THE USE OF ADMINISTRATIVE RECORDS TO ESTIMATE WAGES AND SALARIES FOR SMALL BUSINESSES IN SMALL AREAS

E.B. Dagum, M.A. Hidiroglou and M. Morry, Statistics Canada

1.0 INTRODUCTION

Through the past decade, there has been considerable attention given to the statistical uses that can be made of administrative records. Increasing emphasis is being placed on the use of these records to produce economic statistics for small areas for which no reliable survey estimates exist. This paper discusses problems encountered with the reconciliation of data from two administrative files considered as possible sources for producing small area statistics, related to Business Income and to Wages and Salaries.

Statistics Canada was given access to Income Tax data for the purpose of statistical analysis through the Statistics Act of 1971. The objective was to provide data in lieu of survey data, to improve cost efficiency and to reduce response burden. In the first phase of the present project we investigated the possibility of providing small area estimates on unincorporated businesses. Tax data for unincorporated businesses (T1) have been transcribed by Statistics Canada since 1973. The T1 tax return is the individual tax return that is filed annually by all individuals taxable in Canada. The tax filers which report a business income are of particular interest for the purposes of the present project. From these, statistics relating to unincorporated businesses in Canada can be obtained. Such statistics can be used to estimate the structure (i.e., the breakdown by Standard Industrial Code, province and size) of the unincorporated business universe for a given year. The transcription of these tax records at Statistics Canada is stored on a file known as the COMBINED-MASTER.

Revenue Canada transcribes tax data for unincorporated tax filers to be used in their auditing procedures (for the purposes of their potential for audit through the examination of the returns). The resulting file is known as the COMSCREEN Master file. The COMSCREEN may be regarded as a universe file for unincorporated tax filers that have declared Gross Business Income over $25,000. The COMBINED-MASTER is a 10% sample for unincorporated filers with Gross Business Income between $10,000 to $25,000, roughly a 25% sample for filers with Gross Business Income between $25,000 to $500,000 and a 100% sample for filers with Gross Business Income over $500,000. The COMSCREEN and COMBINED-MASTER contain a number of economic variables which are comparable in concept: these are Sales, Capital Cost Allowance, Net Profit, Gross Profit, Filer's Share of the Net Profit for filers that are involved in a partnership. The COMBINED-MASTER has a number of additional economic variables which are not transcribed on the COMSCREEN file. These are Wages and Salaries, Inventories and Assets. Thus one file (COMSCREEN) is more complete in terms of coverage for businesses with income between $25,000 and $500,000 but contains less information, while the second file (COMBINED-MASTER) has all of the variables of interest but only on a sample basis. Estimates of the variables missing from the COMSCREEN file can be obtained in one of two ways. One way is to use domain estimation by weighting up the records on the COMBINED-MASTER. The other way is to obtain relationships between these variables and variables common to both files using the COMBINED-MASTER and applying them to the same variables on the COMSCREEN file.

In order to provide breakdowns on Gross Business Income and Wages and Salaries, the two files had to contain industrial and geographical classification codes compatible with Statistics Canada standards. Since the classification codes for the COMBINED-MASTER file were assigned at Statistics Canada, these standards were most likely met. It was therefore important to compare the classification codes generated on the COMSCREEN at Revenue Canada to those on the COMBINED-MASTER to determine if the COMSCREEN data could be tabulated. Although the economic variables transcribed by the two agencies are comparable in concept, a numerical comparison had to be carried out to measure the level of agreement. Results of the comparability of classification codes and economic data on the two files are presented in Section 2.

The above comparison indicated whether Gross Business Income could be tabulated using the more complete Revenue Canada file for tax filers with Gross Business Income between $25,000 to $500,000. To obtain estimates of Wages and Salaries missing on the COMSCREEN, regression techniques were investigated using explanatory variables on the COMBINED-MASTER common to both files. Section 3 describes the steps involved in this analysis. Based on the results of the regression analysis, several estimators provided in the literature are considered in Section 4 to estimate Wages and Salaries for small areas. Section 5 gives a summary of the conclusions.

