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.

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 Särndal, 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.