Research Note/Note de recherche

Testing The Use of A Hybrid Regionalisation Scheme for Confidential Tax-Filer Data

Carl G. Amrhein
Department of Geography
University of Toronto
Toronto, ON M5S 1A1

In a recent article (Bailly and Coffey 1994) make a strong case for the importance of scale effects in applied regional science. In a plea for a more applied regional science, three principles are presented, including:

> *Principle 3.2: When treating space and time, it is necessary to recognize the effects of employing varying scales of analysis.* "... the choice of *spatial* scale can exert an important influence upon our work." "Indeed, the flexibility of spatial scale associated with the concept "region" needs to be treated carefully. Consider, for example, the practical implications of formulating and implementing regional development policies in Canada, a vast country with 10 traditional administrative regions, and in Switzerland, a tiny country in which regional policy is administered in over 100 regions!"

The authors might have gone on to address two related issues, the arrange-

This pilot project would not have been possible without the assistance of the staff at the Small Area and Administrative Data Division (SAADD) of Statistics Canada. Due to the extreme sensitivity of the original data, great care was taken to ensure confidentiality. At no time did the author have access to the tax records. All computing based on the original data was performed by SAADD personnel, with the results reviewed by SAADD management before release. The author had access only to summary statistics for the 47 customized zones. The author alone is responsible for the views and conclusions contained herein, and all errors in this paper. The author wishes to express particular appreciation to Dr. John Leyes, Director of SAADD. Funding for this project was provided by the Social Science and Humanities Research Board of the University of Toronto. Additional work was conducted while the author was in residence at the U.S. National Center for Geographic Information and Analysis at the Department of Survey Engineering, University of Maine at Orono.

Canadian Journal of Regional Science/Revue canadienne des sciences régionales, XVII:2 (Summer/été 1994), 259-274.
ISSN: 0705-4580 Printed in Canada/Imprimé au Canada

ment of data from different sources (referred to as the zoning problem), and the administrative restrictions placed on sources of data (usually lumped together as confidentiality requires). As the demand for higher resolution data increases, both due to the need for more frequent and more detailed data, analysts are forced to demand alternative sources of data. In Canada, this demand has created a "market" for various administrative data sources. However, confidentiality constraints often require that these data be aggregated prior to their release. As in the use of census data, it is important to establish at the outset that the aggregation of these data do not introduce unacceptable levels of error into the analysis.

Two related issues underlay the analysis in this paper: first, the resolution (or scale) of the original data and, second, the errors introduced into spatial analysis when other than the original, micro-level data are used. Many believe that better analytical results are obtained with higher resolution (that is, smaller and smaller reporting units) data. Evidence of this belief is contained in the increasing use of micro-data (Fox et al. 1989; Herzog et al. 1992; Miron 1992; Shumway 1992; Waldorf 1992). In the limit, then, the best results are to be obtained when the observations in the data set represent individual responses. However, with the exception of special surveys, socio-economic data are usually reported at some level of aggregation. In most cases, the aggregates are defined by either postal code geography (blocks, routes, sortation zones, etc.) or census geography (census tracts, divisions, etc.). While there has been very dramatic evidence reported on the possible range of values for certain statistics that might (Openshaw and Taylor 1979) or might not (Amrhein and Flowerdew 1992, Amrhein 1992) arise in the use of aggregated data, the evidence is derived from data sets that, at best, represent small area aggregates. What has been lacking is a data set comprised of individual responses that can then be systematically aggregated for subsequent demographic analysis.

As mentioned above, it has been suspected, for many years, that the census geography that underlies much of the demographic analysis in social sciences must have some effect (that is, introduce error) on the numerical and interpretative results obtained (see, Shaw 1985). While there are many ways in which such error might be introduced into spatial analysis (see, Deichmann, Goodchild and Anselin 1992), of particular concern here is the effect of the arrangement of continuous space into defined regions for purposes of data reporting. This effect has been referred to in various ways in different disciplines including the Modifiable Areal Unit Problem (MAUP), the spatial aggregation effect, and the ecological fallacy (see, Dudley 1991, Fotheringham and Wong 1991, Amrhein 1994 and Wrigley 1994 for recent reviews of these problems). Furthermore, while there might be no theoretical reason to tolerate such error (see, Tobler 1991), the vast majority of the social data used will be available only at some level of aggregation, and the error must be managed if analytical results are to be improved.

This paper presents a study designed to simultaneously exploit the information contained in the Income Tax Filer Data set maintained by the SAAD Division of Statistics Canada, and to add another empirical example to the literature searching for the existence, nature, and extent of the aggregation effect.