2.0 COMPARISON OF ELEMENTS BETWEEN THE COMBINED-MASTER AND COMSCREEN FILES FOR TAX YEAR 1981

One of the objectives of the small area project is to produce estimates of Gross Business Income, Capital Cost Allowance, Net Profits, Wages and Salaries and Total Assets at the subprovincial level. The first three items can be found on both the COMBINED-MASTER and COMSCREEN files while the last two items can only be found on the COMBINED-MASTER file. The COMBINED-MASTER is a file resulting from the coding of tax returns for incorporated and unincorporated (T1) filers by Tax Record Access at Statistics Canada based on a pre-specified sample and a sampling algorithm. The COMSCREEN is an audit file created by Revenue Canada based on a sample selected from the universe of self-employed tax filers. All businesses with Gross Income over $25,000 are on the COMSCREEN file.

The COMSCREEN is made up of basically two parts: the first part is the T1 portion as keyed by Revenue Canada and the second part is made up of three segments transcribed at Revenue Canada using tax data pertaining to the three major businesses of a tax filer. For the COMBINED-MASTER, transcription of selected data items is done for each business belonging to the tax filer. The transcript of tax filers with only one business will be referred to as a single record.

Since the COMBINED-MASTER is roughly a one in four sample for businesses whose Gross Business Income lies between $25,000 to $500,000 and a 100% sample for all businesses whose Gross Business Income is above $500,000, the comparison of data elements between the COMBINED-MASTER and COMSCREEN files will be done for businesses whose Gross Business Income (GBI) falls in the range of $25,000 and over.

Single T1 records from the COMBINED-MASTER (200,016 records) were matched to single records from the COMSCREEN File (483,534 records) using Social Insurance Number (SIN) as matching key. The match resulted in 136,982 records which could be used to compare similar fields between the two files.

The fields of interest on the COMBINED-MASTER and COMSCREEN that are to be compared are:

1. **Standard Industrial Classification (SIC)**
2. **Standard Geographical Classification (SGC)**
3. **Selected economic variables**

2.1 Standard Industrial Classification Comparisons

Some of the coding differences for the 1980 Standard Industrial Classification (SIC) between the COMSCREEN and COMBINED-MASTER files were investigated. The coding differences implied that the comparisons beyond the two digit level of SIC would not be meaningful. Thus, comparisons were carried out at the Major Division (1 digit SIC) and Major Group (2 digit SIC) level only.

Table 1 provides a summary of the agreement between Statistics Canada (STC) and Revenue Canada (RCT) SIC coding at the Major Division level. From Table 1, one observes that there is an overall agreement of 78%. The industries that show good agreement are Construction, Transportation, Retail, Accommodation Service, and Logging and Forestry. The poorest areas of agreement are in Fishing, Mining, Manufacturing, Wholesale and Real Estate.

TABLE 1. Summary of Agreement between STC and RCT Coding at the Major Division Level (Controlling for STC Coding)*

Major Division	Agreement
Total	78%
A. Agriculture	73%
B. Fishing	45%
C. Logging and Forestry	84%
D. Mining	59%
E. Manufacturing	52%
F. Construction	90%
G. Transportation	92%
H. Communication	69%
I. Wholesale	41%
J. Retail	91%
K. Finance and Insurance	75%
L. Real Estate	44%
M. Business Service	78%
O.. Educational Service	83%
P. Health and Social	89%
Q. Accommodation Service	93%
R. Other Services	86%

STC=Statistics Canada RCT=Revenue Canada

Some of the major coding differences can be summarized as follows: i) for businesses coded to Manufacturing by Statistics Canada, 29% of the records are coded to Retail by Revenue Canada. ii) for businesses coded to Communication by Statistics Canada, 19% of the records are coded to Transportation by Revenue Canada. iii) for businesses coded to Wholesale by Statistics Canada, 32% of the records are coded to Retail by Revenue Canada. iv) for businesses coded to Real Estate by Statistics Canada, 31% of the records are coded to Construction by Revenue Canada.

The level of agreement between the Revenue Canada and Statistics Canada coding is reduced to 68% when the Major Group level is compared. The differences in coding of the Major Groups occur within and between the Major Divisions.

2.2 Comparison of Standard Geographical Classification Code

Area codes are identifiers which are associated with the units of an area system. For each set of spatial units which comprise an area system, there may be one or more set of area codes. The Standard Geographical Classification code (SGC) is one of the many possible sets of area codes which may be applied to provinces, census divisions, and census subdivisions. The SGC can be broken up into several components, the first two of which are the province code and the census division codes within the province. Census division is a general term applying to counties, regional districts, regional municipalities, and five other types of geographical areas made up of groups of subdivisions.

