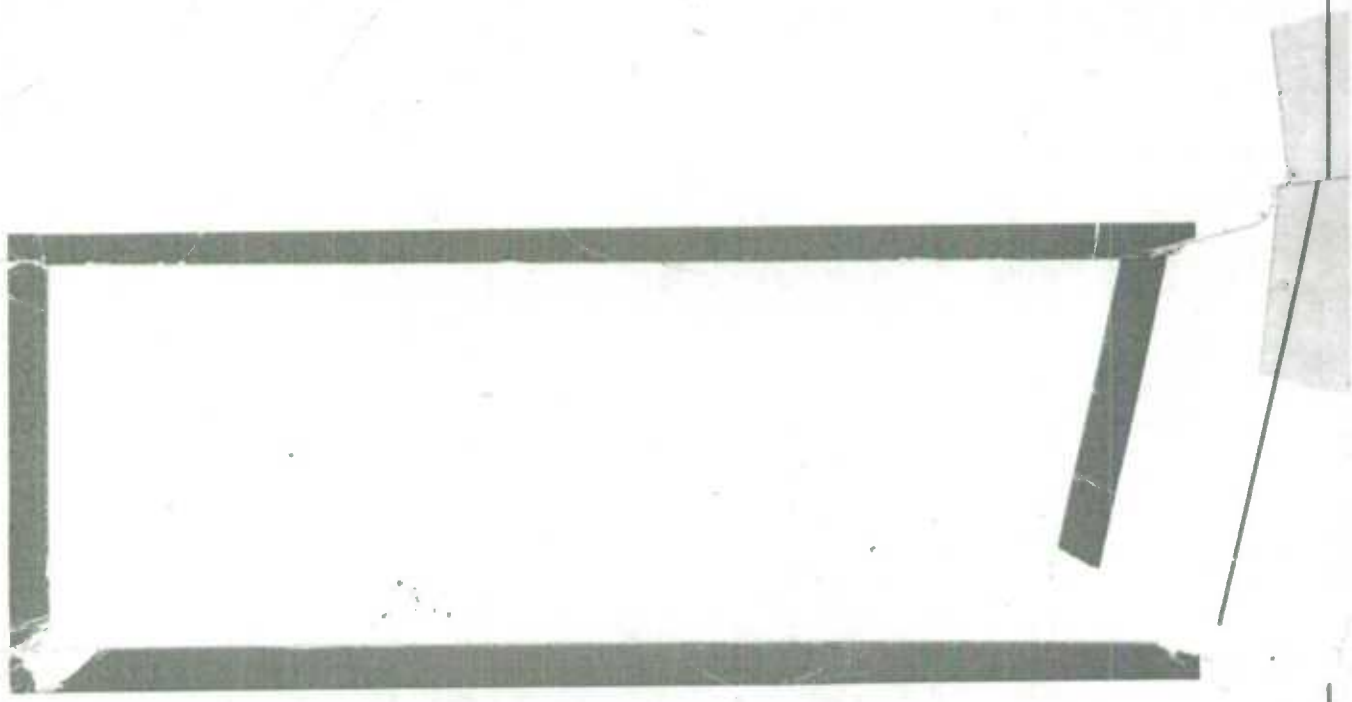
Statistics Canada Statistique Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes entreprises

11-617
no.85-44
c 2

Canada

#5042y

PROBLEMS ASSOCIATED WITH THE ESTIMATION OF SMALL AREA BUSINESS DATA

M.A. Hidiroglou

Working Paper No. BSMD 85-044E

STATISTICS CANADA STATISTIQUE CANADA
SEP 20 1993
LIBRARY
BIBLIOTHÈQUE

PROBLEMS ASSOCIATED WITH THE ESTIMATION OF SMALL AREA BUSINESS DATA

by

M.A. Hidiroglou

Statistics Canada

Invited Paper presented at the Statistical Society of Canada Meetings in Winnipeg, June 2-4, 1985

Problèmes Reliés à l'Estimation des Données sur les Entreprises pour les Petites Régions.

M.A. Hidiroglou

Sommaire

L'estimation des données sur les entreprises pour les petites régions présentes quelques problèmes techniques. Dans le cadre de l'étude présente, nous nous sommes adressés à plusieurs questions. Celles-ci ont inclu, la compatibilité des filières administratives; le dévelopement de modèles de regression afin d'appliquer des méthodes synthétiques ou "pseudo-synthetiques" pour l'estimation des petites régions; et une investigation sur les caractéristique de plusieurs estimations pour les petites régions.

Table of Contents

PROBLEMS ASSOCIATED WITH THE ESTIMATION OF SMALL AREA BUSINESS DATA

by

M. A. Hidiroglou

Statistics Canada

1.0 INTRODUCTION

In the recent years, Statistics Canada has been given the mandate to produce small area estimates for a number of statistics of interest. The production of such statistics is being put into place by using administrative files and/or survey files which already exist. The production of small area statistics from such files can vary in complexity. It may be as simple a matter as geocoding the files, provided that their coverage is satisfactory and that their variables correspond conceptually to the required statistics, and tabulating them at the small area level. It may be as complex as using several files together, working out relationships between required variables from some of the smaller files which contain a large number of variables and applying these relationships on the complete files (in terms of coverage) which contain a smaller number of variables in order to produce the required statistics. Some of the smaller files, which contain a sample of the population of interest, may in some instances, support small area estimation. Such instances occur when the number of sampled observations within a small area is high enough to produce statistics with an acceptable coefficient of variation. Such cases will occur on an infrequent basis because the sampling is usually performed at a higher level than the small areas. In the case of the production of business statistics for small areas, which will be the concern of this paper, the sampling occurs at the Canada or provincial level cross-stratified with a non-geographical variable. This variable may be business income, number of employees, or Standard Industrial Classification.

In the course of developing small area statistics using administrative and/or survey files, there are a number of issues which arise. One of the basic questions is the definition of a small area. Its definition depends on several factors which are as follows. Firstly, there must exist geocoding systems which can successfully convert an address into a required geographical sub-partition of a large area (province). Secondly, these sub-partitions must correspond to user requirements for the production of statistics of interest. Thirdly, the published data must respect the norms of confidentiality set by the organization. Fourthly, the ensuing published numbers must be reliable. Finally the statistics, usually available at a much higher level, must be disaggregated into the small areas of interest: this concern is part of a larger study described by Sande (1984) which will not be addressed in this paper. The choice of the files and of the associated variables is also important. Several administrative files were proposed by Gigantes (1983) as a means to produce economic statistics of interest for small areas. They included administrative tax files produced by Revenue Canada (Comscreen, Self Employed Income File, CORPAC) and administrative tax files produced by Statistics Canada (Combined.Master, ADMIN). These files contain a number of variables of interest which include financial information for unincorporated tax filers (T1) and incorporated tax filers (T2) in Canada.

For the purposes of the present study, the possibility of providing small area estimates on unincorporated businesses was investigated using two administrative files. One file known as the Combined.Master is compiled at Statistics Canada. Access to the returns of the unincorporated businesses which constitute this file was given to Statistics Canada through the Statistics Act of 1971 for the purpose of statistical analysis. Such statistics can be used to estimate the structure (i.e. breakdown by Standard Industrial Code, province and business income size) of the unincorporated business for a given year. The transcription of these tax records at Statistics Canada is stored on the Combined.Master. An excellent description of the composition, construction and uses of the Combined.Master is given by Darcovich (1982). The Combined.Master is a 10% sample for unincorporated filers with Gross Business Income between

$10,000 to $25,000, a 25% sample for unincorporated filers with Gross Business Income between $25,000 to $500,000 and a 100% sample for filers with Gross Business Income over $500,000. There is a constant 2% longitudinal sample of tax filers across the years which is used to study the properties of a tax filer over time. In addition, some records may be pre-specified in advance by using the identification of tax records transcribed in previous years. Some of the characteristics of the transcribed variables on this file have been studied by Sande (1978). The resulting weighting procedures associated with the sample selection have been described in the multiple frame methodological context by Bankier (1983). The other file, obtained through Revenue Canada, is known as the Comscreen file. This file, which is used by Revenue Canada for auditing unincorporated tax filers, may be regarded as a universe file for unincorporated tax filers that have declared Gross Business Income over $25,000.

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 other file (Combined.Master) has all the variables of interest on a sample basis.

Estimates of the variables missing from the Comscreen file can be obtained in one or 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. Using one of these two ways, small area estimates of Wages and Salaries, Total Assets or Inventories can then be produced at the subprovincial level for specific industrial classifications. It must be

noted that the other variables common to both files can be tabulated off the Comscreen file taking the partnership factor into account at the small area level provided that they are compatible.

In order to provide industrial and geographical breakdowns on Gross Business Income or Wages and Salaries for example, the two files had to contain these classification codes compatible with Statistics Canada standards. It was therefore important to compare the classification codes generated on the Comscreen at Revenue Canada to those on the Combined.Master to determine their differences. Although the economic variables transcribed in common 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. Implications of differences between the two files on the estimation strategy are also provided in this section.

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 (the variable of interest in the current study) 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 and some new ones are considered in Section 4 to estimate Wages and Salaries for small areas: whenever possible, the conditional properties of these small area etimators are also discussed. Section 5 reports on the results of a simulation study that investigates the properties and behaviour of these small area estimators. Finally, Section 6 provides a summary of the study.

2.0 COMPARISONS BETWEEN THE TWO ADMINISTRATIVE FILES FOR TAX YEAR 1981

2.1 Some characteristics of the administrative files

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 for given levels of industrial classification groupings. The possible use of the Statistics Canada Combined.Master file and the Revenue Canada Comscreen file was investigated in order to determine whether those objectives could be achieved. It was therefore important to study the characteristics of these administrative files and to establish their differences in terms of the common information that they carried.

An unincorporated (T1) tax filer may operate several revenue generating sources under one Social Insurance Number (S.I.N.). These revenue sources may be business income, professional income, commission income, farming income, fishing income or rental income. Comscreen is created as a result of including in the auditing process all self-employed tax filers with incomes above established thresholds differentiated by type of income. The Combined.Master, on the other hand, only includes on a sample basis self-employed tax filers with Gross Business Income over $10,000. For tax year 1981, one of the characteristics of Comscreen was that a maximum of three income sources could be transcribed into separate segments of the transcript form. Any secondary business with expenses less than $5,000 were not copied. If more than three Profit and Loss statements were attached to a return, information from the three statements which report the largest gross income were transcribed. For the Combined.Master, however, all selected tax filers have each one of their associated business income transcribed. This difference in transcription procedures between the two files has no impact on single income tax filers, however, it may

underrepresent single or multiple business income tax filers on the Comscreen file because the largest sources of income may not be Business Income. For tax year 1981, 91.6% of the tax filers (T1) had a single business, 3.6% had two businesses and 4.8% had three or more businesses. These figures indicate that within the 8.4% of tax filers (T1) owning two or more businesses, there could be underrepresentation on the Comscreen. The extent to which this underrepresentation occurs can be determined by summing the business income reported on the three segments of the Comscreen file and comparing them to the corresponding Gross Business income total also provided within this file.

For purposes of this study, the comparison of elements between the two files was done only for single business tax filers (T1) whose Gross Business Income fell in the range $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 as matching key. The match resulted in 136,982 records which could be used to compare similar fields between the two files. The reasons for not having a 100% match between the Combined.Master and Comscreen files were as follows. The Combined.Master contained a number of incorporated records (T2) not found on the Comscreen, the Combined.Master contained a number of unincorporated records (T1) with business income less than $25,000 most likely not found on the Comscreen, some SIN were not valid on the Combined.Master. The fields of interest to be compared on the resulting matched file were: Standard Industrial Classification (SIC), Geographical Classification (SGC) and selected economic variables. Results of these comparisons are given in detail in Hidiroglou, Morry and Vaillancourt (1984a). A summary of the findings of this study is provided here.

2.2 Standard Industrial Classification Comparisons

The Industrial Classification level is at the four-digit code on the Combined.Master. For the Comscreen, the level may vary between two

digits to four digits. Some of these coding differences for the 1980 Standard Industrial Classification codes between the two files were investigated. The resulting differences documented by Hidiroglou (1984b) indicated that comparisons beyond the two digit level of SIC would not be meaningful. Therefore, the comparisons were carried out at the Major Division (1 digit SIC) and Major Group (2 digit SIC) level only. 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 within the Manufacturing Division).

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. 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.

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

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

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.

An alternative way of evaluating the discrepancies between the industrial classification codes of the two files is by measuring the impact of the coding differences when tabulating the same variables according to the two sets of codes. This type of analysis was carried out using Gross Business Income as the variable to be summed up by major division appearing on the two files.

Table 2 shows the two sets of totals as well as the percentage difference between the totals due to the discrepancies in SIC coding.