The next section of this research note describes the data set, its origins, and the variables available for analysis. In addition, the customized regionalisation used in the analysis will be described. The third section will compare statistics from the original data set with the same statistics derived from the customized regionalisation. Any potential aggregation effects will be noted. Next, the aggregated data will be analyzed. A brief demographic analysis of the families in the sample will be conducted. In this section, comparisons over time will be made with respect to particular variables in the data set, as well as a discussion on income dynamics and a section on the relationships between variables (Product Moment Correlations) in the data set. *These sections are not meant to act as a comprehensive demographic analysis of the data or to search for new insights on population dynamics*, but rather to check the usefulness of the data and to suggest the significant potential that the data present in terms of possible future and more comprehensive analyses. The final section will summarize the findings of this pilot study.

The Data Set

The Small Area and Administrative Data Division of Statistics Canada collects a variety of data sets, including automobile registration, family allowance payments, Canada Pension payments, and the Income Tax-Filer Longitudinal Sample. The Tax Filer Data (TFD) set was started in 1982 as a ten percent sample of income tax records. The individuals in the original sample have been tracked in subsequent years. Individuals leave the annual sample through death, immigration out of Canada, or not filing a return. This data set, in certain respects, is similar to the unemployment insurance data set used in several well-known migration studies (for example, Courchene 1974; Grant and Vanderkamp 1976). The tax-filer data set used here contains less information on work history, but more information on family structure and the income dynamics of spouses. Both data sets experienced similar constraints imposed by confidentiality requirements.

Currently, five years of data are available, including a migration record for the periods 1982-1983 through 1986-1987. Forty-two different socio-economic variables are extracted from each tax return. The sample used in this pilot study is a special tabulation of a portion of the original TFD set.

To generate a sample with a manageable number of observations, attention was restricted to south and central Ontario, roughly from a line between Sault St. Marie and Sudbury south to the Great Lakes (includes postal Forward Sortation Areas (FSAs) and Rural Postal Codes beginning with letters K, L,

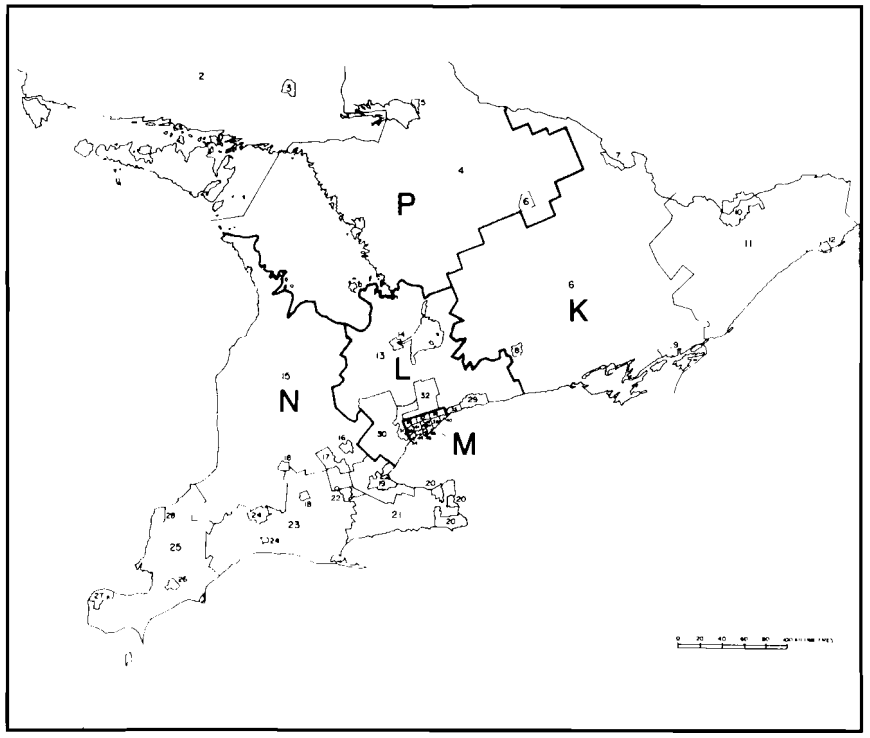

FIGURE 1 Study Area in Southern Ontario

M, N, and P -see Figure 1). From all the observations in this region, a two percent sample was drawn in which tax returns from marriages consisting of a male and female were joined to form a family record. In other words, the pilot sample consists of traditional nuclear families containing a husband, wife, and possibly dependents. This sampling process produced 31,140 family records (60,800 individuals) in each year 1984, 1985, and 1986. The years 1982 and 1983 are excluded from the pilot study due to data incompatibilities that require extensive data processing to remove.