In the present context, census divisions will be the small area of interest. In order to obtain census codes from the COMSCREEN file, there are three address sources which can be converted to these codes. These are: the filer's address, the filer's postal code and the locality code. The locality code is a five digit code assigned by Revenue Canada. The first three digits identify province, county or census division and selected larger municipalities. The last two digits identify the municipality within the county or census division based on the SGC. These address sources are also found on the COMBINED-MASTER file. There are an additional two address sources transcribed on the COMBINED-MASTER file which are: the business address and the business postal code. Since the small area estimates must refer to business activity and since the census divisions on the COMSCREEN file can be assigned only using the tax filer's address or locality code, it is of interest to assess how well a census division code can be derived from a filer's address or locality code as opposed to a business address. If business addresses are in general close to filer's addresses one can use the filer's address as a good proxy to describe the location of the business activity.

In order to determine the level of agreement between these different sources, some 200,000 records on the COMBINED-MASTER file were processed through conversion tapes which would assign an SGC to each of the five existing address sources. The correspondence at the provincial and census division level is provided in Table 2.

As can be noted from Table 2, the level of

TABLE 2. Level of Agreement for Assigning SGC Codes for Different Address Sources

Source of Comparison	Number of Records	Province Agreement	Census Division Agreement
1. Business Postal Code vs Business Address	10,627	0.998	0.984
2. Business Postal Code vs Filer's Postal Code	13,977	0.994	0.953
3. Business Postal Code vs Filer's Address	12,443	0.994	0.949
4. Business Postal Code vs Locality Code	11,452	0.994	0.948
5. Business Address vs Filer's Postal Code	45,682	0.995	0.936
6. Business Address vs Filer's Address	47,833	0.996	0.949
7. Business Address vs Locality Code	42,673	0.994	0.942
8. Filer's Postal Code vs Filer's Address	150,130	1.000	0.991
9. Filer's Postal Code vs Locality Code	140,842	0.998	0.980
10. Filer's Address vs Locality Code	139,344	0.998	0.998

agreement between the different address sources is very good for provinces. For census divisions the level of agreement, although lower than the one obtained for provinces is quite good. The filer's address seems to be as good a proxy as the locality code for obtaining the census division given that the business address is regarded as the best source. The correspondence between the addresses and their associated postal codes for assigning a census division code is quite good. For COMSCREEN there is a filer's postal code for 90% of the records, a filer's address for 99.9% of the records and a locality code for 99.9% of the records. Consequently, almost all records on the COMSCREEN file could be assigned a census division code based on the filer's address. Furthermore, according to Table 2, this code would be a good proxy for the one that would have been obtained if a business address had been available.

2.3 Comparison of Economic Variables on the Two Files

There are a number of variables on the COMSCREEN file which may be used to obtain estimates on income, counts of businesses within specified small areas or as auxiliary information to predict some variables of interest. Revenue Canada and Statistics Canada transcribe similar types of data from the T1 tax returns and their associated financial statements onto the COMSCREEN and COMBINED-MASTER files respectively. On the COMSCREEN file, these items are known as Sales, Gross Profit Capital Cost Allowance, Net Profit and Tax Payer's Share of Partnership of Net Profit. The corresponding items on the COMBINED-MASTER file are known respectively as Gross Business Income, Gross Profit, Depreciation Total, Net Profit or Loss and Filer's Share of Net Profit or Loss. The definitions of the corresponding variables vary slightly between the two government agencies. In order to assess the extent to which the two sets of figures differ, five ratios were formed and their distribution tabulated. A brief summary of the comparisons is provided in Table 3.

The distribution of ratios between 0.9 and 1.1 provided for the pairs of variables below show that the agreement is quite good. Two measures of agreement are provided for this range. One (including zeroes) where zeroes for one of the variables in the pair is allowed to enter into the computation of the ratio. The other (excluding zeroes) where only non-zero entries for both variables are allowed to enter into the computation of the ratio. The strict exclusion of zeroes from both pairs of variables is probably a better measure of agreement than the alternative measure of agreement (including zeroes). The results are rather encouraging, especially concerning the variable of major interest, Gross Business Income for which 98% of the non-zero values are within 10% (the ratio fell between .9 and 1.1) of the corresponding COMSCREEN value indicating that the more complete COMSCREEN file contains basically the same variable.