TABLE 2. Sum of GBI by STC and RCT Major Division (in $000's)

Major Division	Sum Using STC SIC Coding	Sum Using RCT SIC Coding	Percentage Difference
Logging and Forestry	144,237	136,190	- 5.5
Mining	13,962	16,155	+ 15.7
Manufacturing	394,788	264,338	- 33.0
Construction	1,667,841	1,681,853	+ .8
Transportation	640,735	736,606	+ 14.9
Communication	32,103	42,221	+ 31.5
Wholesale	367,504	252,745	- 31.2
Retail	4,185,043	4,432,538	+ 5.9
Finance and Insurance	26,163	34,535	+ 31.9
Real Estate	130,671	75,310	- 42.3
Business Service	123,375	140,222	+ 13.6
Educational Service	11,285	20,553	+ 82.1
Health and Social	46,827	61,803	+ 31.9
Accommodation	910,086	901,524	- .9
Other Services	591,518	669,773	+ 13.2

According to Table 2, the largest percentage discrepancies occur in industries that account for a small fraction of the total business activity.

The three largest industries, i.e. construction, retail trade and Accommodation show a very close agreement (percentage difference below 5%). There are, however, a few industries with relatively high weight such as manufacturing and wholesale that differ by as much as 30% in Business Income, suggesting that using the Comscreen file instead of the Combined.Master will give significantly different totals in some major divisions.

The differences observed in the Standard Industrial Classification between the two files can be fairly large. If the Comscreen is to be used as an auxiliary source of data for small area estimation, the universe counts of businesses at the chosen level of SIC and sub-provincial cross-classification should reflect Statistics Canada's SIC coding. There are two ways of doing this. One way, would be to match the Combined.Master records to the Comscreen records thereby obtaining a matched file and unmatched file of records. The matched portion could be used to work out the distribution of STC SIC codings within an RCT SIC code, using this distribution to impute STC SIC codes on the unmatched portion on each of the RCT SIC codes. The matched portion STC SIC codes could then replace the RCT SIC codes. The resulting distribution of the transformed Comscreen RCT SIC codes would then have the same distribution as the STC SIC codes. The application of this distributional approach for imputing SIC codes could be applied by working out the conditional distributions at a given level of disaggregation (i.e.: province by income subgroups between $25,000 and $500,000). This approach would assign to each record on the Comscreen file a unique SIC code and the resulting distribution would be like the one on the Combined.Master. Another approach is to obtain conditional distributions of the STC SIC coding for each RCT SIC coding and apply these to similar variables on the Comscreen file at the required small area level. To define matters more precisely, let there be L strata at some level of disaggregation of the Comscreen file (this disaggregation may be province by income subgroups between $25,000 and $500,000 for instance). Let z_{ij} be the sum of variables on the Comscreen file for matched records which have the i-th SIC code at Statistics Canada and the j-th SIC code at Revenue Canada (i=1, ..., L; j=1, ..., L): the variable may be one (to obtain a count) or Gross Business Income. In order to obtain the calibrated sum for small area "a" on the Comscreen file use

$$\hat{Z}_{ai.} = \sum_{j=1}^{L} (z_{ij} / z_{i.}) \, Z_{a.j}$$

where $z_{i.} = \sum_{j=1}^{L} z_{ij}$ and $Z_{a.j}$ is the

sum of the "z" variable for the a-th small area and j-th RCT SIC code. The definition of "z" will depend on whether counts or quantitative variables are used for the synthetic estimation process in the small areas.

2.3 Comparison of Geographical Classification Codes

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 sets 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 census division codes within 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 3.

TABLE 3. 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

As can be noted from Table 3, the level of 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 3, this code would be a good proxy for the one that would have been obtained if a business address had been available.

2.4 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 area 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 4.

TABLE 4. 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	Range 0.9 to 1.1 Excluding Zeroes
1.	Gross Business Income (SIC) vs Sales (RCT)	92%	98%
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%

The distribution of ratios between 0.9 and 1.1 provided for the above pairs of variables showed 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 0.9 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 were

drawn. Firstly, 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. The Comscreen SIC coding can be made to reflect the STC SIC coding using the procedure suggested in Section 2.2. Secondly, on the basis of records that contain both filer's address and business address, it is evident that the business activity takes places 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 Gross Business Income and Wages and Salaries according to geographic locations derived from filer's addresses would be basically equivalent to tabulating these economic variables according to the location of the business activity. Finally, 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.

3.0 MODELLING WAGES AND SALARIES

The use of the Comscreen file as an auxiliary file of population information entails the application of relationships between variables of interest on the Combined.Master by applying these to the Comscreen file variables. These relationships were obtained by regrouping the four digit 1980 Standard Industrial Codes (SIC) on the Combined.Master into 18 Major Divisions and into 76 major groups separately. 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 Gross Business Income (GBI) range between $25,000 to $500,000 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 transcribed for all tax filers with GBI over $500,000.

Previous earning models described by Lillard and Willis (1979), 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. The following is a summary of the analyses found in Hidiroglou (1984c).

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.

Table 5: Percentage of Filers with Non-Zero Wages and Salaries by Major Division and by Income Classes (GBI)

Major Division	Income class (thousands of dollars)										
	$25-$50	$50-$100	$100-$150	$150-$200	$200-$250	$250-$300	$300-$350	$350-$400	$400-$450	$450-$500	overall
Agriculture	59.8	79.1	88.9	90.5	78.9	100.0	87.5	100.0	100.0	100.0	73.4
Fishing	50.0	60.0	66.7	--	--	--	--	--	--	--	57.6
Logging and Forestry	40.8	66.4	86.5	80.6	89.5	96.4	88.2	92.3	83.3	100.0	61.3
Mining	32.9	71.4	84.0	62.5	60.0	50.0	50.0	--	--	--	54.6
Manufacturing	57.3	79.6	88.9	89.7	95.2	92.6	97.9	100.0	94.7	100.0	76.9
Construction	52.5	68.9	77.9	79.8	80.0	83.0	80.0	78.9	78.6	91.3	63.6
Transportation	40.0	57.6	71.5	87.4	92.5	93.3	88.9	100.0	91.7	100.0	53.4
Communication	59.3	80.2	85.4	92.9	100.0	100.0	100.0	--	100.0	--	70.3
Wholesale	40.8	56.3	59.1	68.9	66.2	75.2	71.4	70.3	82.0	73.9	63.4
Retail	49.9	65.1	74.1	78.6	83.4	86.9	88.7	88.9	92.4	93.1	74.0
Finance and Insurance	12.5	10.0	11.1	14.3	16.7	50.0	--	100.0	100.0	--	17.6
Real Estate	32.2	31.5	30.3	38.2	35.2	36.1	30.8	30.8	33.3	25.0	34.2
Business Service	35.5	58.3	79.8	81.1	75.9	76.2	87.5	86.7	50.0	88.9	50.5
Educational Service	56.4	78.3	93.3	100.0	--	--	--	100.0	--	--	68.9
Health and Social	75.2	87.4	90.1	84.4	90.0	83.3	83.3	75.0	80.0	75.0	81.4
Accomodation	72.2	83.2	83.9	96.7	96.7	98.3	98.1	100.0	98.7	97.8	89.0
Other Services	63.4	81.0	86.6	89.0	84.5	86.4	90.0	90.6	61.5	88.0	72.0
All Industries	51.7	68.1	77.3	81.2	83.8	86.2	87.0	86.4	88.6	91.0	68.7

As can be observed from Table 5, for most major divisions, the proportion of filers showing Wages and Salaries increases as the business income increases. Plots of the average 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 was 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 revealed that there could be a good deal of variation of Wages and Salaries for similar business income, with the scatter of the Wages and Salaries increasing as Gross Business Income increased. This suggested that a transformation of the data would 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 coefficient of determination $\bar{R}_p^2$ term, indicated that the square root transformation (models i or ii) was the best. Furthermore, the intercept term was not sufficiently significant to include in the model, resulting in the appropriateness of a ratio-type estimator for Wages and Salaries..

Having settled on the transformation, 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 changed over these intervals. As a result of the scatter plots, regression models (transformed), were tried out using Wages and Salaries in a linear and/or quadratic form or segmented polynomial form in the vector of independent variables. In order to determine if provinces had different fits, the provinces were added in to the regression as dummy variables. The following conclusions were reached at using the above fits. For the most part, the fits were linear within each major division and they differed in slope between provinces in the majority of the cases. For those major divisions which had a combination of linear and quadratic terms, although the addition of the quadratic term was at times statistically significant, the adjusted coefficient of determination ($\bar{R}^2_p$) increased only slightly. As a result of these fits, it was therefore decided that Wages and Salaries could be (for the most part) predicted using Gross Business Income as auxiliary information. Disaggregating the fits from major divisions to major groups did not significantly improve the fits.

A number of other variables common to Comscreen and to the Combined .Master files showed high correlation with Wages and Salaries. The best fits using these other variables (which included Depreciation, Net Profit, Gross Profit, and Gross Professional Income) were found using a stepwise regression procedure. These fits were done for major divisions, and major groups both at the Canada and provincial levels. The results showed that the addition of more auxiliary variables did not significantly improve the fit once the most important auxiliary variable had been taken into account. It was found that this variable was Gross Business Income with the exception of the major division Retail for which the regression using Gross Profit gave the highest $\bar{R}^2_p$. The conclusions drawn from the disaggregation of the fits for major divisions from the Canada level to the provincial level was that the $\bar{R}^2_p$ improved for some provinces in some instances while it worsened in others.

The investigation concerning the modelling of Wages and Salaries led to the following conclusions. Firstly, among the available auxiliary variables, Gross Business Income is the one most strongly related to Wages and Salaries. Secondly, in fitting the regression, one line without the intercept term provides the best fit. Thirdly, 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. Finally, 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

4.1 General Description

Wages and Salaries for small areas may be obtained in several ways using the Combined.Master and Comscreen files. If the overall sampling rate is high enough and the number of sampled units within the small area, which in our case is a cross-classification of Census Division and Major Division, then the expansion estimator (sum of weighted up data off the Combined.Master) may be good enough to produce satisfactory precision, provided that the observed number of observations is close to the expected number of observations.

The expansion estimator can be improved using the known population domain sizes (obtained from Comscreen) and observed sample domain sizes (obtained from Combined.Master). The resulting estimator, the post-stratified estimator will be denoted as POS. If the expansion (EXP) or post-stratified estimations are not 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. If 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 if the sample take within a small area is small, because strength will be "borrowed" from other small areas if the assumption holds.

Production of synthetic estimates requires the use of both the Comscreen and the Combined.Master. Two types of synthetic estimators may be contemplated, which will be referred to as the count-synthetic (COUNT-SYN) and the ratio-synthetic (RATIO SYN). 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, 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. Furthermore, it is assumed that the Comscreen file is a complete file for tax filers with Business Income over $25,000.

Procedures to correct for the bias potentially produced by the purely synthetic methods have been proposed by Schaibble (1979), Fay and Herriot (1979), Drew, Singh and Choudhry (1982), Sarndal (1984), Hidiroglou and Sarndal (1984), Battese and Fuller (1984), and Srinath and Hidiroglou (1985). These bias correction procedures involve the sum of weighted linear combinations of synthetic and direct estimators. The approach used for arriving at these weights differs amongst the aforementioned authors. Fay and Herriot (1979) model the sample means of the dependent variable on the vector of sample means of the independent variable and determine sample-

based weights that reflect the uncertainty of a linear regression fit over small area means and the sampling variability of the sample mean of the based weights that reflect the uncertainty of a linear regression fit over small area means and the sampling variability of the sample mean of the dependent variable. Battese and Fuller (1984) and Sarndal (1984), model the data across all areas using regression procedures and correct the synthetic estimators for the bias by differencing the direct estimator and synthetic estimator at the GBI group within area level, weighting up this difference and adding it to the synthetic estimator portion. The major difference between the Sarndal and Battese-Fuller procedures is that the latter makes uses of a nested-error regression model which reflects the between and within small area-variances as well as the number of observed sample units falling within each area. Drew, Singh, Choudhry (1982), and Srinath and Hidiroglou (1985) use a weighted linear combination of the small area mean and synthetic mean, basing their weighting on the observed sample counts within each small area of interest.

In what follows, the properties for some of the above mentioned estimators will be studied both theoretically (Section 4.2) and through the results obtained from a simulation study (Section 5).

4.2 Properties and Problems Associated with Small Area Estimators

For the particular problem of estimating Wages and Salaries from economic administrative files, some notation will be introduced in order to reflect both the sampling and various estimation strategies. To this end, suppose that a population of N consists of A mutually exclusive and exhaustive small areas labelled a = 1, ..., A. For each small area 'a', units are further classified into I mutually exclusive industrial groupings. Suppose that the area by industrial cross-classification can be further classified into G mutually exclusive and exhaustive income classes, labelled j = 1, ..., G. This labelling gives a three-way cross-classification into AIG possible cells with N_{aig} population members in the aig-th cell, with a corresponding sample count n_{aig} in a simple random sample of size n. To simplify the

notation, the subscript 'i' representing industrial classification will be dropped bearing in mind that the estimation is done at the small area by industrial classification detail. For aggregation across a subscript, that subscript is replaced by '.'; thus

$N_{a.} = \sum_{g=1}^{G} N_{ag}$ is the population size for the a-th area for a particular industrial classification of interest. The sample aggregates $n_{a.}$ are similarly defined. The variable y will be used to denote Wages and Salaries while the variable x will denote the Gross Business Income.