Each family record consists of three fields - family data, husband data and wife data. A subset of the forty-two available variables is selected for this analysis. These twenty-seven variables are defined as:

Family Variables

FSZ	family size, number of people in the family
FK17	number of children 17 years old or younger

FK18	number of children 18 years old or older
FWRK	number of workers in the family
FCOMP	family composition flag (not useful with aggregate data) = 1 if married with two filers = 2 if married with one filer = 3 if family is a common-law family
FY	total family income less eligible expenses
FERN	family earnings
FUI	family unemployment insurance compensation
FALL	family allowance benefits
FRT	family retirement income
FOTH	other family income

Husband (first letter H), and Wife (first letter W) variables

HBR, WBR	year of birth (not used due to data problems)
HDTH, WDTH	year of death (not used since nearly everyone is still alive)
HY, WY	income less eligible expenses
HERN, WERN	earnings
HUI, WUI	unemployment insurance compensation
HRT, WRT	retirement income
HFALL, WFALL	family allowance (declared by the husband, if working)
HOTH, WOTH	other income

Customized Regionalisation

In order to meet the requirements of confidentiality, the 31,140 individual family records must be aggregated, with only summary statistics reported. Since the data are located by the six character postal code, census geography is not easily utilized. Instead, a customized regionalisation scheme is adopted that aggregates zones using the first three characters of the postal code (FSA or rural postal zone). The aggregation criteria are designed to: preserve confidentiality, produce a manageable number of aggregates given the available resources (around fifty zones), and minimize the loss of information. This last criterion is operationalised by preserving the "ruralness" or "urban-ness" of the zones aggregated.

The aggregation process produces forty-six customized zones categorized as urban, rural, or fringe (in the case of zones surrounding the metropolitan Toronto region). A forty-seventh "region" contains the records, in each year,

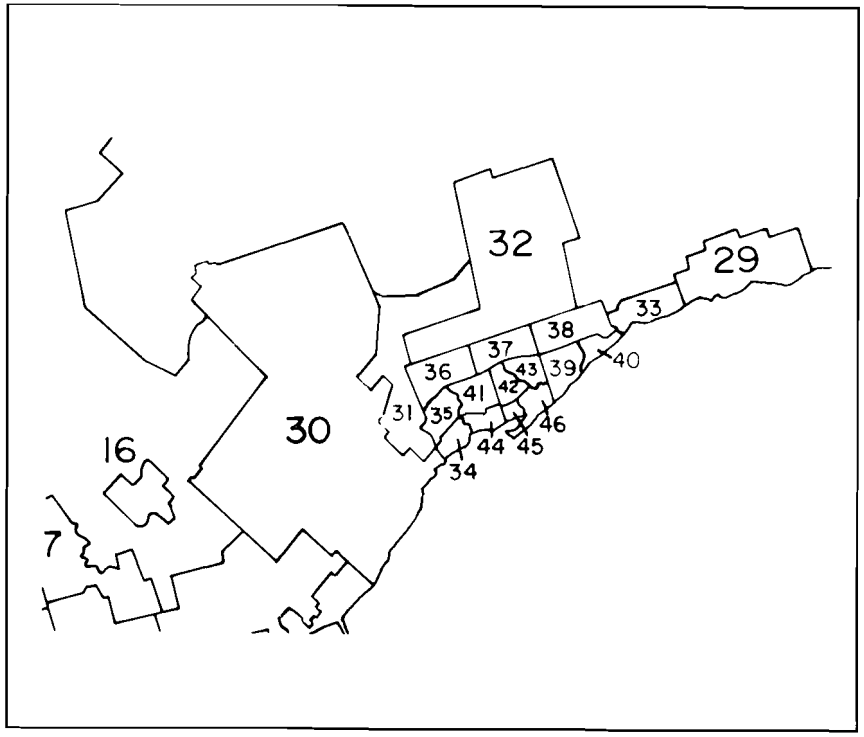

FIGURE 2 Metropolitan Toronto and Surrounding Areas

of families not resident in the Ontario region for a given year, but were resident in Ontario in one of the other years and thus eligible for inclusion in the sample. Figures 1 and 2 show the details of the aggregation process in which the 31,140 original observations are aggregated into 47 regions. Planning Divisions defined by the Planning Department of Metropolitan Toronto are used to define zones within the boundaries of the formally defined Metropolitan Toronto region.