After having carried out the comparison of industrial and geographical classification codes as well as the comparison of related economic variables on the COMBINED-MASTER and COMSCREEN file, the following conclusions can be drawn: 1. There is a relatively good agreement between the two files on industrial classification codes at the major division level (78%). This agreement deteriorates to 68% when the major group level is compared, consequently tabulating income and other variables from the COMSCREEN should not go beyond the 1 digit SIC breakdown. 2. On the basis of records that contain both filer's address and business address, it is evident that the business activity takes place in the general location of the residence of the filer as shown by a 95% agreement between the census divisions corresponding to the two addresses. Thus tabulating Income and Wages and Salaries according to geographic locations derived from filer's addresses will be

TABLE 3. Summary of Distribution of Ratios for Selected Variables from the COMBINED-MASTER (STC) and COMSCREEN (RCT)

Comparisons	Range 0.9 to 1.1	
	Including Zeroes	Excluding Zeroes
1. Gross Business Income (SIC) vs Sales (RCT)	92%	95%
2. Gross Profit (RCT) vs Gross Profit (STC)	87%	89%
3. Depreciation (STC) vs Capital Cost Allowance (RCT)	85%	91%
4. Net Profit (SIC) vs Net Profit (RCT)	82%	87%
5. Partnership' share (STC) vs Partnership' share (RCT)	92%	96%

basically equivalent to tabulating these economic variables according to the location of the business activity. 3. Concerning the COMSCREEN and COMBINED-MASTER economic variables pertaining to the same concept the discrepancies are within the 10% range for over 90% of the entries, suggesting that totals obtained from the COMSCREEN would come close in value to corresponding totals from the COMBINED-MASTER.

After having verified that the industrial and geographic classification codes were of acceptable quality it was possible to proceed with the estimation of the variables for small area in particular those of Wages and Salaries that were only available on a sample basis.

3.0 MODELLING WAGES AND SALARIES

To gain some experience in modelling Wages and Salaries, the COMBINED-MASTER file was used. The four digit 1980 Standard Industrial Codes (SIC) on the COMBINED-MASTER were regrouped into 18 major divisions and into 76 major groups separately. Major divisions represent the highest level of aggregation of industries, each division representing one of these broad types of activity (e.g., Agricultural Industries, Forestry Industries,..., Mining Industries, Manufacturing Industries, etc.). Major groups are an aggregation of industries which are subsets of the major divisions (e.g., Food Industries, Beverage Industries in the Manufacturing Division). This regrouping of SIC was necessary to make the analysis more manageable and to investigate whether estimates of Wages and Salaries for unincorporated businesses (T1) could be produced at these levels of industrial aggregation for selected small areas.

The analysis was restricted to tax filers which had declared a $25,000 to $500,000 Gross Business Income (GBI) range for two reasons. Firstly, COMSCREEN does not have any transcribed information for tax filers with GBI less than $25,000 and secondly, the COMBINED-MASTER has Wages and Salaries information transcribed for all tax filers with GBI over $500,000.

Previous earning models described by Lillard and Willis (1978), Greenless, Reece and Zieschang (1982), Betson and Van der Gag (1983), and Little and Samuhel (1983) related the logarithm of Wages and Salaries to demographic data which included data items such as education and work experience, race, urbanity region, one digit occupational codes, weeks worked, and hours per week worked. In our context, the auxiliary information which can be used as part of the regression model is in the form of economic (or accounting) variables associated with tax returns from businesses. They include variables such as Gross Business Income, Net Profit, Gross Profit and Depreciation.

3.1 Some Characteristics of Wages and Salaries

The $25,000 to $500,000 Gross Business Income range was split up into ten classes and cross-tabulations were obtained for the frequency of records with non-zero Wages and Salaries within each major division and across all major divisions. It was observed that for most major divisions, the proportion of filers showing Wages and Salaries increases as the business income increases.

In terms of estimation of Wages and Salaries, this implies that the records on the COMSCREEN have to be imputed with zero Wages and Salaries using the distribution obtained from the COMBINED-MASTER by income class group. The chosen distribution would be assumed to hold for filers that have no Wages and Salaries in the small areas.