The simplest estimator to use is the expansion estimator(EXP) which utilizes only the sample data in the small area and industrial classification. For the a-th small area, it is given by:

$$t_a\ (EXP) = \frac{N}{n} \sum_{g=1}^{G} \sum_{k=1}^{n_{ag}} y_{agk.} \qquad (4.2.1)$$

For a given sample realization, this estimator is unconditionally unbiased for $Y_{a..}$, the true population total for the a-th small area. However, conditionally, it is biased for a given sample realization of size $n_{a.}$. Similarly its usual variance estimator is conditionally biased while being unconditionally unbiased. This property of conditional biasedness is not desirable because for given sample realizations of different sizes, the estimate of total will be above or below the true population value conditionally: the worst example is producing an estimate (N/n) times bigger the true total $Y_{a..}$ when $n_{a.} = N_{a.}$. An excellent description for using the conditional argument in survey sampling has been given by Rao (1985). The conditional bias for EXP given a sample realization is:

$$B\left[t_a\ (EXP) \mid n_{a.}\right] = \left(\frac{N}{n}\ n_{a.} - N_{a.}\right)\ \bar{Y}_{a..} \qquad (4.2.2)$$

while the conditional bias for the estimator of unconditional variance is:

$$B\left[v_a\ (EXP) \mid n_{a.}\right] = \frac{N(N-n)}{n(n-1)} \left\{ S_a^2 \left[(n_a - 1) + \left(1 - \frac{n_a}{n}\right)\left(1 - \frac{n_a}{N_a}\right) - \frac{n-1}{N-1}\,(N_a - 1) \right] + \bar{Y}_{a..}^2 \left[n_a \left(1 - \frac{n_a}{n}\right) - \frac{n-1}{N-1}\, N_a \left(1 - \frac{N_a}{N}\right) \right] \right\} \qquad (4.2.3)$$

where $\bar{Y}_{a.} = Y_{a..} / N_a$ and $S_a^2 = (N_a - 1)^{-1} \sum_{i=1}^{N_a} (y_{ai} - \bar{Y}_{a.})^2$

Note that for both the expressions (4.2.2) and (4.2.3), the conditional bias is equal to zero whenever n_a is equal to its expectation ($n_a = n N_a / N$).

An estimator which is both conditionally and unconditionally unbiased ($n_{ag} \geq 1$ for all a's and g's) is the post-stratified estimator given by:

$$t_a \text{ (POSG)} = \sum_{g=1}^{G} (N_{ag} / n_{ag}) y_{ag} \qquad (4.2.4)$$

For $n_{ag} = 0$, it may be defined arbitrarily - a synthetic estimator for instance. For the post-stratified estimator, the number of post-strata (G) will be dictated by the homogeneity of the units within each of the post-strata. To see this point, define two superpopulation models which reflect the behaviour of the data within each of the post-strata. The two models are

Model 1: $y_{agk} = b_2 + e_{agk}$; $\varepsilon_1 (e_{agk}) = 0$; $\varepsilon_1 (e^2_{agk}) = \sigma^2_a$,

Model 2: $y_{agk} = b_{ag} + e^*_{agk}$; $\varepsilon_2 (e^*_{agk}) = 0$; $\varepsilon_2 (e^{*2}_{agk}) = \sigma^2_{ag}$

where $b_{a1} \neq \ldots \neq b_{aG}$ and $\sigma^2_{a1} \neq \ldots \neq \sigma^2_{aG}$

for $a = 1, \ldots, A$; $g = 1, \ldots, G$ and $k = 1, \ldots, N_{ag}$.

These models may be used to compare the conditional variance for one group $V \left[t_a \text{ POS 1)} \mid n_a \right]$ versus the conditional variance for G groups, $V \left[t_a \text{ (POSG)} \mid n_{ag} \right]$. Under model 1, the difference is

$$\varepsilon_1 \left\{ V \left[t_a \text{ (POSG)} \mid n_a \right] - V \left[t_a \text{ (POS1)} \mid n_{ag} \right] \right\}$$ which can

be shown to be equal to

$$\frac{1}{n_a} \left\{ N_{a1} \left(\frac{n_a - n_{a1}}{n_{a1}} \right)^{1/2} - \sum_{g=2}^{G} N_{ag} \left(\frac{n_a - n_{ag}}{n_{ag}} \right)^{1/2} \right\}^2$$

$$- \frac{1}{n_a} \left\{ \sum_{g \neq h} N_{ag} N_{ah} \left[1 - \left(\frac{n_a - n_{ag}}{n_{ag}} \right)^{1/2} \left(\frac{n_a - n_{ah}}{n_{ah}} \right)^{1/2} \right] \right\} \qquad (4.2.5)$$

Now, the first part of this equation is a square and hence greater than zero. It remains to show that all the cross-product factors in its second part are negative for all g and h (g=h). That is, one must prove that

$$1 - (n_a / n_{ag} - 1)^{1/2} \; (n_a / n_{ah} - 1)^{1/2} \leq 0. \qquad (4.2.6)$$

To this end, let $n_{am} = \max (n_{ag}: g=1, \ldots, G)$. If $n_{am} \leq 0.5\, n_a$, then equation (4.2.6) must hold. If $n_{am} > 0.5\, n_a$, let $n_{am} = cn_a$ where $c > 0.5$. The maximum that the remaining n_{ag}'s ($g \neq m$) can attain is $(1-c)\, b_a$. Now for this case, equation (4.2.6) again holds. Hence, no grouping is required when model 1 holds.

Under model 2, the difference between the two conditional variances of t_a (POSG) and t_a (POS1) can be shown to be equal to

$$\sum_{g=1}^{G} N_{ag}\, \sigma_{ag}^2 \; (N_a/n_a - N_{ag}/n_{ag}) + (N_a^2/n_a - N_a) \; \frac{N_a}{N_a - 1} \sum_{g=1}^{G} W_{ag} \; (b_{ag} - \bar{b}_a)^2 \qquad (4.2.6)$$

where $W_{ag} = N_{ag} / N_a$ and $\bar{b}_a = (\sum_{g=1}^{G} N_{ag}\, b_{ag}) \; / \; N_a$. Equation (4.2.6) may be positive or negative, depending on the configurations of n_{ag} vis-à-vis b_{ag}. This implies that the number of post-strata required to achieve optimality may conditionally be different between each sample realization.

Given that t_a (POSG) is both conditionally and unconditionally unbiased, for $n_{ag} \geq 1$, how should its variance be computed to yield meaningful confidence intervals? Conditionally or unconditionally? To simplify matters, assume that one group is used, G=1. The unconditional variance estimator of t_a (POS1) minus its associated conditional analogue for a given sample realization n_a is

$$v\,[(t_a\,(\mathrm{POS1})] - v\,[t_a\,(\mathrm{POS1}) \mid n_a]$$

$$= N_a^2 \left\{ \sum_{k=1}^{n_a} (y_{ak} - \bar{y}_a)^2 \right\} \left\{ \frac{1}{n_a} \left(\frac{1}{N_a} - \frac{n}{N\, n_a} \right) \right\}. \qquad (4.2.7)$$

Expression (4.2.7) is positive if $n_a > \frac{n}{N} N_a$. This implies that confidence intervals based on the unconditional variance will be too long for sample size realizations n_a greater than the expected size $n N_a / N$ and too short otherwise. Therefore, confidence intervals should be computed using conditional variances in order to achieve the nominal levels.

Synthetic estimators which have been considered in the present context are of the form

$$t_a \text{ (SYNG)} = \sum_{g=1}^{G} Z_{ag.} \left(\bar{y}_{.g.} / \bar{z}_{.g.} \right) \qquad (4.2.8)$$

where the "z" variable may be a count or auxiliary variable related to Wages and Salaries (Gross Business Income). The unconditional bias of such an estimator is given by:

$$B \left[t_a \text{ (SYNG)} \right] = \sum_{g=1}^{G} \left(\frac{\bar{Y}_{.g.}}{\bar{Z}_{.g.}} - \frac{\bar{Y}_{ag.}}{\bar{Z}_{ag.}} \right) Z_{ag.} \qquad (4.2.9)$$

where $Z_{ag.} = \sum_{k=1}^{N_{ag}} z_{agk}$, $\bar{Z}_{ag.} = Z_{ag.} / N_{ag}$

and $\bar{Z}_{.g.} = \left(\sum_{a=1}^{A} Z_{ag.} / \sum_{a=1}^{A} N_{ag} \right)$ is the synthetic "Z"

component. This form of the synthetic estimator can be improved in the current context, noting that there has been a 25% sample (quite a high sampling rate). The improved synthetic estimator is of a form in accordance with Royall (1978). It is given by

$$t_a \text{ (PREDG)} = \sum_{g=1}^{G} y_{ag.} + \sum_{g=1}^{G} \left(Z_{ag.} - z_{ag.} \right) \frac{\bar{y}_{.g.}}{\bar{z}_{.g.}} \qquad (4.2.10)$$

and its unconditional bias is

$$B \left[t_a \text{ (PREDG)} \right] = \left(1 - \frac{n}{N}\right) B \left[t_a \text{ (SYNG)} . \right]$$

t_a (PREDG) is based on a predictive type of argument which can be summarized as follows. An estimator of a given variable within a small area is the sum of the units falling within the small area plus a predicted portion based on the synthetic estimator for elements not sampled and within the small area. This formulation of small area estimation points to a strategy which combines direct and synthetic estimators according to some pre-set criteria. Several strategies for achieving this will now be reviewed. The estimators are of the form,

$$t_a\,(\text{MIXG}) = \sum_{g=1}^{G} \left(W_{1ag}\,\bar{y}_{.g.} + W_{2ag}\,\bar{y}_{ag.}\right) \qquad (4.2.11)$$

where W_{1ag} and W_{2ag} are weights which depend on the strategy used.

In the above context, Sarndal (1984) proposed asymptotically design unbiased estimators that incorporate auxiliary information through the use of the generalized regression technique. For the two special cases of this technique included in this paper, since the sampling is simple random, the weights are of the form

$$W_{1ag}\,(\text{SARG}) = \left(\frac{Z_{ag.}}{\bar{z}_{.g.}} - \frac{N}{n}\,n_{ag}\,\frac{\bar{z}_{ag.}}{\bar{z}_{.g.}}\right) \qquad (4.2.12)$$

and

$$W_{2ag}\,(\text{SARG}) = \frac{N}{n}\,n_{ag}$$

This estimator is nearly unbiased, however, it has two drawbacks; (i) its variance can be considerable in some small domains and (ii) it can take on negative values in situations that do not allow such values. For the case of G=1, negative values will occur whenever W_{1a} is negative and $|W_{1a}\,\bar{y}_{...}| > |W_{2a}\,\bar{y}_{a..}|$. This undesirable feature has been substantially reduced by Hidiroglou and Sarndal (1984), by applying a dampening factor which slightly biases the original Sarndal procedure and brings down the mean square error dramatically. These weights are of the form:

$$W_{1ag}\,(\text{MSARG}) = \left(\frac{Z_{ag.}}{\bar{z}_{.g.}} - F_a\,\frac{N}{n}\,n_{ag}\,\frac{\bar{z}_{ag.}}{\bar{z}_{.g.}}\right) \qquad (4.2.13)$$

and

$$W_{2ag}\ (\text{MSARG}) = F_a \frac{N}{n} n_{ag} \ ,$$

with

$$F_a = \begin{cases} E_a / n_{a.} & \text{if } n_{a.} \geq E_a \\ n_a / E_a & \text{if } n_{a.} < E_a \end{cases}$$

where $E_a = n\,(N_{a.} / N)$.

As can be observed from (4.2.13), the likelihood of obtaining negative values is reduced with this modification. The estimator (4.2.11) with the weights given by (4.2.12) will be denoted as t_a (SARG). This estimator is conditionally biased for a given sample realization $n_{a.}$ within a small area "a" and unconditionally asymptotically unbiased for $0 \leq n_a \leq N_a$. This conditional bias is given by

$$B\left[t_a\ (\text{SARG}) \mid n_{a.}\right] = \sum_{g=1}^{G} N_{ag} \left(\frac{N}{n}\frac{n_a}{N_a} - 1\right)\left(\bar{Y}_{ag.} - \bar{Z}_{ag.}\frac{\bar{Y}_{.g.}}{\bar{Z}_{.g.}}\right) \tag{4.2.14}$$

The conditional bias of t_a (SARG) will be bigger than the one associated with t_a (SYNG)

if $$\sum_{g=1}^{G} N_{ag} \frac{N}{n}\frac{n_a}{N_a}\left(\bar{Y}_{ag.} - \bar{Z}_{ag.}\frac{\bar{Y}_{.g.}}{\bar{Z}_{.g.}}\right) > 0 \ .$$

The conditional bias associated with t_a (MSARG), the estimator (4.2.11) with weights given by (4.2.13) is

$$B\left[t_a\ (\text{MSARG}) \mid n_{a.}\right] = \begin{cases} 0 & \text{for } n_a \geq E_a \\ \sum_{g=1}^{G} N_{ag}\left(F_a \frac{N}{n}\frac{n_a}{N_a} - 1\right)\left(\bar{Y}_{ag.} - \bar{Z}_{ag.}\frac{\bar{Y}_{.g.}}{\bar{Z}_{.g.}}\right) & \text{for } n_a < E_a \end{cases} \tag{4.2.15}$$

The interesting point to note about t_a (MSARG) is that it is conditionally unbiased above the expected sample size E_a within a small area, yet, conditionally biased below that value. It must also be noted that when $n_a = 0$, t_a (SYNG), t_a (SARG) and t_a (MSARG) have all the same conditional bias. Since $F_a \leq 1$, the likelihood of getting negative estimates of totals has been drastically reduced for $n_a \leq E_a$. For $n_a \geq E_a$, the likelihood of obtaining negative values is even more remote. Another advantage of t_a (MSARG) over t_a (SARG) is that for $n_a \geq E_a$, meaningful unbiased conditional confidence intervals can be provided. For $n_a < E_a$, and as n_a approaches zero, the conditional confidence intervals are not meaningful since the estimator becomes more biased: this, however, is a problem shared by any type of estimator in the form given by expression (4.2.11).