Comparing the Customized Regions to the Original Data

Table 1 contains information on the number of observations in the sample in each year, and the number of observations in each category of region after aggregation.

As seen in the table, the population in the Urban plus Fringe categories comprises 69%, 72%, and 76% of the sample population in each year respectively. The increase, in part, is the result of a declining proportion of people

TABLE 1 Distribution of the Sample (and Frequency) Among the Region Types After Aggregation

Region Type	1984	1985	1986
Urban			
Metro Toronto	6,350	6,430	6,610
	(20.4)	(20.7)	(21.2)
Other Urban	10,430	10,840	11,400
	(33.5)	(34.8)	(36.6)
Urban Fringe	4,810	5,160	5,690
	(15.4)	(16.6)	(18.3)
Rural	6,440	6,690	7,000
	(20.7)	(21.5)	(22.5)
Other (outside Ontario)	3,110	2,020	420
	(10.0)	(6.5)	(1.3)
Total	31,140	31,140	31,120
	(100)	(100)	(100)

outside the Ontario region. Note that the proportion of the population in rural areas is also increasing slightly over time. By 1986, all but 420 families in the sample reside in Ontario.

Before examining the effect of the aggregation into forty-seven zones, it is important to check the internal integrity of the sampled data. One problem concerning the birth dates was alluded to earlier. Most of the birth years for the wives are either 1800 or 1900, suggesting a coding problem.

The three panels of Table 2 present a pair of more sophisticated tests. First, the sum of all the income fields on the tax return (FERN+FUI+FALL+FRT+FOTH with similar calculations for husband and wife) should very closely approximate, in turn, the sum of the FY, HY, and WY variables in the sample. When the number of family tax returns in a region is used as a weight, the sums of the weighted regional averages of all income fields should closely approximate the sum of weighted averages of the FY, HY, and WY fields. The second test compares the original 31,140 observations to the sum of the weighted averages. If the aggregation process is effective in preserving the information contained in the income variables, any difference between the original data and the weighted averages should be small and attributable to the random effect of accumulated random roundings of the regional data.

The results in Table 2a suggest that the sum of the observed values in the various income fields closely approximates the value in the total income field (FY, HY, or WY). However, two points are noteworthy. First, for the combined family information, there is an increasing difference over time between the sum of the income variables, and the total income variable, from 1.01% in 1984 to 2.67% in 1986. While 2.67% may not be enough to raise concerns, it does indicate that over time, variability is being introduced into the data in a

TABLE 2a Comparison of the Sum of Income Fields with HY, FY, or WY Based on the Original 31,140 Observations. Differences are Calculated as a Percentage of the Sum of Fields

	1984		1985		1986	
	Sum of Fields	HY, WY or FY	Sum of Fields	HY, WY or FY	Sum of Fields	HY, WY or FY
Family	1,370,903	1,357,084	1,484,675	1,460,983	1,529,031	1,488,326
difference (%)	1.01		1.60		2.67	
Wife	333,451	321,002	363,377	350,068	386,661	370,914
difference (%)	3.73		3.67		4.07	
Husband	947,751	935,292	1,023,141	1,009,800	1,038,796	1,015,682
difference (%)	1.31		1.30		2.22	

TABLE 2b Comparison of the Sum of Income Fields with HY, FY or WY Based on the Weighted Aggregated Results. Differences are Calculated as a Percentage of the Sum of Fields

	1984		1985		1986	
	Sum of Fields	HY, WY or FY	Sum of Fields	HY, WY or FY	Sum of Fields	HY, WY or FY
Family	1,374,239	1,358,995	1,485,612	1,462,400	1,526,906	1,487,300
difference (%)	1.11		1.56		2.59	
Wife	333,597	321,101	358,393	350,310	386,698	370,990
difference (%)	3.75		2.26		4.06	
Husband	963,432	936,329	1,024,874	1,018,300	1,039,115	1,016,300
difference (%)	2.81		0.06		2.20	

TABLE 2c Comparison of the Sum of Income Fields and HY, FY or WY Based on the Original 31,140 Observations with Weighted Aggregated Results. Table Values are Differences Calculated as a Percentage of the Value Associated with the Original 31,140 Observations

	1984		1985		1986	
	Sum of Fields	HY, WY or FY	Sum of Fields	HY, WY or FY	Sum of Fields	HY, WY or FY
Family (%)	0.24	0.14	0.06	0.10	0.14	0.07
Wife (%)	0.04	0.03	1.40	0.07	0.01	0.02
Husband (%)	1.66	0.11	0.17	0.84	0.03	0.06

manner that accumulates. If this trend continues into future years, at some point the total income variable may no longer be a suitable surrogate for the separate income variables. Secondly, the difference for the wife's income is consistently higher than either family or husband information. This difference may reflect the deductibility of child care expenses for working women, but requires further verification.