Plots of the averages of non-zero Wages and Salaries within Business Income classes within major divisions at the Canada level versus their corresponding Business Income averages were obtained in order to observe the relationship between those two variables at an aggregated level. The resulting plots indicated that there is a good relationship across all major divisions for those two variables. Plots of the raw values for non-zero Wages and Salaries versus their associated Gross Business Income indicated that there could be a good deal of variation of Wages and Salaries for similar business income, the scatter of the Wages and Salaries increased as Gross Business Income increased. The scatter plots point to a transformation of the data to make it more homogeneous.

3.2 Transformation of Data

Previous analyses indicated that Gross Business Income was the most strongly correlated variable with non-zero Wages and Salaries across the majority of major divisions and major groups.

Scatter plots of non-zero Wages and Salaries versus Gross Business Income indicated that the spread of non-zero Wages and Salaries increased as Gross Business Income increased. This suggested that a transformation of the data might be required. Since Gross Business Income was the most correlated variable with non-zero Wages and Salaries, five models were tried out using this auxiliary variable. The models were fitted at the Canada level by Major Division. The models were:

i) $SALWAG = INT1 + SLOPE1 * GBI + E1$;

ii) $SALWAG = SLOPE2 * GBI + E2$

iii) $SALWAG/\sqrt{GBI} = INT3/\sqrt{GBI} + SLOPE3 * \sqrt{GBI} + E3$

iv) $SALWAG/\sqrt{GBI} = SLOPE4 * \sqrt{GBI} + E4$

v) $LOG(SALWAG) = INT5 + SLOPE5 * LOG(GBI) + E5$

where SALWAG= Wages and Salaries, GBI=Gross Business Income, INT = intercept, SLOPE = slope of the regression, E = error term.

Examination of the standardized residuals and the adjusted $\bar{R}^2_p$ term, indicated that the square root transformation was the best. This transformation pointed to the appropriateness of a ratio-type estimator. Furthermore, the intercept term was not sufficiently significant to include in the model.

3.3 Search for Fits

Scatter plots of the ratio of the mean of non-zero Wages and Salaries to the mean of the Gross Business Income within selected intervals of the Gross Business Income were obtained to determine whether these ratios were constant over the Gross Business Income intervals or whether they increased or decreased over the intervals or whether there existed breaks between which these ratios were constant. These scatter plots which were done at the major division by province cross-classification indicated that a mixture of these conditions could exist depending upon the major division and provincial cross-classification.

As a result of the above scatter plots, eight regression models were fitted to reflect if the conclusions drawn from the scatter plots held. These models were: i) $SALWAG/\sqrt{GBI}$ vs. $\sqrt{GBI}$:Linear ii) $SALWAG/\sqrt{GBI}$ vs. $\sqrt{GBI}$, $(GBI)^{3/2}$: Linear & Quad-

ratic iii) $SALWAG/\sqrt{GBI}$ vs. $(GBI)^{3/2}$: Quadratic iv) $SALWAG/\sqrt{GBI}$ vs. $\sqrt{GBI}$:Linear, break at $100K Gross Business Income v) - viii) Models i) through iv) with provinces added as dummy variables to test if the fits differed by province.

From Table 4, one concludes that for the most part, the fits are linear within each major division and that they differ in slope between provinces in the majority of the cases. For those major divisions which have a combination of linear and quadratic terms, although the addition of the quadratic term is statistically significant, the adjusted coefficient of determination ($\bar{R}_p^2$) is increased only slightly.

TABLE 4. Summary of the Fits by Major Division

Major Division	Best Fit at the Canada Level	Provinces Signi. Different	$\bar{R}_p^2$	Number of Observations
1. Logging & Forestry	Linear*	Yes*	0.74	980
2. Mining	Linear*	No*	0.62	88
3. Manufacturing	Linear*	Yes*	0.69	2,987
4. Construction	Linear*	Yes*	0.64	12,585
5. Transportation	Linear*	Yes*	0.57	3,992
6. Communication	Linear*	No	0.72	248
7. Wholesale	Linear & Quadratic*	Yes*	0.46	1,396
8. Retail	Linear & Quadratic*	Yes*	0.51	18,545
9. Finance & Insurance	Break at 100K	No	0.83	15
10. Real Estate	Linear*	Yes*	0.52	303
11. Business Service	Linear*	Yes*	0.64	534
12. Educational Service	Linear & Quadratic*	Yes**	0.74	91
13. Health & Social	Linear*	Yes**	0.70	378
14. Accommodation	Linear & Quadratic*	Yes*	0.76	6,005
15. Other Services	Linear*	Yes*	0.62	5,127

* : Significant at the 1% level.
**: Significant at the 5% level.