Srinath and Hidiroglou (1985) eliminate the problem of negative estimates of totals by constructing weights of the form:

$$W_{1ag}(SH) = \frac{(Z_{ag} - \hat{Z}_{ag})(\hat{z}_{ag} / z_{ag} - 1) Z_{ag}}{\hat{Z}_{ag}(z_{ag} / \hat{z}_{ag}) \bar{z}_{ag} + (Z_{ag} - \hat{Z}_{ag})(\hat{z}_{ag} / z_{ag} - 1) \bar{z}_{.g}}$$

$$W_{2ag}(SH) = \frac{\hat{Z}_{ag}(z_{ag} / \hat{z}_{ag}) Z_{ag}}{\hat{Z}_{ag}(z_{ag} / \hat{z}_{ag}) \bar{z}_{ag} + (Z_{ag} - \hat{Z}_{ag})(\hat{z}_{ag} / z_{ag} - 1) \bar{z}_{.g}} \qquad (4.2.14)$$

where $\hat{Z}_{ag} = z_{ag}(Z_{..} / z_{..})$; $\hat{z}_{ag} = Z_{ag}(z_{..} / Z_{..})$

These weights have the following characteristics. Firstly, when no sampled units fall within the small area, the synthetic estimator is used. The expansion estimator (or ratio analogue is used only if the number of sampled units within the small area represents the expected value: that is,

$z_{ag} / Z_{ag} = z_{..} / Z_{..}$. The bias of this estimator can be reduced and it can be made consistent, $t_{a.}$ (SHG) = $Y_{a..}$ when $n_{ag} = N_{ag}$ for all g=1, ..., G, by making use of the Royall type of estimator given by expression (4.2.10). The weights then become

$$W_{1ag}(P/SH) = \frac{(Z_{ag} - \hat{Z}_{ag})\ (\hat{z}_{ag} / z_{ag} - 1)\ (Z_{ag} - z_{ag})}{\hat{Z}_{ag}\ (z_{ag} / \hat{z}_{ag})\ \bar{z}_{ag} + (Z_{ag} - \hat{Z}_{ag})\ (\hat{z}_{ag} / z_{ag} - 1)\ \bar{z}_{.g}}$$

$$W_{2ag}(P/SH) = \frac{\hat{Z}_{ag}\ (z_{ag} / \hat{z}_{ag})\ Z_{ag} + n_{ag}\ (Z_{ag} - \hat{Z}_{ag})\ (\hat{z}_{ag} / z_{ag} - 1)\ \bar{z}_{.g}}{\hat{Z}_{ag}\ (z_{ag} / \hat{z}_{ag})\ \bar{z}_{ag} + (Z_{ag} - \hat{Z}_{ag})\ (\hat{z}_{ag} / z_{ag} - 1)\ \bar{z}_{.g}}$$

(4.2.15)

Battese and Fuller (1984) obtained conditional total estimates for small areas using a regression approach whose error structure reflected a nested between and within area error component. The form of their weights for a Royall type of estimation and one group (G=1) is:

$$W_a(P/BF) = \left[(Z_a - z_a) - (N_a - n_a)\ d_a\ \bar{z}_{a.}\right] / \bar{z}_{..}$$

(4.2.16)

$$W_{2a}(P/BF) = \left[n_a + (N_a - n_a)\ d_a\right]$$

where d_a is a dampening factor $(0 \leq d_a \leq 1)$ within each area determined as a function of the estimated between and within area components of variance, and the observed sample sizes n_a. The variable d_a is obtained in an optimal manner and reflects by how much the synthetic estimator should be corrected with the difference of the sample mean and synthetic value. The case $d_a = 1$ yields an estimator which is similar to t_a (SAR1).

Outliers in the present context, that is simple random sampling without replacement, should be identified and their weight should be modified in order to decrease their upward impact on the conditional bias. The approach suggested here is the one proposed by Hidiroglou and Srinath (1981): assign a weight of one to outlying observations and re-adjust the weights for the remaining units in order that the sum of the weights adds up to the known population total count. Also, outlying observations must not

be included in the computation of the purely synthetic estimates. The estimation for the estimator (4.2.11) would be regarded as being done for the population units excluding the sampled outlier units, plus the outlier units counting for themselves. The estimators would be of the form

$$t_a\ (\text{MIXG/OUT}) = \sum_{g=1}^{G} (W^*_{1ag}\ \bar{y}^*_{.g.} + W^*_{2ag}\ \bar{y}_{ag}) + \sum_{g=1}^{G} \sum_{k=1}^{n^o_{ag}} y^o_{ag}, \quad (4.2.17)$$

where n^o_{ag} is the number of outlier units falling within area "a" and group "g", y^o_{ag} is the y value for these outlying units. The star '*' over the other symbols denoted that the weights have been adjusted and that the averages exclude the outlying values.

4.3 Estimating Small Areas Across Time

In the previous section, some of the most widely used as well as some new small area estimation techniques have been discussed. These methods are valid for a given time period and do not relate to any other time period. It is known, however, that large organizations compile administrative files on a regular basis. This regular compilation can be used to advantage if there is a good correlation between the same variables across time.

In our present context, the larger and more complete file (Comscreen), provides auxiliary information such as counts or sums of variables at very low level of disaggregation, while the other file (Combined.Master) provides relationships such as domain means or domain ratios. The following notation will summarize the situation at hand. Assume that at time t, we have

$N(t)$ - the size of the overall universe,

$N_a(t)$ - the size of the 'a'th small area population,

$N_{ag}(t)$ - the size of the 'a'th small area and 'g'th group population,

$n(t)$ - the size of the sample,

$n_a(t)$ - the size of the observed sample within the 'a'th small area,

$n_{ag}(t)$ - the size of the observed sample within the 'a'th small area and 'g'th group,

$y_{agk}(t)$ - the observed dependent value for the 'k'th unit within small area 'a' and the 'g'th group,

$\underset{\sim}{x}_{agk}(t)$ - the observed p-dimensional vector for the independent variables associated with the 'k'th unit within small area 'a' and the 'g'th group.

Assume that data is available for both administrative files for a period of time spanning $T_1 \leq t \leq T_2$. Populations will undergo births and deaths between two given time periods, with a block remaining common through time. Hence, the population at time t (current) can be expressed as the union of population units common to time s (previous) and time t plus the births since time s(s t). That is

$$P(t) = C(t,s) \cup B(t,s)$$

where C(t,s) = common units between times 's' and 't' and B(t,s) = births since time 's' at time 't'. Note that in terms of set notation, $C(t,s) = P(t) \cap P(s)$ and $B(t,s) = P(t) \cap \bar{P}(s)$. Population $P(T_2)$ can be expressed as the union of intersecting sets with populations $P(T_i)$ (i=1, ..., T_2-1) and the birthed population B (T_2, T_2-1). The correlation across time for a given variable "z" is most likely to be the largest for time periods that are close, that is:

$$\rho\,[z(t), z(s)] > \rho\,[z(t), z(s-1)]$$

where ρ is the correlation coefficient.

Let E(t) be the sample drawn at time t. Bearing in mind, that the correlation is strongest for units closer in time, $P(T_2)$ the population at time T_2 can be decomposed as the following union of non-overlapping subsets,

$E^*(T_2, T_2\text{-}i)$ $(i=0, 1, \ldots, T_1)$ and $P^*(T_2)$ where

$$E^*(T_2, T_2) = E(T_2),$$

$$E^*(T_2, T_2\text{-}i) = E(T_2\text{-}i) \cap P(T_2) \cap \bar{C}(T_2\text{-}i) \quad (i = 1, \ldots, T_1),$$

$$C(T_2\text{-}i) = \pi_{j=0}^{i-1} \left[\cup E^*(T_2, T_2\text{-}j)\right] \quad (i = 1, \ldots, T_1),$$

and

$$P^*(T_2) = P(T_2) \cap \left\{ \pi_{j=0}^{T_1} \left[\cup C(T_2\text{-}j)\right] \right\}.$$

The above decomposition of $P(T_2)$ is one which regroups the population into layers of units which are most highly correlated with the current time period T_2. This decomposition can be represented in bar form as follows:

$P^*(T_2)$	$E^*(T_2, T_2\text{-}T_1)$	. . .	$E^*(T_2, T_2\text{-}1)$	$E^*(T_2, T_2)$

Note that $P(T_2) = P^*(T_2) \cup \left\{ \pi_{j=0}^{T} \left[\cup E^*(T_2, T_2\text{-}j)\right] \right\}$

$$P^*(T_2) \cap E^*(T_2, T_2\text{-}j) = \emptyset \text{ (the null set)}, \quad j=0, \ldots, T_1;$$

$$E^*(T_2, T_2\text{-}i) \cap E^*(T_2, T_2\text{-}j) = \emptyset; \quad i=0, \ldots, T_1; \quad j=0, \ldots, T_2; \quad i \neq j.$$

In order to estimate relationships between two time periods s and t, $T_1 \leq s < t \leq T_2$, the following subsets are defined:

$$E(T_2, T_2\text{-}i) = E(T_2) \cap E(T_2\text{-}i); \quad i = 0, \ldots, T_1.$$

For a given small area 'a' and group 'g', the population $P_{ag}(T_2)$ can also be decomposed into the union of non-overlapping subsets (all may be non-empty), $P^*_{ag}(T_2)$, $E^*_{ag}(T_2, T_2 - T_1)$, ..., $E^*_{ag}(T_2, T_2-1)$, $E^*_{ag}(T_2, T_2)$ whose union is $P_{ag}(T_2)$. These subsets are defined as previously over the cross-classification of small areas and groups. Denote as $n^*_{ag}(T_2, T_2-i)$, i=0, ...,T_1), the observed sample sizes within the 'a'th small area and 'g'th group common to periods T_2 and T_2-j. The non-observed set, the one that is never sampled from times T_1 through to T_2 is of size $n^*_{ag}(T_2) = N_{ag}(T_2) - \sum_{j=0}^{T_1} n^*_{ag}(T_2, T_2-j)$.

At time T_2, only the set $E^*(T_2, T_2)$ has been observed, with the sets $E^*(T_2, T_2-T_1)$ ... $E^*(T_2, T_2-1)$ having been observed at some prior time period. In order to obtain the y- value for each of the non-observed subsets $E^*(T_2, T_2-i)$ (i = 1, ..., T_1), the following lagged regression model can be used:

$$y_{agk}(T_2) = \underset{\sim}{z}_{agk}(T_2-i)\,\underset{\sim}{\beta} + e_{agk}(T_2)$$

where e_{agk} is distributed i.i.d. with mean zero and variance

$$\sigma^2(T_2),\ \underset{\sim}{z}_{agk}(T_2-i) = (y_{agk}(T_2-i),\ \underset{\sim}{x}_{agk}(T_2-i),\ \underset{\sim}{x}_{agk}(T_2))$$

The 'agk'th observation can then be obtained as $\hat{y}_{agk}(T_2) = z_{agk}(T_2)\hat{\underset{\sim}{\beta}}$. A predictive Royall type of estimation for the 'a'th small area is then

$$t_a(T_2) = \sum_{g=1}^{G} \left\{ \sum_{k\in E(T_2)} y_{agk}(T_2) + \sum_{t=T_2-T_1}^{T_2-1} \sum_{k\in E^*(T_2,\ T_2-t)} \hat{y}_{agk}(T_2) + \sum_{k\in P^*(T_2)} \tilde{y}_{agk}(T_2) \right\}$$

where for $k \in P^*(T_2)$, $\tilde{y}_{agk}(T_2)$ is estimated using one of the previously mentioned estimation procedure given in Section 4.2.

5.0 RESULTS FROM THE EMPIRICAL STUDY

In order to study the properties of the various estimators discussed in the previous section, a simulation was undertaken. The province of Nova Scotia was chosen as our population with N = 1,678 unincorporated tax filers. The small areas of interest were 18 Census divisions within that province. The major industrial groups studied within these areas were Retail (515 units in the population), Construction (496 units in the population), Accommodation (114 units in the population and the remaining industries grouped into Others (553 units in the population). The variable of interest was Wages and Salaries (available on a sample basis) and the auxiliary variables were either counts or Gross Business Income (available on a 100% basis). The overall correlation coefficients between Wages and Salaries and Gross Business Income were 0.42 for Retail, 0.64 for Construction, 0.78 for Accommodation and 0.61 for Others.