Table 2b provides a similar comparison using the weighted regional values from the forty-six aggregates. The differences closely match those found in the previous table. Of note, however, is that the change in the differences arising from the use of aggregate values is not consistent. Note, for example, the increase in the difference for husbands in 1984, the decrease to a very small value in 1985, and the stability of the 1986 value. The conclusion from these tables is that while, on aggregate, the weighted regional aggregates based on the total income variable are suitable surrogates in the absence of access to the original tax returns, there is a differential impact on the accuracy of the estimates that varies by year and sex. Caution is clearly warranted in extending these results to the entire sample across Canada.

Table 2c presents a comparison of the income variables contained in Tables 2a and 2b. Any difference indicated in Table 2c can only be attributed to the aggregation of the original 31,140 family tax records, to 47 weighted regional values weighted by the number of family tax returns found in the region. Fortunately, these percentage differences are quite small at the regional level, providing support for the further use of these customized regions.

In spite of these encouraging results, it should be recognized that aggregation affects vary with the statistics used in the analysis. Preliminary results in Amrhein and Flowerdew (1992) suggest that while simple statistics such as means are resistant to aggregation affects, more complicated statistics such as regression coefficients can demonstrate dramatic aggregation effects, even with data derived from purely random distributions. This finding is explicit in Amrhein's rules of aggregation (Amrhein 1994). Further evidence using British Census data is found in Amrhein and Flowerdew (1993).

In other words, the simple descriptive analysis performed here may suffer from little or no aggregation effect, but the results from a regression analysis using the same data may be very misleading. In addition, while the results are encouraging at the scale of South and Central Ontario, other analyses indicate that there may be greater variation among the forty-six regions that comprise Ontario in this study. This indicates a differential impact at the level of the customized regions that might be due to aggregating, a result that is contrary to the usual expectation of reduced variance in aggregated census data (see, Wrigley 1994).

Demographic Analysis

The results in the previous section are sufficiently encouraging to use the customized regionalisation for a rudimentary demographic analysis of the families in the sample. Several different pieces of information can be obtained from these data. It is important to note that this simple analysis is intended only to demonstrate the viability of this customized regionalisation as a means of accessing otherwise unaccessible data. This analysis is not intended to pro-

TABLE 3a Average of Regional Means

Variable	1984	1985	1986
FSZ	3.31	3.31	3.39
FK17	0.99	1.00	1.02
FY	43,980	47,028	47,748
FERN	35,289	37,374	38,223
FUI	844	778	722
FRT	2,676	3,165	3,565
HY	30,259	32,724	32,793
WY	10,430	11,378	12,004

TABLE 3b Weighted Average of Regional Means

Variable	1984	1985	1986
FSZ	3.33	3.33	3.35
FK17	1.02	1.02	1.03
FY	43,641	46,962	47,792
FERN	35,444	37,782	38,578
FUI	826	764	694
FRT	2,498	3,052	3,478
HY	30,068	32,702	32,656
WY	10,312	11,250	11,921

duce new insights into southern Ontario population dynamics.

Comparison Over Time

Given the customized regionalisation adopted, a reasonable first step is to compare selected variables over time. Table 3 compares eight variables over the three years. Table 3a presents the average of the regional means while Table 3b presents the weighted average of the regional means.

Overall, the similarity between Tables 3a and 3b is not surprising given the results in Amrhein (1994) and Amrhein and Flowerdew (1993). It is worth noting that while the trends for each variable in Tables 3a and 3b are the same over time, there are differences in the timing of changes. This is seen clearly in the case of variables FSZ and especially FK17. In the latter case, the regional average reaches 1.02 in 1986, but the weighted average never falls below 1.02 in any of the three years. While the absolute values differ little, a difference in the expected number of children per family of 0.03 (0.99 in the 1984 regional average versus 1.02 in the 1984 weighted average) translates into about 1,000 dependents for the sample, and many more for the province

(this is a 2% sample) or nation (the original data set is a 10% sample). Education and child-care planners would likely see the difference as increased transfer payments or an increased number of schools and child-care facilities. While Table 3 suggests that overall, there may be no dramatic differences between the regional and weighted average, there are subtle differences.