The linear model $SALWAG/\sqrt{GBI}$ versus $\sqrt{GBI}$ was fitted by major group at the Canada level with and without the provinces used as dummy variables. The disaggregation of major division into major groups did not significantly improve the fits.

The auxiliary variable Gross Business Income (GBI) was the most highly correlated variable with Wages and Salaries for most major divisions and major groups and their cross-classification with provinces. A number of other variables common to CONSCREEN and to the COMBINED-MASTER files showed high correlation with Wages and Salaries. However, they were also very highly correlated with Gross Business Income, therefore, once GBI was included in the regression, they were not used as additional explanatory variables in a multiple regression to avoid multicollinearity.

The selected auxiliary variables that were fitted in a one variable regression were: Depreciation, Net Profit, Gross Profit, Partnership (a zero-one variable) Gross Business Income, Gross Professional Income, the Square of Gross Business Income.

It was found that with the exception of the major division Retail for which the regression using Gross Profit gave the highest $\bar{R}_p^2$ the best fit was obtained using Gross Business Income as the explanatory variable.

The search for the best one variable model was also carried out at the province level within each major division.

The conclusions were that the disaggregation of the fits for major divisions from the Canada level to the provincial level improves $\bar{R}_p^2$ for some provinces in some instances and worsens $\bar{R}_p^2$ in others. Some of the major divisions such as Mining, Finance and Insurance, and Educational Service do not contain enough data to be disaggregated from the Canada to the provincial level. For these major divisions, the fits should be done at the Canada level.

The investigation concerning the modelling of Wages and Salaries led to the following conclusions: 1) among the available auxiliary variables Gross Business Income is the one most strongly related to Wages and Salaries; 2) in fitting the regression one line without the intercept term provides the best fit; 3) it is necessary to divide the regression equation through by the square root of GBI to make the residuals homoscedastic - this latter transformation makes estimation via regression equivalent to applying a ratio-type estimator to obtain estimates of Wages and Salaries; 4) the optimal level at which modelling should be carried out is at the major division by province cross-classification (pooling across provinces in industries with not enough data points per province).

4.0 ESTIMATORS OF WAGES AND SALARIES FOR SMALL AREAS

Wages and Salaries for small areas may be obtained in several ways using the COMBINED-MASTER and CONSCREEN files. If there are enough sampled units within the small area, which may be for example a cross-classification of Census Division and Major Division, the direct estimator (sum of weighted up data of the COMBINED-MASTER) may be good enough to produce satisfactory precision.

If the direct estimator is not sufficiently good enough, synthetic estimation (Gonzalez 1973) may be used to produce small areas estimates. For synthetic estimation, an unbiased estimate is obtained from the COMBINED-MASTER sample for a large area; when this estimate is used to derive estimates for sub-areas on the assumption that the small areas have the same characteristics as the larger areas, these estimates are identified as synthetic estimates. When the assumption that small areas resemble large areas fails, the synthetic estimator becomes design biased. Despite the bias, we may gamble on the synthetic estimate, because strength will be 'borrowed' if the assumption holds. Procedures to correct for this bias have been proposed by Särndal (1981), given that the sample sizes falling in the small areas are relatively large.

Production of synthetic estimates requires the use of both the CONSCREEN and the COMBINED-MASTER. Two types of synthetic estimators may be contemplated, which will be referred to as the count-synthetic and the ratio-synthetic. As was pointed out previously, the relationship between Wages and Salaries and Gross Business Income is such that a ratio-type estimator is most appropriate. The

count-synthetic estimator being the simplest of its kind should be considered for purposes of comparison. For a given province and industrial grouping, the data on the COMBINED-MASTER and the COMSCREEN files are split into Gross Business Income (GBI) groups. For the count-synthetic, mean Wages and Salaries are obtained for each of these GBI groupings within a provincial and industrial cross-classification from the COMBINED-MASTER file and multiplied by the population counts within the areas for the corresponding provincial and industrial cross-classification on the COMSCREEN file. For the ratio-synthetic, proportions of Wages and Salaries totals to Gross Business Income totals are obtained for the GBI groupings within a provincial and industrial cross-classification from the COMBINED-MASTER file and multiplied by the GBI population totals within the areas for the corresponding provincial and industrial cross-classification on the COMSCREEN file. The use of synthetic estimation assumes that the industrial coding and that the Gross Business Income between the two files are comparable. Furthermore, it is assumed that the COMSCREEN file is a complete file for tax filers with Business Income over $25,000.