For the Monte Carlo simulation, 500 samples, each of size 419, were selected using simple random sampling without replacement from the target population of 1,678 unincorporated tax filers. The selected sample units were classified into type of industry and Census Division. Two types of income groupings were used: (1) G = 3 income classes given by \$25K - \$50K, \$50K - \$150K, and \$150K - \$500K; (2) G = 1 given by \$25K - \$500K. The main findings are summarized in three layers. The first layer is the highest level of aggregation at the Industrial level. The second layer is aggregated at selected Industrial by Census Division level. The third layer is aggregated by summarizing selected statistics for each sample realization within the Industrial by Census Division level.

For each layer, the main findings are discussed with respect to (a) relative bias of the estimators; (b) relative efficiency; (c) root mean square error. Denoting as R the number of times a sample was included in a given cross-classification, the relative bias is computed as

$$\overline{R\,B}\,(\text{ESTIM}) = \frac{1}{R}\sum_{r=1}^{R}\left[\frac{t^{(r)}\,(\text{ESTIM})}{Y} - 1\right]$$

where $t^{(r)}$ (ESTIM) is the value for a particular estimator ESTIM of the r-th Monte Carlo sample and Y is the known population total.

The relative efficiency of t(ESTIM) with respect to the direct estimator (EXP) is computed as

$$\overline{EFF}\ (ESTIM) = \left\{ \frac{\overline{MSE}\ (EXP)}{\overline{MSE}\ (ESTIM)} \right\}^{1/2}$$

where $\overline{MSE}\ (ESTIM) = \frac{H}{R} \sum_{r=1}^{R} \left[t^{(r)}\ (ESTIM) - Y \right]^2$

Note that $\overline{RB}$ (ESTIM) and $\overline{MSE}$ (ESTIM) have been given for layer 3. For the other layers, the averaging process was done over necessary classifications required for collapsing to the highest levels.

Estimators requiring counts as auxiliary information are denoted as (ESTIM/C) while estimators using Gross Business Income as auxiliary information are denoted as (ESTIM/R). The estimators that are considered in this study are the post-stratified estimators (POS1/C and POS1/R), the expansion estimator (EXP), the Srinath-Hidiroglou estimators (SH1/C and SH1/R), the modified Sarndal estimator (MSAR1/C, MSAR3/C, MSAR1/R, MSAR3/R), the Battese-Fuller estimator (BF1/R) and the synthetic estimator (SYN1/C, SYN1/R, SYN3/C, SYN3/R).

Tables 6 and 7 examine the performance of the estimators in terms of relative bias and efficiency respectively at the layer 1 level of analysis. In terms of bias, the purely synthetic estimators (SYN) display the most bias followed by the nested error model estimator (BF). The unbiased estimator EXP shows negligible relative bias (0.04). The post-stratified estimator (POS) has a large relative negative bias for "Accommodation", but this is due to a non negligible probability of getting no sampled units in the cell. In "Retail", the effect of using a regression estimator for the modified regression estimator (MSAR) is to increase the relative bias while in "Accommodation" the revise trend occurs. In terms of relative bias, the positive weight estimator (SH) behaves roughly like the modified regression estimator (MSAR). The effect of increasing the number of income classes from G=1 to G=3 has been to lower the relative bias for all the estimators considered.

TABLE 6: Relative Bias for the Estimators

Estimators		Industrial Group			
		Retail	Construction	Accommodation	Others
1.	EXP	0.023	0.019	0.036	0.017
2.	POS1/C	-0.082	-0.048	-0.264	-0.012
3.	POS1/R	-0.021	-0.043	-0.256	-0.033
4.	SYN1/C	0.111	0.049	0.353	0.188
5.	SYN3/C	0.167	0.062	0.199	0.157
6.	SYN1/R	0.246	0.038	0.240	0.147
7.	SYN3/R	0.195	0.040	0.199	0.144
8.	SH1/C	0.035	0.010	0.187	0.042
9.	SH1/R	0.113	0.012	0.170	0.037
10.	MSAR1/C	0.039	0.012	0.182	0.063
11.	MSAR3/C	0.069	0.020	0.114	0.050
12.	MSAR1/R	0.111	0.012	0.138	0.046
13.	MSAR3/R	0.085	0.013	0.120	0.045
14.	BF1/R	0.174	0.019	0.154	0.089

TABLE 7: Relative Efficiency for the Estimators With Respect to EXP

Estimators		Industrial Group			
		Retail	Construction	Accommodation	Others
1.	POS1/C	1.345	1.352	1.293	1.177
2.	POS1/R	1.241	1.860	1.860	1.641
3.	SIN1/C	1.880	1.465	2.074	1.714
4.	SYN3/C	1.954	2.016	2.493	1.945
5.	SYN1/R	2.269	1.925	3.265	2.041
6.	SYN3/R	2.367	1.923	3.116	2.016
7.	SH1/C	1.513	1.539	1.705	1.347
8.	SH1/R	1.669	2.154	3.395	1.863
9.	MSAR1/C	1.641	1.645	1.809	1.445
10.	MSAR3/C	1.804	2.098	2.383	1.762
11.	MSAR1/R	1.788	2.194	3.352	1.893
12.	MSAR3/R	1.857	2.183	3.190	1.891
13.	BF1/R	2.077	2.374	3.555	2.335

In terms of relative efficiency, all estimators are significantly better than the unbiased estimator EXP. The post-stratified estimator (POSI/R) based on regression is more efficient than the one based on the counts (POSI/C), especially for those industrial divisions where the correlation between Wages and Salaries and Gross Business Income is high. The synthetic estimators (SYN) are, on the average, more efficient than the post-stratified estimators. For all estimators concerned, the division of the income classes into 3 domains as opposed to 1 domain, improves the efficiency significantly in terms of counts. However, this disaggregation for the regression based estimators may or may not improve the efficiency. This observation is in agreement with the theoretical results obtained for the post-stratified estimators (see Section 4.2). The division of income classes for the estimators based on the counts will have the effect to approximate the best fit obtained via regression on auxiliary information. The uniformly most efficient estimator over the range of mixture estimators is the nested error model (BF1/R). However, it gains efficiency at the expense of acquiring more bias. The positive weight estimator based on the regression relationship (SH1/R) behaves like the modified regression estimator (MSAR1/R) and it is sometimes more efficient (Accommodation). For all estimators using a regression relationship, their efficiency is highest for those industrial groups with the highest correlation between Wages and Salaries and Gross Business Income (Accommodation).

Tables 8, 9, and 10 examine the performance of the estimators in terms of efficiency at the layer 2 level of analysis. Retail and Construction are industrial groups with low and medium correlation for areas of similar size, while Accommodation is an industrial group with high correlation for areas that are smaller than those for the previously mentioned industrial groups. For all the estimators concerned, there is a tendency for the Root Mean Squared Error to increase as the mean sample take increases. This tendency, however, is not consistent, and depends on the magnitude of the bias associated with the individual small area. For all industrial groups concerned, the nested error model (BF1/R) estimator has smaller Root Mean Squared Error than the corresponding mixture estimators. The relative gain of BF1/R over the modified regression estimator MSAR1/R is however not dramatic. The separation of the income groups for the synthetic estimator using the regression (SYN/R) provides no clear advantage. It depends on

TABLE 8 : Industrial Group - Retail Areas: 18 Census Divisions in Nova Scotia. Root Mean Squared Error of Each of Ten Estimators Over 500 Repeated Simple Random Samples from the Entire Population

Area	Mean Sample Take	EXP	POS1/C	POS1/R	SYN1/R	SYN3/R	SH1/C	SH1/R	MSAR1/C	MSAR1/R	MSAR3/C	MSAR3/R	BF1/R
13	0.478	25.1	15.6	15.8	28.5	20.5	2.9	22.5	2.9	23.6	9.7	16.9	23.3
18	0.542	31.6	17.5	17.3	20.5	13.6	9.1	17.5	10.1	17.6	13.6	13.5	16.8
1	1.762	56.6	46.9	87.0	24.5	21.5	29.8	28.6	28.1	33.0	26.5	27.9	31.3
16	2.244	56.9	46.1	49.6	93.3	73.0	32.1	49.4	30.0	54.0	39.5	41.7	54.9
14	2.798	62.2	53.3	73.6	5.2	19.4	45.6	36.8	42.3	37.4	35.4	37.0	30.7
4	3.024	75.0	60.4	70.0	15.1	16.0	43.9	33.2	40.8	38.4	27.3	34.7	28.4
3	3.896	102.3	82.8	159.1	13.3	12.0	55.0	54.3	49.5	59.5	50.9	48.3	52.1
15	4.212	89.9	59.2	69.1	56.8	51.4	44.1	49.4	41.9	51.1	43.4	44.5	46.7
2	5.448	206.3	156.9	159.3	87.5	114.5	131.9	104.8	125.7	94.9	112.8	101.2	75.5
8	5.642	109.8	92.9	82.1	63.5	54.5	73.6	62.4	66.3	58.6	60.7	59.4	49.3
5	5.932	120.6	98.3	103.7	37.6	35.9	76.8	73.1	70.2	70.7	71.3	63.1	52.8
6	7.628	110.9	75.4	74.5	24.8	35.4	65.8	47.9	63.2	46.1	58.5	41.8	36.1
11	8.346	153.3	111.8	134.8	58.9	48.4	98.9	102.9	87.3	100.0	82.0	88.8	71.9
7	8.610	186.7	134.1	190.8	87.8	90.3	115.3	129.4	103.9	114.8	96.8	109.6	91.0
10	8.920	148.6	97.1	100.6	49.9	44.8	85.4	85.4	77.3	80.9	76.6	73.2	61.2
12	10.576	215.8	147.8	169.7	124.0	132.1	133.5	133.2	128.1	118.5	116.4	110.5	99.3
17	23.950	284.9	218.5	188.0	112.1	87.6	210.5	175.6	193.3	163.0	170.1	165.7	144.6
9	24.640	270.1	201.0	174.5	88.3	71.1	193.9	159.4	176.4	144.1	155.0	145.9	144.0

TABLE 9: Industrial Group Construction Areas: 18 Census Divisions in Nova Scotia. Root Mean Squared Error of Each of Ten Estimators Over 500 Repeated Simple Random Samples From the Entire Population

Area	Mean Sample Take	EXP	POS1/C	POS1/R	SYN1/R	SYN3/R	SH1/C	SH1/R	MSAR1/C	MSAR1/R	MSAR3/C	MSAR3/R	BF1/R
13	1.290	64.2	43.4	39.8	21.2	21.8	25.6	17.8	21.1	16.1	28.7	16.7	15.7
16	1.958	61.6	42.2	46.7	8.0	9.2	26.2	26.6	26.7	29.1	21.3	29.7	21.0
15	2.010	66.7	49.7	44.5	12.1	12.5	34.7	20.2	34.9	21.1	24.0	21.2	18.7
4	2.046	152.5	121.2	86.9	83.7	82.6	80.2	63.3	80.3	70.0	73.6	69.3	63.2
1	2.068	81.0	62.2	54.3	39.8	40.4	37.5	29.2	37.4	31.2	26.2	32.2	29.4
18	2.114	97.1	68.1	59.2	17.0	17.4	41.1	26.6	41.1	27.2	35.4	28.2	23.3
14	2.794	144.8	116.2	94.4	19.5	21.8	77.8	37.6	68.3	36.1	30.6	38.1	33.2
3	3.036	155.1	136.0	80.4	46.8	47.5	95.9	49.7	89.3	54.7	58.1	54.8	43.9
5	4.502	135.3	108.1	64.8	19.4	19.8	81.0	43.3	73.0	41.1	58.0	40.2	32.4
8	6.620	162.4	114.1	76.9	52.3	52.0	94.2	57.6	83.9	54.6	63.5	54.2	42.5
2	6.818	181.7	140.9	91.5	41.1	42.9	113.7	72.7	98.8	65.8	67.7	68.5	49.4
12	6.988	167.1	117.6	115.9	153.4	158.6	93.8	101.7	85.2	109.5	92.0	110.8	106.7
11	7.632	181.6	141.9	89.0	24.7	25.8	120.3	64.9	107.6	59.5	59.5	59.6	44.1
10	9.606	265.6	187.2	139.6	70.5	80.7	167.1	124.6	151.7	113.1	138.7	114.1	85.0
6	10.216	397.4	283.2	179.0	354.7	351.8	263.3	180.5	254.2	191.8	188.1	192.0	211.2
7	11.922	248.4	177.3	113.9	50.3	53.2	163.1	96.1	147.7	90.5	88.7	93.5	73.7
17	13.238	255.7	167.4	112.0	48.2	51.3	157.7	98.1	143.9	92.9	99.4	95.5	74.3
9	29.154	365.7	285.4	239.4	137.5	130.7	279.4	230.6	268.0	216.7	233.9	215.4	193.7

TABLE 10: Industrial Group: Accommodation Areas: 18 Census Divisions in Nova Scotia. Root Mean Squared Error of Each of Twelve Estimators Over 500 Repeated Simple Random Samples From the Entire Population