The overall regional economy described by these data show a number of desirable trends (recall that the Ontario economy in the middle 1980s was healthy and growing). All the income variables are rising except unemployment earnings. The decrease in unemployment earnings indicates a growing economy with increasing job opportunities (rather than increasing discouraged workers as seen in the early 1990s). Increasing retirement income and increasing numbers of dependents capture the aging adult population and the increasing number of births to the tail-end of the baby boom population.

These summary statistics, however, mask variation within the study region. An analysis looking at the distribution among the forty-six aggregates shows that the Northern sections of the study area are characterized by more children (higher values for FK17) and greater unemployment insurance income for both husbands and wives (and, obviously, families). In contrast, the Southern portion of the study area, especially Metro Toronto, exhibits fewer children, more workers per family (higher values of FWRK) and greater total earnings for husbands and wives (and families). For example, zone two (a rural northwestern area including Elliot Lake) consistently has one of the highest values of FK17 (around 1.30), while zone forty-two (Metro Toronto Planning District 4, in the City of Toronto) has one of the lowest values (around 0.63). At the same time, zone forty-two has one of the lowest values of FUI ($300) each year. The highest values of FUI (around $1,400) are found in Northern urban areas such as Sault St. Marie and North Bay. These subregional trends are consistent for each year, and across the three years.

Income Dynamics

The SAADD data provide an effective means of examining the income dynamics of husbands versus working wives. Table 4 shows the ratio of HY/WY for each year.

Note that Table 4a shows the *total income* in the sample. That is, the total income of husbands is divided by the total income of wives, while Table 4b shows the ratio of the aggregate with the highest (lowest) mean husband income, to the highest (lowest) mean wife income. Table 4a, therefore, shows aggregate results, while Table 4b gives some idea of the best and worst cases.

Over three years, the total income reported by husbands declines from 2.91 to 2.73 times the total income reported by wives. Since HY also increases over the three year period the decreasing income ratio indicates relative gains in total income by working wives. The aggregate results, however, mask the

TABLE 4a Ratio of Husband's Total Income to Wive's Total Income (HY/WY)

	1984	1985	1986
HY/WY	2.91	2.88	2.73

TABLE 4b Ratio of Highest and Lowest Mean Regional Incomes (HY/WY)

Year	High Incomes	Low Incomes
1984	2.97	3.45
1985	2.76	3.21
1986	3.26	2.83

true magnitude of the differences. As expected, all of the regions recording high incomes are in the Metro Toronto region, while the regions recording low incomes are on the Northern boundary of the study region. The income ratios on the margins of the distribution are high, often close to or greater than three. Relative changes that underlay these ratios are equally dramatic (for example, the high husband salary increased almost $20,000, or 34% from 1985 to 1986, while the comparable high wife's salary increased less than $3,000, or 13%).

Relationships Between Variables (Product Moment Correlations)

Table 5 contains the significant Pearson Product Moment Correlation statistics for most of the variables in the sample for 1984 (calculations for 1985 and 1986 data show similar results and are not reported here). All of the values shown in the table are statistically significant (at $\alpha = 0.05$).

The results in Table 5 can be used to further check the internal integrity of the data. For example FY should and is highly and positively correlated with other income variables such as HY and WY. A similar relationship exists between FRT, HRT, and WRT, and FUI, HUI, but not WUI. Similarly, since family allowance is an explicit function (linear for small numbers of children) of the number of children, it should be, and is, highly and positively correlated with FSZ and FK17. HY and WY are positively correlated, capturing the two income dynamics in Southern Ontario. In the same fashion, there are high negative correlations among FSZ, FRT, and HRT; FK17, FRT, and WRT; FUI and FY.

Other bivariate relationships are not necessarily the result of government program rules (as in family allowance) or variable definition. For example, FUI and WRT are positively related suggesting that increasing retirement among wives is related to increasing unemployment among husbands! Since wives are retired, the unemployment benefit flows to the husband. In addition,

TABLE 5 Product Moment Correlation Statistic for Regional Means of Selected Variables in 1984