More sophisticated procedures which incorporate mixtures of direct and synthetic estimation using regression have been suggested by Särndal (1981), and Fay and Herriot (1979). In the context of the present study, the regression estimation again reduces to ratio-type estimation. Särndal models the data across all areas using regression procedures and corrects the synthetic estimators for the bias by comparing weighted up estimates of the GBI groups at the area level and across all areas. Fay and Herriot model area means, by fitting a regression to each small area, and form a weighted average of the sample and regression estimate for each small area. They adjust the weights to reflect the relative magnitudes of the average lack of fit of the regression and the variance of the sample estimate. The advantage of Särndal's procedure over strictly synthetic estimation is that estimates of reliability can be attached with the small area estimates. This is an important factor which can be taken into account when one decides how to group the small areas and the industries into classifications that are publishable.

5.0 CONCLUSIONS

This study set out to investigate if the variable Wages and Salaries that is only available on a sample basis can be estimated reliably for small areas through the use of a universe file that contains related auxiliary variables.

Examination of the two available administrative data sources indicated that the two files were compatible in terms of industrial classification at the major division level, geographic coding at the census division level and in the content of the economic variables present. This comparability enabled the application of the relationship derived from the sample file to the auxiliary variable on the universe file to obtain improved estimates of Wages and Salaries for industries in small areas. The relationship with the auxiliary variable Gross Business Income was strongest at the major division industrial breakdown by province. The square root of GBI transformation necessary led to ratio-type estimation.

A subsequent simulation study by the present authors in co-authorship with Rao and Särndal (1984) indicated that all of the small-area estimators proposed in this paper showed improved efficiency over the direct estimation. This suggests that enriching the sample data with information available on the administrative universe file will be a viable alternative for producing reliable small area estimates.

REFERENCES

(1) Betson, D. and Van Der Gaag, 3. (1983). Working married women and their impact on the distribution of welfare in the United States. Working paper, Institute for Research on Poverty, University of Wisconsin.

(2) Dagum, E.B., Hidiroglou, M.A., Morry, M., Rao, J.N.K., and Sarndal, C.E., (1984), Evaluation of Alternative Small Area Estimators Using Administrative Records. Presented at the Annual Meeting of the American Statistical Association, Philadelphia; Aug. 13-16.

(3) Fay III, R.E. and Herriot, R.A. (1979). Estimates of Income for Small Places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, 405-410.

(4) Gonzalez, M.E. (1973). Use and evaluation of synthetic estimates. Proceedings American Statistical Association, Social Statistics Section, 33-36.

(5) Greenless, W.S. Reece, J.S. and Zieschang, K.D. (1982). Imputation of missing values when the probability of response depends on the value being imputed. Journal of the American Statistical Association, 77. 251-261.

(6) Lillard, L.A., and Willis, R.J. (1978). Dynamic Aspects of Earning Mobility. Econometrics, 46, 985-1011.

(7) Little, Roderick, J.A. and Samuhel, Michael, E., (1983). Alternative Models for CPS Income Imputation. Presented at the Annual Meeting of the American Statistical Association, 85-90.

(8) Särndal, C.E. (1981). When robust estimation is not an obvious answer: The Case of the synthetic estimator versus alternatives for small areas. Proceedings American Statistical Association, Survey Research Section, 710-712.

For further information, see also:

(9) Gonzalez, M.E. and Hoza, C. (1978). Small area estimation with application unemployment and housing estimates. Journal of the American Statistical Association, 73, 7-15

(10) Hidiroglou, M.A. (1984). Exploratory Analyses Performed on the COMBINED-MASTER file. Technical report, Statistics Canada.

(11) Hidiroglou, M.A. (1984). Some characteristics of SIC coding between COMSCREEN and the COMBINED-MASTER. Technical report, Statistics Canada.

(12) Hidiroglou, M.A., Morry, M. and Vaillancourt, C. (1984). Comparison of Elements between the COMBINED-MASTER and COMSCREEN files. Technical report, Statistics Canada.

(13) Holt, D., Smith, I.M.F. and Tomberlin, T.J. (1979). A model based approach to estimation for small subgroups of a population. Journal of the American Statistical Association, 74, 405-410.