.3% of married males and 93.5% of married females were matched. (In Table 4.1, results from the matching procedure, by sex, presence of spouse's SIN and marital status are shown). Since fewer married females than males usually file, this can explain the difference between the two sexes. The high matching percentage for married females is a very encouraging result.

The matching percentages in the "other" category (divorced, separated) with a non-zero spouse's SIN field were over 50%. These percentages should increase when an all-Canada file is used, since people tend to relocate following a separation or a divorce [1].

2) **Matched Couples**

2.1) Married Couples

To assess the validity of the married couples formed, postal codes and surnames were compared, age difference between spouses was considered and finally, the occurrence of double links--SIN (male) = spouse's SIN (female) and vice versa--was investigated [2]. These results show a very high degree of consistency between spouses for these fields (see Tables 4.2 to 4.5).

There were 16,108 couples (85.4% of all married couples) with double links, equal postal code and equal five first characters of surnames. Then there were 2,564 couples (13.6% of all married couples) with only one difference (different postal codes, different five first characters of surnames or single link-- 41.6% of these cases involved a missing postal code or spouse's SIN. In other words, 99.0% of all married couples had, at the most, one difference. Remarkably, there was only one married couple with different postal codes (no missing postal codes), "single" link (no missing spouse's SIN) and different five first characters of surname.

The few couples with an age difference of over 20 years were investigated and did not show high rates of differences in the auxiliary fields compared. Mostly it seemed that errors may have occurred in the year of birth field [3].

Differences for these fields are not a sure sign of an invalid match, but consistencies like those experienced here do indicate a high rate of valid couples.

2.2) Couples With At Least One Not Presently Married Spouse

There were 881 matched records with at least one not presently married spouse.

About 40% of these records were from separated spouses. A minority of the separated couples had the same address. Some other couples had one spouse widowed and the other one married. From a preliminary study on the Prince Edward Island 1982 file, most of the married filers in this kind of link were not alive. There were also couples formed of previously married filers (divorced, divorced and remarried, etc.) and of filers that seemingly got married between December 31st (reference date for the marital status) and the time of filing.

3) **Remaining Unmatched**

As far as age and marital status are concerned, a large proportion of the unmatched filers were young and single (see Tables 4.6 and 4.7). Also most of these did not show the presence of dependents.

There was a large concentration of married males (45 years old and over) that were not matched to females. This can be explained by the fact that females in this age group had a lower labour force participation rate than younger females [4] and also were less likely to file returns to claim the Child Tax Credit. In fact, 80% of males over 45 years of age showed the presence of a dependent spouse with no income.

There was also a larger concentration of wi-

dowed, separated and divorced females than males.

4) Comparisons with the 1981 Census

The data for husband-wife families were compared to the 1981 Census data (income for 1980). The tax data, at this point, include only families where both spouses filed a tax return. Also, income reported by children is not part of the tax family income but is included in the Census family income and the total income definitions are not exactly the same for the two sources. [5] Even if the ultimate goal of the project is to produce family data at the small intraprovincial area (Census divisions, federal electoral districts), results were only compared at the provincial level, since the small area geography of the tax file is still plagued with mailing address problems. It was felt undesirable to mix geography and family estimation problems at such an early stage of the project.

Overall, the coverage of the matched married couples was higher for younger than older families and for middle income compared to extreme income families (see Tables 4.8 and 4.9). In fact, for couples with the husband less than 55 years old, the coverage of the tax data was 86% of the census figure.

In the study of the unmatched records, the possibility of having couples where only one spouse filed a return was addressed. To estimate the number of these couples, the number of unmatched married filers with the presence of a dependent spouse with no income was used. Tables 4.10 and 4.11 update the two previous tables using the estimate of couples with only one reporting spouse. Overall, the coverage of the taxfiler families increased from 70.8% to 79.1%. For the families with husbands less than 55 years old, the coverage increased to 90%. The (husband) age group (55-64) increased the most from the addition of couples with one reporting spouse; its representation went up by 22.4%. The representation of the (husband) age group (65&+) remained very low. The representation of the very low-income (<$5000) families went up by 33.5%, the biggest jump of all income intervals. The representation of the high income and of the interval ($5000-$9999) remained low.