Area	Mean Sample Take	EXP	POS1/C	POS1/R	SYN1/R	SYN3/R	SH1/C	SH1/R	MSAR1/C	MSAR1/R	MSAR3/C	MSAR3/R	BF1/R
4	0.226	12.4	6.5	6.5	1.5	1.7	10.5	1.3	10.5	1.3	2.4	1.6	1.3
1	0.252	33.7	16.8	16.8	7.2	7.5	3.0	6.3	3.0	6.3	2.7	6.6	6.3
18	0.516	16.9	9.9	23.5	38.7	40.0	19.3	35.3	19.2	38.7	47.1	41.6	35.3
3	1.016	29.6	21.0	19.5	9.6	4.8	30.6	10.9	30.5	10.9	12.8	10.7	8.7
14	1.044	226.8	160.1	119.0	63.2	63.0	136.8	47.6	138.4	60.0	90.6	56.4	52.9
10	1.258	47.1	31.5	28.0	19.5	17.8	27.1	12.9	27.0	12.0	19.0	12.2	12.5
2	1.370	86.4	71.2	46.4	27.1	27.3	48.0	19.0	41.4	17.1	25.1	21.3	17.1
6	1.488	72.3	62.5	42.4	9.4	7.3	50.4	22.8	46.1	20.6	33.3	19.8	17.7
7	1.526	145.5	93.7	96.8	63.1	60.8	55.7	45.0	57.2	41.2	42.7	42.5	41.9
8	1.538	118.6	82.0	67.0	35.8	38.6	47.7	30.4	41.7	28.5	46.0	30.7	27.2
15	1.540	243.3	203.4	132.1	41.2	52.5	155.7	58.7	139.8	62.5	102.7	65.2	54.8
12	1.802	129.5	102.3	58.4	32.3	35.2	67.5	28.5	64.3	33.7	33.8	33.6	28.5
5	2.040	123.2	91.6	61.7	31.0	35.6	60.7	28.0	60.6	26.6	47.6	34.7	25.4
11	3.056	102.6	75.9	67.8	63.6	62.1	65.9	53.5	65.1	50.5	72.6	52.4	48.1
17	3.084	172.1	158.7	85.5	36.6	42.0	105.0	44.4	94.9	44.9	54.9	46.0	46.5
9	6.828	224.9	166.9	109.5	74.8	81.0	139.1	85.1	128.4	80.4	106.8	88.2	79.1

how well a uniform regression model holds throughout the small areas. The post-stratified estimator based on the regression becomes better than the one based on the counts as the correlation increases. This conclusion is also true for all other estimators. The expansion estimator is the worst estimator since its Root Mean Squared Error is higher than the one associated with the remaining estimators.

This third layer of analysis is given in the graphs that are displayed in Appendix A. The behaviour of the expansion (EXP), ratio-synthetic (SYN1/R), modified regression ratio (MSAR1/R), Battese-Fuller (BF1/R), post-stratified (POS1/R) and positive weight (SH1/R) estimators is investigated from the point of view of conditional bias and mean squared error within each of selected Census divisions by industrial group cross-classifications. These cross-classifications display a mixture of bias (positive, negative, negligible) and correlation (low, medium, high). The summary statistics of each sample size realization is averaged over the number of Monte-Carlo samples that have achieved it. Several points about the behaviour of the various estimators may be noted from these graphs.

From the relative conditional bias graphs, it can be seen that the following holds. The conditional bias of the ratio-synthetic estimator (SYN1/R) is fairly constant throughout the sample size take range. The modified regression-ratio (MSAR1/R) is biased in the direction of the ratio-synthetic below the expected sample take and nearly unbiased above. The behaviour of the positive weight estimator (SH1/R) is very similar to that of the modified regression-ratio estimator. The nested error model estimator (BF1/R) tends to be biased in the direction of the ratio-synthetic if the bias of the latter estimator is considerable. Its bias, for any given sample size take, is higher than the bias associated with the other mixture estimators (MSAR1/R and SH1/R). For all the mixture estimators concerned, their bias seems to decrease as the correlation increases. The post-stratified (POS1/R) estimator is nearly unbiased throughout the whole sample size taken range. The expansion estimator is as expected from theory, biased negatively below the expected sample size take and positively above.

Turning to the conditional Root Mean Squared Error (R.M.S.E.) graphs, it is

evident that the expansion estimator is the worst estimator due to its high R.M.S.E. throughout the sample size take range. Its behaviour is quadratic, with the R.M.S.E. growing bigger as the sample size take departs from the expected size take. In order to summarize all the estimators concerned in terms of their general behaviour throughout the provided graphs, Table 11 ranks them from low R.M.S.E. to high R.M.S.E. taking into account, industry, bias and correlation.

TABLE 11: Summary Behaviour of the Estimators Based on Selected Industrial and Census Division Classifications

Industrial Group	Census Division	BIAS	Correlation	RANKING					
				1	2	3	4	5	6
RETAIL	2	< 0		BF1/R	SYN1/R	MSAR1/R	SH1/R	POS1/R	EXP
	3	≃ 0	LOW	SYN1/R	BF1/R	SH1/R	MSAR1/R	POS1/R	EXP
	8	> 0		BF1/R	MSAR1/R	SH1/R	SYN1/R	POS1/R	EXP
CONSTRUCTION	6	< 0	MEDIUM	POS1/R	SH1/R	MSAR1/R	BF1/R	SYN1/R	EXP
	7	≃ 0		SYN1/R	BF1/R	MSAR1/R	SH1/R	POS1/R	EXP
ACCOMMODATION	9	> 0	HIGH	SYN1/R	BF1/R	MSAR1/R	SH1/R	POS1/R	EXP

As evidenced from this table, the dominant estimator in terms of R.M.S.E. is the nested error model (BF1/R): it loses to the synthetic ratio estimator in those instances where the bias for specific small area and industry is negligible. The effect of higher correlation is to group the mixture

estimators tighter in terms of R.M.S.E. Turning back to the root conditional M.S.E. graphs, the ranking given in Table 11 holds throughout most of the range. The nested error model seems to lose power near the left-hand tail of the graphs. As sample size taken increases, the R.M.S.E. of the mixture estimators comes closer: this is to be expected because the post-stratified portion of these estimators becomes more dominant as sample size taken increases. The post-stratified estimator R.M.S.E. behaves basically as a decreasing monotonic function as sample size taken increases. This estimator, for most of the graphs, does not dominate the mixture and synthetic estimators.

6.0 Conclusions

The estimation of small area business statistics is not a straightforward matter. Within the limits of the current investigation, several concerns were addressed. These included the compatibility of administrative files compiled by different organizations; the modelling of data using regression techniques in order to apply synthetic or "pseudo-synthetic" procedures for small area estimation; and the investigation of the properties of several small area estimators.

The use of administrative files such as Comscreen, which contain counts or auxiliary information on a 100% basis, is important because they form the basis of benchmark variables required for synthetic estimation. The comparability of such files to these (such as the Combined.Master) which contain the variables of interest on a sample basis is therefore crucial in order to use the auxiliary information available on a 100% basis. With respect to the current application of these files for small area estimation, the estimation of Wages and Salaries, the data common to both files was fairly comparable. The discrepancy of the Industrial Coding between the two files can be synthetized by controlling on the coding of one of them. The strength of the regression relationships between Wages and Salaries and other variables varied according mainly to Industrial Code.

In terms of the available estimators for small area estimation, the following

conclusions can be made. The best estimators in terms of low bias and root mean squared error are the mixture estimators. For relatively high sample take within a small area, the post-stratified estimator may be superior to these mixture estimators. Amongst the mixture estimators, the estimators provided by Hidiroglou and Sarndal (1984), and Srinath and Hidiroglou (1985) offer a good compromise for low bias and root mean squared error. The nested error model estimator suggested by Battese and Fuller (1984) has consistently achieved the smallest root mean squared error at the expense of bias. Pure synthetic estimators may be very biased for some small areas. The expansion estimator is conditionally badly biased above and below the expected sample take within a small area, and it also displays the highest root mean squared error. The use of this estimator using the weights at a high level of aggregation is not recommended. The mixture estimators show the most gains in terms of reliability (root mean squared error) if the auxiliary information is well correlated with the variable of interest. Time series procedures can strengthen the small area estimation if the correlation across time between the same variables is greater than the correlation of these variables with other variables within a given time period.

ACKNOWLEDGEMENTS

The author would like to thank Gerry Horner and Conrad Bordeleau for their work on linking and gecoding the required administrative files; John Hunton for patiently producing the computer graphs; Nicky Smalldridge for typing the manuscript; E.E. Strauss, Chief and A.V. Winkworth, Director of Business Survey Methods Division for supporting this research.

References

Bankier, M.B. (1983), "Estimators Based on Several Stratified Samples with Applications to Multiple Frame Surveys", Proceedings of the Section on Survey Research Methods of the American Statistical Association Meeting, Toronto, Ontario.

Battese, G.E. and Fuller, W.A. (1984). An Error Components Model for Prediction of County Crop Areas Using Survey and Satellite Data. Survey Section, Statistical Laboratory, Iowa State University, Ames.

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.

Darcovitch, N. (1982). Current Procedures and Proposals for the Acquisition and Usage of Income Tax Data. Statistics Canada Report.

Drew, J.D., Singh, M.P. and Choudhry, G.H. (1982). Evaluation of Small Area Estimation Techniques for the Canadian Labour Force Survey. Survey Methodology, 8, 17-47.

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.

Fuller, W.A. and Harter, R. (1985). The Multivariate Components of Variance Model for Small Area Estimation. Paper presented at the Small Area Symposium, May 22-24, Ottawa, Canada.

Gigantes, T. (1983). Small Area Business Data - A Proposal. Statistics Canada Report.

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

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, 251-161.

Hidiroglou, M.A. and Srinath, K.P. (1981). Some estimators of a population total from simple random samples containing large units. Journal of the American Statistical Association, 76, 690-696.

Hidiroglou, M.A., Morry, M. and Vaillancourt, C. (1984). "Comparison of Elements between the Combined.Master and Comscreen files - Tax Year 1981". Statistics Canada.

Hidiroglou, M.A. (1984). "Some Characteristics of SIC coding between Comscreen and the Combined.Master". Statistics Canada Report.

Hidiroglou, M.A. (1984). "Exploratory Analyses performed on the Combined.Master file". Statistics Canada Report.

Hidiroglou, M.A. and Sarndal, C.E. (1984). Experiments with Modified Regression Estimators for Small Domains. Statistics Canada Report.

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

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.

Rao, J.N.K. (1985). Conditional Inferences in Survey Sampling, Statistics Canada Report.

Royall, R.M. (1978). Prediction Models in Small Area Estimation. In Synthetic Estimates for Small Areas (C. Steinberg, ed.), pp 63-87. National Institute on Drug Abuse, Research Monograph 24. U.S. Government Printing Office, Washington, D.C.

Sande, I.G. (1978). Analysis of 74 and 75 Tax Record Access Annual Files. Statistics Canada Report.

Sande, I.G. (1984). Small Area Business Data Development Project. Paper presented at the Bureau of the Census - Statistics Canada Interchange, Washington, D.C.

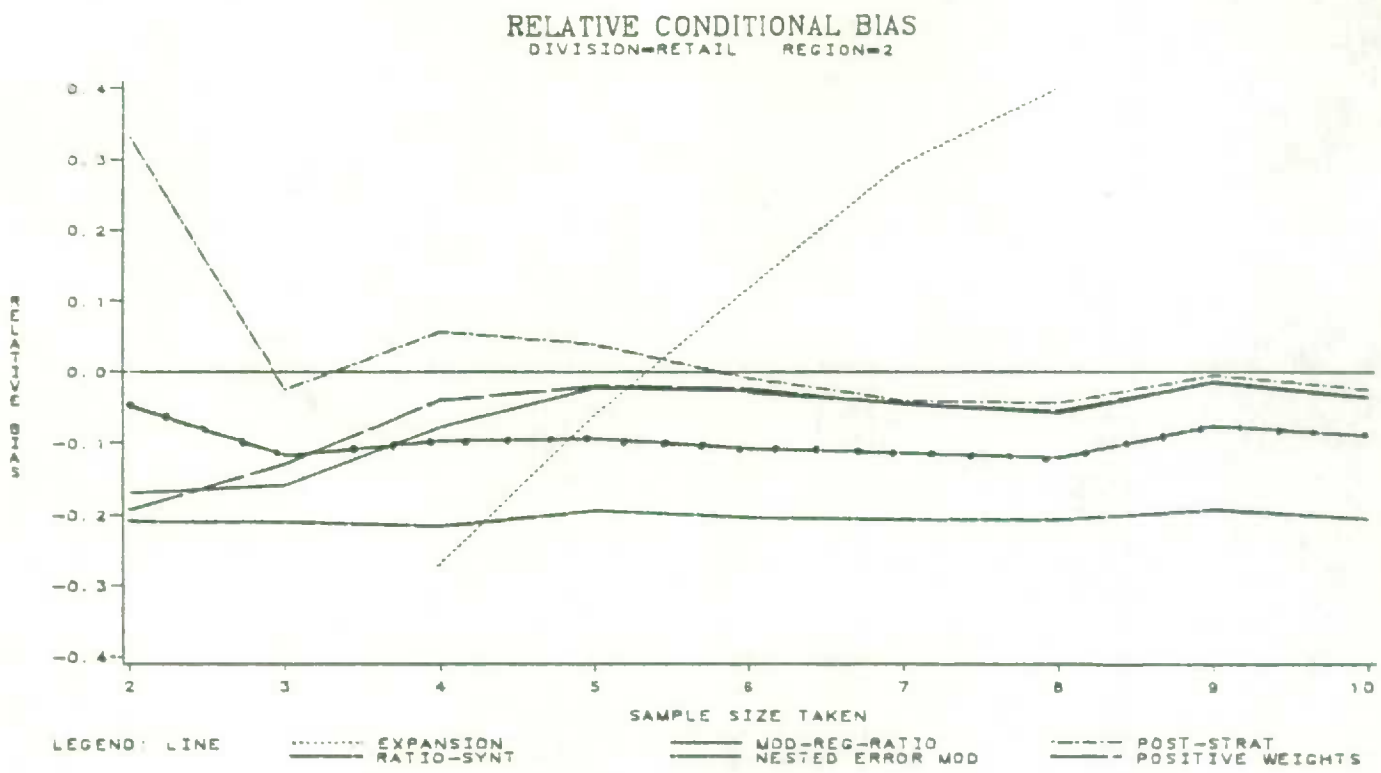
Sarndal, C.E. (1984). Design-consistent versus model-dependent estimators for small domains. Journal of the American Statistical Association, 79, 624-631.