	REG	FSZ	FK17	FWRK	FY	FUI	FRT	FALL	HY	HUI	HRT	WY	WUI	WRT
REG	n.a.	n.a.	-0.3	0.5	0.6	-0.5	n.a.	-0.3	0.5	-0.5	n.a.	0.6	n.a.	n.a.
FSZ	n.a.	n.a.	0.9	0.4	-0.4	n.a.	-0.7	0.8	-0.3	n.a.	-0.7	-0.4	0.4	-0.6
FK17	-0.3	0.9	n.a.	n.a.	-0.4	n.a.	-0.7	0.9	-0.4	n.a.	-0.7	-0.4	0.5	-0.5
FWRK	0.5	0.4	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	-0.5	-0.5	n.a.	n.a.	n.a.
FY	0.6	-0.4	-0.4	n.a.	n.a.	-0.7	0.5	-0.5	1.0	-0.6	0.5	0.9	-0.3	0.4
FUI	-0.5	n.a.	n.a.	-0.3	-0.7	n.a.	n.a.	n.a.	-0.6	0.9	n.a.	-0.7	-0.4	0.7
FRT	n.a.	-0.7	-0.7	-0.4	0.5	n.a.	n.a.	-0.6	0.6	n.a.	1.0	0.5	-0.4	0.7
FALL	-0.3	0.8	0.9	n.a.	-0.5	n.a.	-0.6	n.a.	-0.4	n.a.	-0.6	-0.5	0.4	-0.5
HY	0.5	-0.3	-0.4	n.a.	1.0	-0.6	0.6	-0.4	n.a.	-0.5	0.6	0.8	-0.3	0.3
HUI	-0.5	n.a.	n.a.	-0.5	-0.6	0.9	n.a.	n.a.	-0.5	n.a.	n.a.	-0.7	n.a.	-0.4
HRT	n.a.	-0.7	-0.7	-0.5	0.5	n.a.	1.0	-0.6	0.6	n.a.	n.a.	0.5	-0.5	0.6
WY	0.6	-0.4	-0.4	n.a.	0.9	-0.7	0.5	-0.5	0.8	-0.7	0.5	n.a.	n.a.	0.5
WUI	n.a.	0.4	0.5	n.a.	-0.3	0.4	-0.4	0.4	-0.3	n.a.	-0.5	n.a.	n.a.	n.a.
WRT	n.a.	-0.6	-0.5	n.a.	0.4	-0.4	0.7	-0.5	0.3	-0.4	0.6	0.5	n.a.	n.a.

Note: REG is the region number, starting with one in the northwest corner of the study region and ending with Metropolitan Toronto (Regions 34 to 46).
Variable FKI8 does not correlate with any other variable at a significant level.
All correlation coefficients shown are significant at 0.05 level at least.
n.a. means that the value is not reported.

there is a negative correlation between FUI and WUI, FUI and WY. Both results suggest that increasing FUI is derived from increasing HUI (as seen before) with wives either working or retired.

The use of the REG variable requires special mention. The extent to which the regional identifier acts as a continuous variable is limited. The numbering is consistent in that low numbered regions are in the North, high numbered regions in the South, with the highest numbers in Metro-Toronto. A better approach might be to adopt longitude and latitude values. The nature of the customized regions, however, makes this approach difficult. For example, some of the urban regions are completely surrounded by rural regions. In addition, to keep the number of regions small, some of the urban regions contain nearby, but disconnected nucleations (for example, zone 20 (St. Catharines-Niagara)). The results in Table 5 do confirm what is readily known about the regional economy. The positive correlation between REG and the income variables is expected (FY,HY, and WY). The negative correlation between REG and FK17, and REG, HUI and FUI were both observed earlier. Thus while the REG variable is not ideal, it does appear to be useful. Caution suggests, however, that it be used carefully.

An alternative approach is to use dummy variables to represent the regions. Dummy variables, however, do not explicitly include the information,

even in the relative sense of the REG variable, of the location of one region with respect to the other regions. Given the regional economic structure of Southern Ontario, this locational information is important.

Finally, sets of dummy variables for each region together with "north or south", "east or west", and "urban or rural" might be constructed for a heavy computational cost.

Conclusions

The Tax Filer data set is not perfect. However, it does provide a data set for analysis into a range of social, economic, and demographic aspects of local populations. Compared to other micro data sets, the Tax Filer data set provides high resolution data over time with a unique level of financial detail linked to family structure. There are a number of weaknesses with the current data set. From a computational standpoint, there are coding errors (birth and death dates for wives), Kingston was eliminated from the sample, and some of the very small urban areas were incorrectly assigned to aggregates. Finally, as mentioned before, meeting the necessary confidentiality requirements has introduced some unknown error component through the random rounding of results. From a demographic analysis standpoint, the existence of only husband/wife nuclear families limits the generalisability of the results. If the goal is to study migration, then a migrant file that links to other members of the family would be necessary in order to properly attribute the migration decision. In spite of these limitations, there is still a great deal of information to be gained from the current sample. After a suitable set of assumptions are tested, a rich set of models can be imagined.