Tables 4.12 and 4.13 present the distribution of the remaining unmatched records for married filers by age and family income level. Over half of the remaining married males were aged 65 and over. However, these could only have increased to 53% the representation of families with husbands aged 65 and over. The small number of high income earners in the remaining unmatched filers indicates that the low representation of high income remains a question. The difference may be due to differences in reporting from the two sources.

When the number of remaining unmatched married males were added to get a minimum estimate of the number of married couples (Tax) [6], it was found that the Tax count amounted to 86.3% of the total husband-wife count (92.1% of the families with husbands between 15 and 64 years of age).

5. PROJECT OUTLOOK

The preliminary results gathered so far are very encouraging for husband-wife families.

It now remains to examine the full potential of the Income Tax File — the next step in this project. The different items to be investigated will include:

- a second matching of couples using other information than the SIN and spouse's SIN (mostly, this step will match couples that did not report their spouse's SINs, using names, addresses, and exemption fields),
- the matching of reporting children to family units using mostly names and addresses; and
- the imputation of non-reporting dependent family members using auxiliary information like the exemption fields and the Child Tax Credit.

After that, tax-based series will be assessed on a larger scale to verify the results found in the preliminary studies, at the national level. The last part of the project will involve studying different techniques for the purpose of modelling the tax-based series into a population-based series.

FOOTNOTES

[1] For all divorce suits brought by females from 1969 to 1979, only 61.2% of the males were still living at the time of the suit in Prince Edward Island. (McKie D.C., et al 1983).

[2] Investigation of the consistency between exemptions and net income and of the first name-spouse's first name link, and comparison of addresses could not be done with the 1980 study file at hand, since these fields were not present.

[3] In 1980 the century of birth was not included in the year of birth field and it seemed that some filers put their age instead of their year of birth. The century of birth is now included since 1982.

[4] Participation rate = Labour Force/total population. Participation rates in Prince Edward Island (1980) were: 65.9% for females aged (20-24), 61.4% for females aged (25-44), and 43.1% for females aged 45-64.

[5] The major sources of income are part of the two definitions, but non-taxable sources of income (welfare payments, war veteran's allowances, etc.) are not included in the tax definition. Furthermore, sources of income like employment expenses and capital gains are included in the tax definition but not in the census definition.

[6] This is a minimum estimate because it uses the hypothesis that all the remaining married females can be matched to one of the remaining married males to form a family.

BIBLIOGRAPHY

McKie, D.C., Prentice B., and Reed P., "Divorce: Law and the Family in Canada." Statistics Canada, Ottawa, February 1983, Catalogue No.89-502E, 280p.

"Social Concepts Directory," Statistics Canada, Ottawa, December 1980, Catalogue No.12-560, 140p.

Jung A, and Turner R., "Family Statistics From the Personal Income File," Statistics of Income and Related Administrative Record Research: 1983, Department of Treasury IRS, October 1983, pp.21-27.

TABLE 4.1 NUMBER OF FILERS AND MATCHED FILERS IN CATEGORIES OF SEX, PRESENCE OF SPOUSE'S SIN AND MARITAL STATUS

NO. OF FILERS (NO. MATCHED)	SPOUSE'S SIN	NO. OF FILERS (NO. MATCHED)	MARITAL* STATUS	NO. OF FILERS (NO. MATCHED)
M A L E S				
34,991 (19,714) **56.3%	Missing	11,894 (347) 3.1%	Married	747 (307) 41.1%
			Widowed	385 (0) 0%
			Single***	10,143 (21) .2%
			Other	619 (46) 7.4%
	Present	23,097 (19,340) 83.7%	Married	22,111 (18,743) 84.8%
			Widowed	109 (31) 28.4%
			Single	87 (58) 66.7%
			Other	790 (508) 64.3%
F E M A L E S				
31,359 (19,701) 62.6%	Missing	10,216 (405) 4.0%	Married	480 (295) 61.5%
			Widowed	1,801 (22) 1.2%
			Single	6,978 (17) .2%
			Other	957 (71) 7.4%
	Present	21,143 (19,296) 91.3%	Married	19,759 (18,636) 94.3%
			Widowed	298 (97) 32.6%
			Single	57 (33) 57.9%
			Other	1,029 (530) 51.5%