Schaibble, W.L. (1979). A Composite estimator for small area statistics. In Synthetic Estimates for Small Areas (C. Steinberg, ed.), pp 36-83. National Institute on Drug Abuse, Research Monograph 24, U.S. Government Printing Office, Washington, D.C.

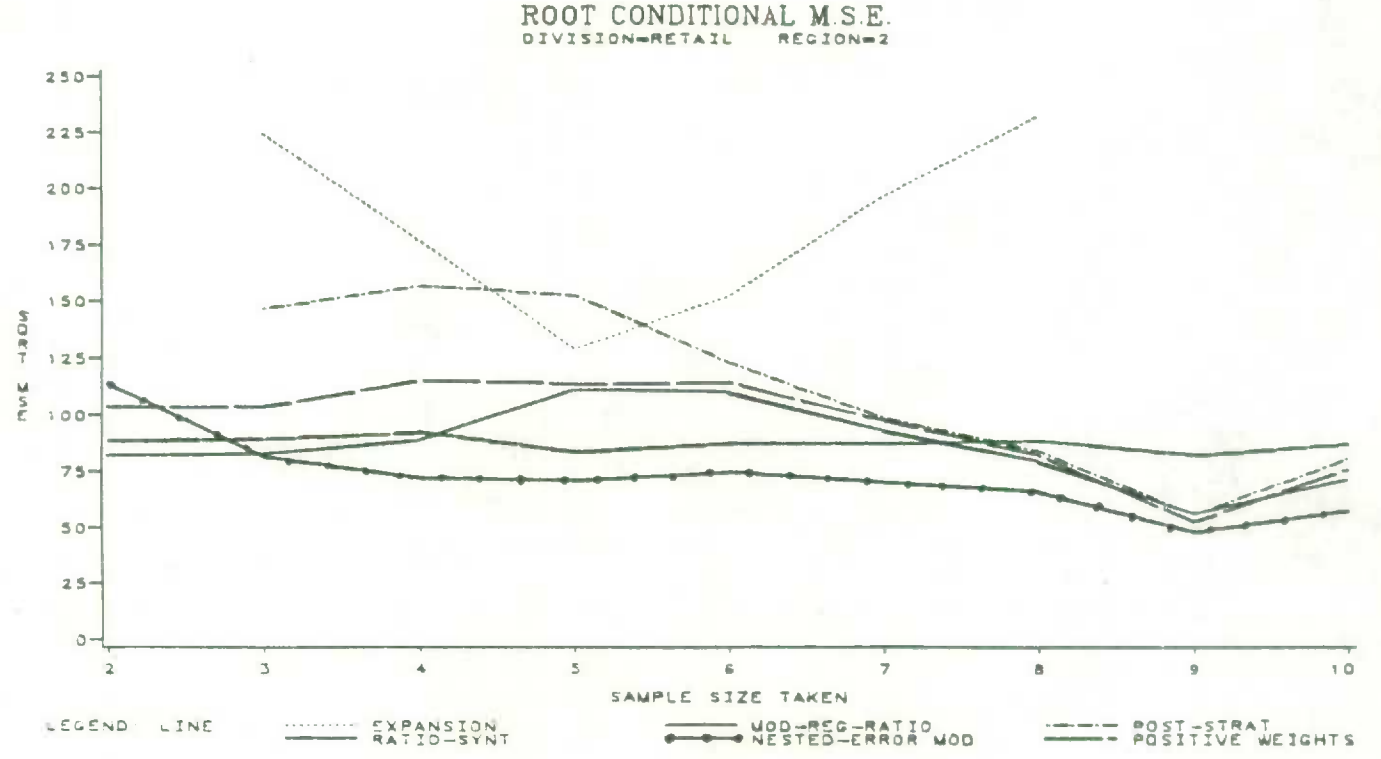
Srinath, K.P. and Hidiroglou, M.A. (1985). Sample Size-Dependent Estimators for Small Areas with Applications to Business Data. Paper presented at the Small Area Symposium, May 22-24, Ottawa, Canada.

APPENDIX A

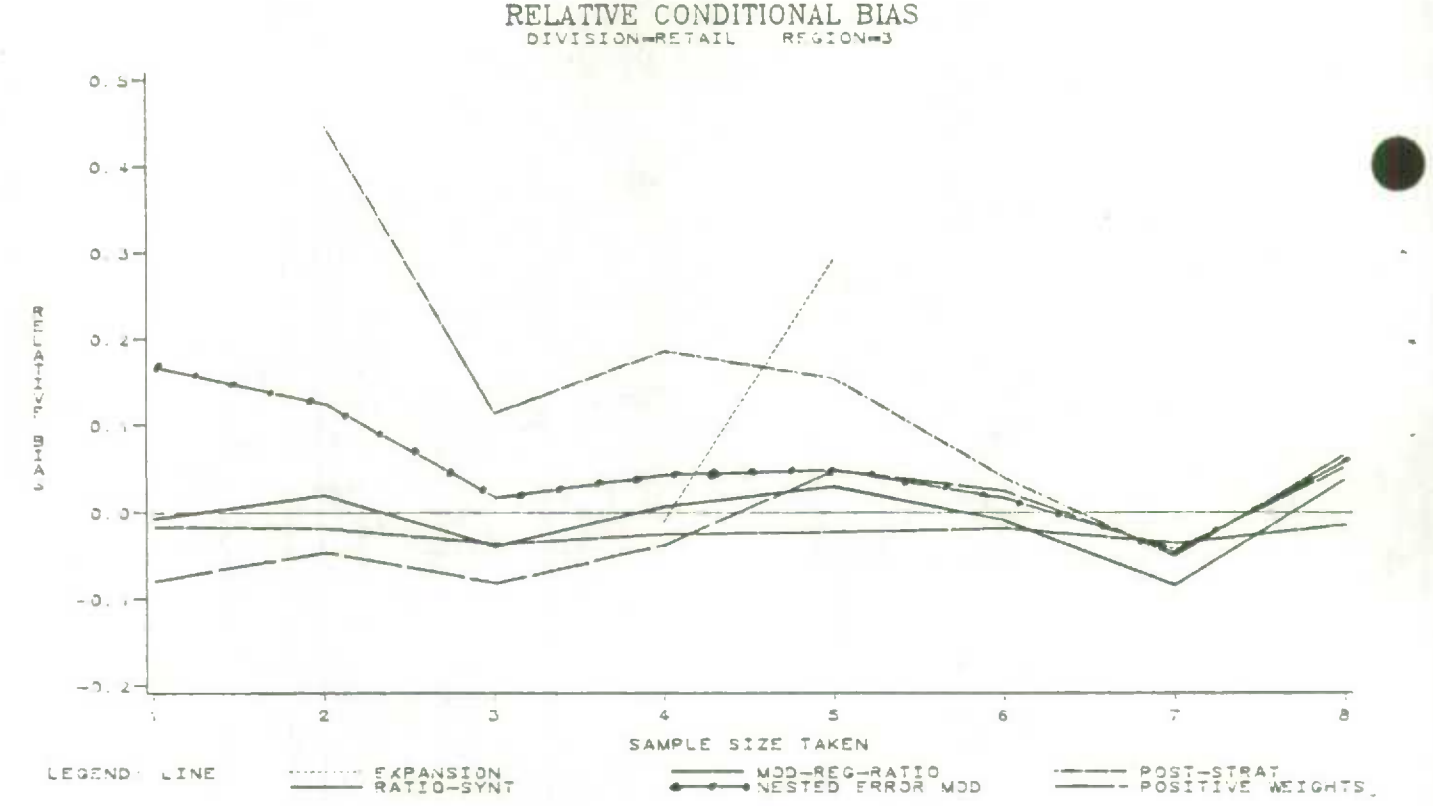
Selected Relative Conditional Bias and Root Conditional Mean Squared Error Graphs