If the wealth of information in the SAADD Tax Filer data set can be made available using something like the customized zones developed for this pilot study, then links could be made with other data sets. Assuming the customized zones are based on some manageable geographical partitioning like postal zones, then current software packages can solve the overlay problem to build comparable partitionings. With such an information system, the real utility of access to a detailed, high resolution spatial data base on an annual interval will become apparent.

Finally, the availability to SAADD of the individual returns provides an ideal situation for the multi-scale, variance-based aggregation filters being developed by Wrigley and his colleagues (Wrigley 1994). If such matrices were built for the Tax Filer data, then analysts would be able to access the statistical information contained in the data, for a range of scales, without violating confidentiality.

References

Amrhein, C. 1992. "Searching for Aggregation Effects Using Age/Sex Specific Migration Data", *The Operational Geographer*, 10: 32-38.

_____. Forthcoming. 1994. "Searching for the Elusive Aggregation Effect: Evidence from Statistical Simulations", *Environment and Planning A*.

Amrhein, C. and R. Flowerdew. 1992. "The Effect of Data Aggregation on a Poisson Regression Model of Canadian Migration", *Environment and Planning A*, 24: 1381-1391.

_____. 1994. "Searching for the Elusive Aggregation Effect: Evidence from British Census Data." Working Paper. Department of Geography, Lancaster University.

Bailly, A. and W. Coffey. 1994. "Regional Science in Crisis: A Plea for a More Open and Relevant Approach", *Papers in Regional Science*, 73: 3-14.

Courchene, T. 1974. *Migration, Income, and Employment: Canada 1965-68.* Montreal: C.D. Howe Institute.

Deichmann, U., M. Goodchild, and L. Anselin. 1992. "Dealing with Errors in Socio-Economic Databases: Selected Findings of a National Research Initiative", *The Operational Geographer*, 10: 12-16.

Dudley, G. 1991. "Scale, Aggregation, and the Modifiable Areal Unit Problem", *The Operational Geographer*, 9: 28-32.

Fotheringham, A. and D. Wong. 1991. "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis", *Environment and Planning A*, 23: 1025-1044.

Fox, W., H. Herzog Jr. and A. Schlottmann. 1989. "Metropolitan Fiscal Structure and Migration", *Journal of Regional Science*, 29: 523-36.

Grant, E. and J. Vanderkamp. 1976. *The Economic Causes and Effects of Migration: Canada, 1965-71.* Ottawa: Economic Council of Canada.

Herzog, H., A. Schlottmann and T. Boehm. 1992. "Migration in Spatial Job Search: A Survey of Empirical Findings". Paper presented to the 39th North American Meetings of the Regional Science Association International, Chicago.

Miron, J. 1992. "Immiseration of the Baby Boom: Earnings, Cohort Size and Female Workforce Participation". Paper Presented to the 39th North American Meetings of the Regional Science Association International, Chicago.

Openshaw, S. and P. Taylor. 1979. "A Million or so Correlation Coefficients: Three Experiments on the Modifiable Areal Unit Problem", in N. Wrigley (ed.), *Statistical Applications in the Spatial Sciences*. London: Pion.

Shaw, R.P. 1985. *Intermetropolitan Migration in Canada.* Toronto: New Canada Publications.

Shumway, J.M. 1993. "Factors Influencing Unemployment Duration with a

Special Emphasis on Migration: An Investigation Using SIPP Data and Event History Methods", *Papers in Regional Science*, 72: 159-176.

Tobler, W. 1991. "Frame Independent Spatial Analysis", in M. Goodchild and S. Gopal (eds.), *Accuracy of Spatial Data Bases*. New York: Taylor and Francis.

Waldorf, B. 1992. "The Determinants of Assimilation and Attachment in the Context of International Migration: The Case of Guestworkers in Germany". Paper presented to the 39th North American Meetings of the Regional Science Association International, Chicago.

Wrigley, N. 1994. "Revisiting the Modifiable Areal Unit Problem and the Ecological Fallacy". Department of Geography, University of Southhampton, forthcoming for the Festschrift for Peter Haggett, edited by Cliff, A.D., Gould, P.R., Hoare, A.G. and Thrift, N.J. to be published by Blackwell: Oxford.