* Missing codes were included in the "other" category.
** Since some filers were matched more than once and only the actual number matched was used (without reference to the number of times, a record was matched), different total matched records for both sexes were found.
*** These do not necessarily have an invalid marital status since this is as of December 31st and they could be married when filing.

TABLE 4.2 COMPARISON OF POSTAL CODES (PC) BETWEEN MARRIED SPOUSES

	PC (MALE) = PC (FEMALE)	PC (MALE) ≠ PC (FEMALE)	AT LEAST ONE PC MISSING	TOTAL
NO. OF COUPLES	16,926 (89.8%)	922 (4.9%)	1,005 (5.3%)	18,853

TABLE 4.3 COMPARISON OF FIVE FIRST CHARACTERS OF SURNAME BETWEEN MARRIED SPOUSES

	FIVE FIRST CHARACTERS SURNAME (FEMALE) = FIVE FIRST CHARACTERS OF SURNAME (MALE)	FIVE FIRST CHARACTERS SURNAME (FEMALE) ≠ FIVE FIRST CHARACTERS OF SURNAME (MALE)	TOTAL
NO. OF COUPLES	18,609 (98.7%)	244 (1.3%)	18,853

TABLE 4.4 AGE DIFFERENCE BETWEEN MARRIED SPOUSES

	AGE DIFFERENCE: ABSOLUTE VALUE						
	0-5 YEARS	6-10 YEARS	11-15 YEARS	16-20 YEARS	21-25 YEARS	26+ YEARS	TOTAL
NO. OF COUPLES	14,140	3,667	768	194	63	21	18,853
% OF COUPLES	75.0	19.5	4.1	1.0	0.3	0.1	100

TABLE 4.5 OCCURRENCE OF DOUBLE LINKS

	DOUBLE LINKS (SIN (M) = SPOUSE'S SIN(F) AND VICE VERSA)	SINGLE LINK (THE TWO SPOUSE'S SIN ≠0)	SINGLE LINK (ONE SPOUSE'S SIN = 0)	TOTAL
NO. OF COUPLES	18,093 (96.0%)	164 (.9%)	596 (3.2%)	18,853

TABLE 4.6 AGE DISTRIBUTION OF UNMATCHED MALES BY MARITAL STATUS

AGE	MARRIED	WIDOWED	DIVORCED	SEPARATED	SINGLE	MISSING	TOTAL
14&-	0	0	0	0	144	13	157
15-24	133	0	5	22	6,375	67	6,602
25-34	268	6	113	109	1,943	1	2,440
35-44	302	15	117	88	563	2	1,087
45-54	637	45	104	68	505	0	1,359
55-64	1,294	124	43	52	376	0	1,889
65&+	1,377	273	26	22	245	3	1,946
TOTAL	4,011	463	408	361	10,151	86	15,480

TABLE 4.7 AGE DISTRIBUTION OF UNMATCHED FEMALES BY MARITAL STATUS

AGE	MARRIED	WIDOWED	DIVORCED	SEPARATED	SINGLE	MISSING	TOTAL
14&-	0	1	0	0	64	6	71
15-24	136	5	15	87	4,711	80	5,034
25-34	290	66	217	238	1,214	1	2,026
35-44	241	129	203	158	265	2	998
45-54	304	294	111	114	196	2	1,021
55-64	302	558	55	43	195	1	1,154
65&+	114	927	31	16	340	5	1,433
TOTAL	1,387	1,980	632	656	6,985	97	11,737

TABLE 4.8 FAMILY COUNTS BY AGE GROUP OF THE HUSBAND AND WIFE

AGE	HUSBAND		WIFE	
	MARRIED COUPLES (TAX)	HUSBAND-WIFE FAMILIES (CENSUS)	MARRIED COUPLES (TAX)	HUSBAND-WIFE FAMILIES (CENSUS)
24&-	1,240 (84.6%)	1,465	2,275 (86.5%)	2,630
25-34	6,234 (91.4%)	6,820	6,729 (88.5%)	7,600
35-44	4,837 (85.8%)	5,640	4,671 (88.1%)	5,300
45-54	3,424 (79.2%)	4,325	3,059 (73.3%)	4,175
55-64	2,128 (54.1%)	3,930	1,585 (41.7%)	3,800
65&+	990 (22.2%)	4,455	534 (17.1%)	3,125
TOTAL	18,853 (70.8%)	26,630	18,853 (70.8%)	26,630

TABLE 4.9 FAMILY COUNTS BY INCOME INTERVALS

INCOME INTERVALS	MARRIED COUPLES (TAX)	HUSBAND-WIFE FAMILIES (CENSUS)
<5,000	551 (67.6%)	815
5,000 - 9,999	1,883 (48.4%)	3,890
10,000 - 14,999	3,846 (77.5%)	4,960
15,000 - 19,999	4,112 (88.7%)	4,635
20,000 - 24,999	3,327 (86.6%)	3,840
25,000 - 29,999	2,148 (69.5%)	3,090
30,000 - 34,999	1,233 (66.5%)	1,855
35,000 - 39,999	694 (49.6%)	1,400
> 39,999	1,059 (49.0%)	2,160
TOTAL	18,853 (70.8%)	26,630

TABLE 4.10 FAMILY COUNTS BY AGE GROUP OF THE HUSBAND, ADJUSTED

AGE OF* HUSBAND	MARRIED COUPLES (TAX) + (ESTIMATE OF COUPLES WITH ONE SPOUSE REPORTING)	HUSBAND-WIFE FAMILIES (CENSUS)
24 &-	1,321 (90.2%)	1,465
25-34	6,407 (93.9%)	6,820
35-44	5,062 (89.8%)	5,640
45-54	3,696 (85.5%)	4,325
55-64	3,007 (76.5%)	3,930
65 &+	1,361 (30.5%)	4,455
TOTAL	21,054 (79.1%)	26,630

* There were 130 couples with only the wife reporting; for these the age of the wife was used as a proxy for the age of the husband.

TABLE 4.11 FAMILY COUNTS BY INCOME INTERVALS, ADJUSTED

INCOME INTERVALS	MARRIED COUPLES (TAX) + (ESTIMATE OF COUPLES WITH ONE SPOUSE REPORTING)	HUSBAND-WIFE FAMILIES (CENSUS)
<5,000	824 (101.1%)	815
5,000 - 9,999	2,388 (61.4%)	3,890
10,000 - 14,999	4,323 (87.2%)	4,960
15,000 - 19,999	4,410 (95.1%)	4,635
20,000 - 24,999	3,506 (91.3%)	3,840
25,000 - 29,999	2,269 (73.4%)	3,090
30,000 - 34,999	1,335 (72.0%)	1,855
35,000 - 39,999	781 (55.8%)	1,400
> 39,999	1,218 (56.4%)	2,160
TOTAL	21,054 (79.1%)	26,630

TABLE 4.12 AGE DISTRIBUTION BY SEX OF THE REMAINING UNMATCHED MARRIED FILERS

AGE	MALE	FEMALE
15-24	62	126
25-34	131	254
35-44	111	207
45-54	194	275
55-64	433	284
65 &+	1,009	111
TOTAL	1,940	1,257

TABLE 4.13 INCOME DISTRIBUTION BY SEX OF THE REMAINING UNMATCHED MARRIED FILERS

INCOME INTERVALS	MALE	FEMALE
<5,000	363	732
5,000 - 9,999	545	319
10,000 - 14,999	388	121
15,000 - 19,999	235	52
20,000 - 24,999	141	24
25,000 - 29,999	59	3
30,000 - 34,999	68	3
35,000 - 39,999	35	0
> 39,999	106	1
TOTAL	1,940	1,257