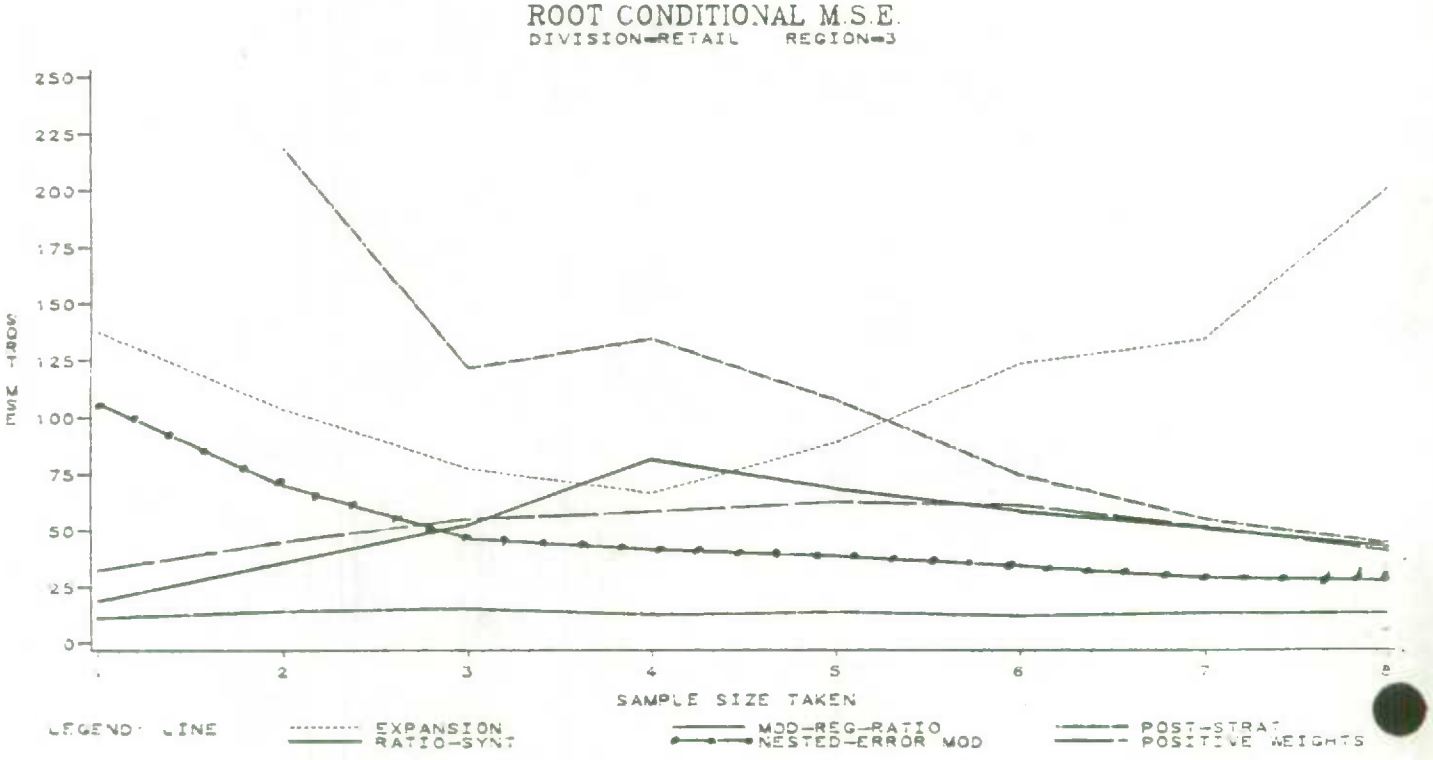
RELATIVE BIAS FOR REGRESSION ESTIMATORS



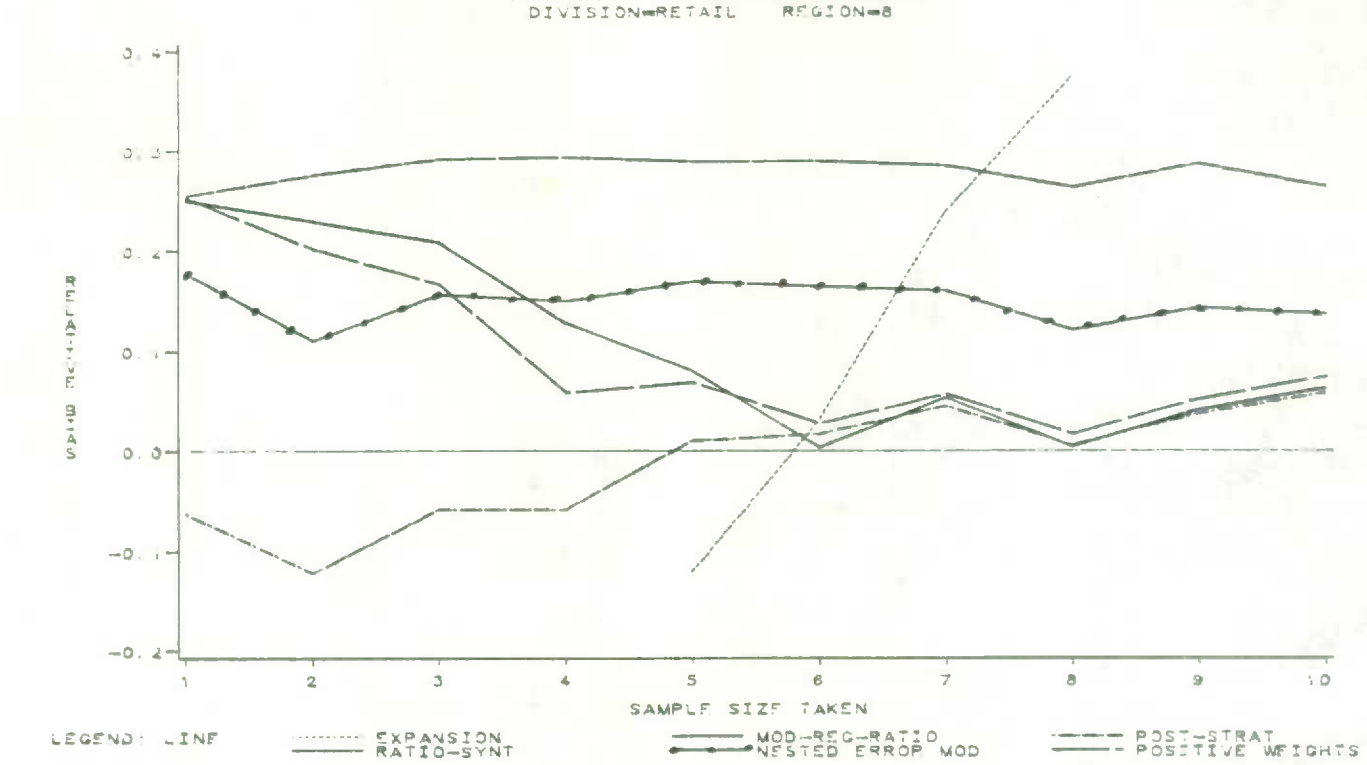
SQUARE ROOT MSE FOR REGRESSION ESTIMATORS



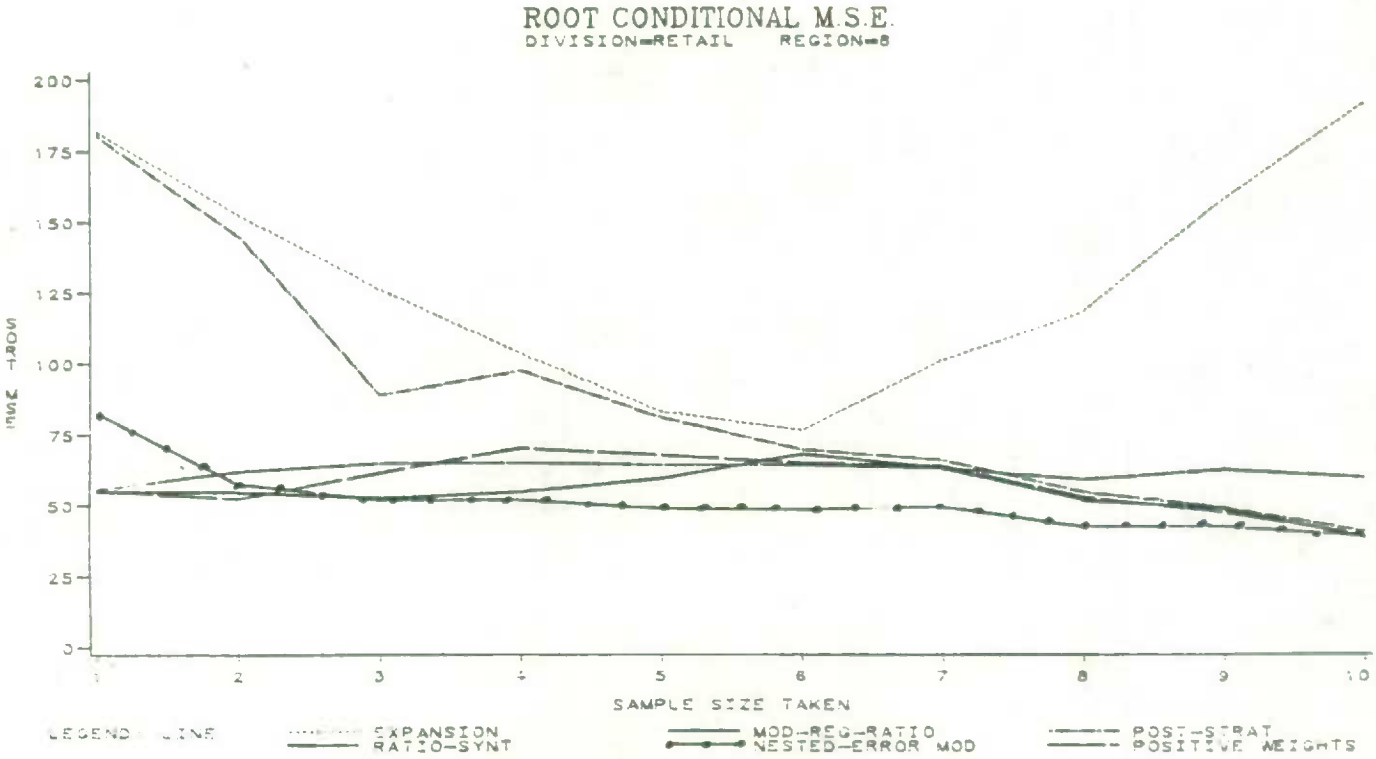
RELATIVE BIAS FOR REGRESSION ESTIMATORS



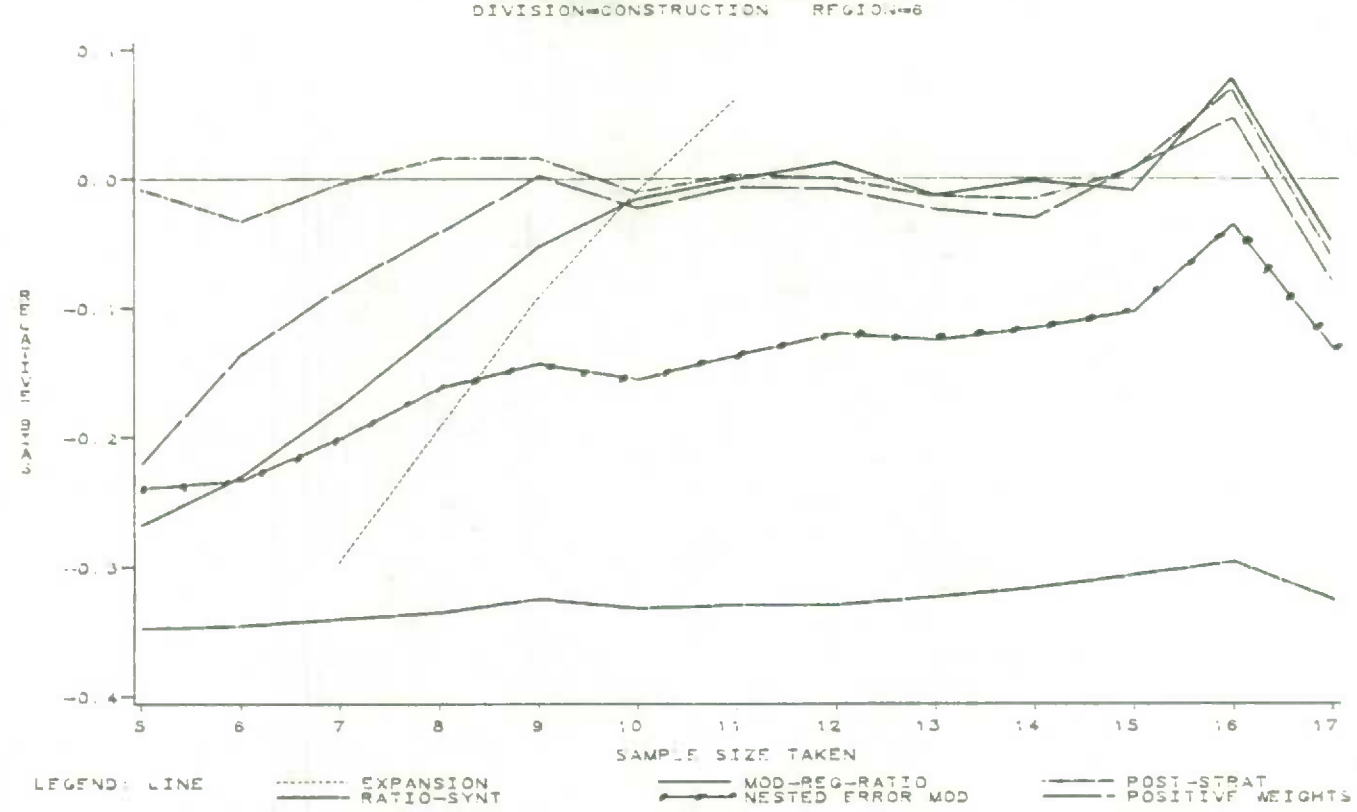
SQUARE ROOT MSE FOR REGRESSION ESTIMATORS



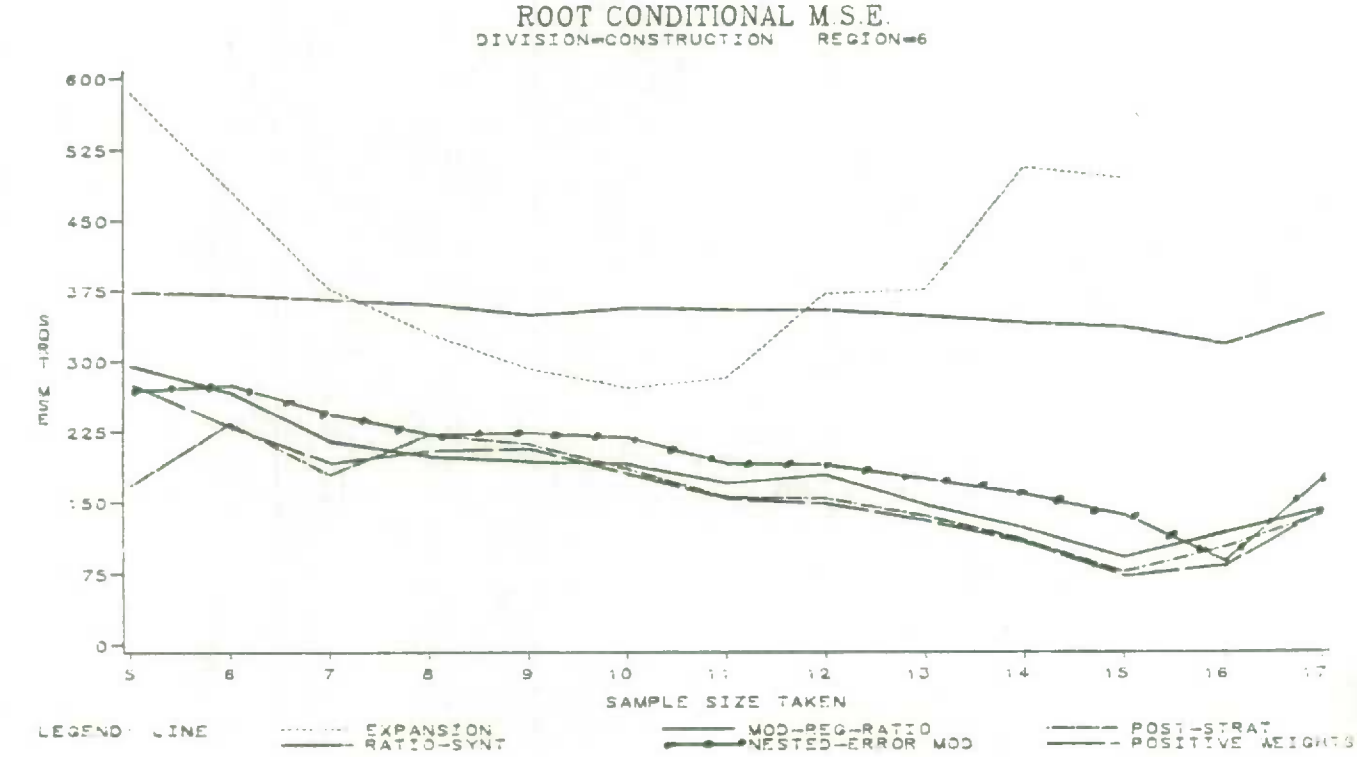
RELATIVE BIAS FOR REGRESSION ESTIMATORS



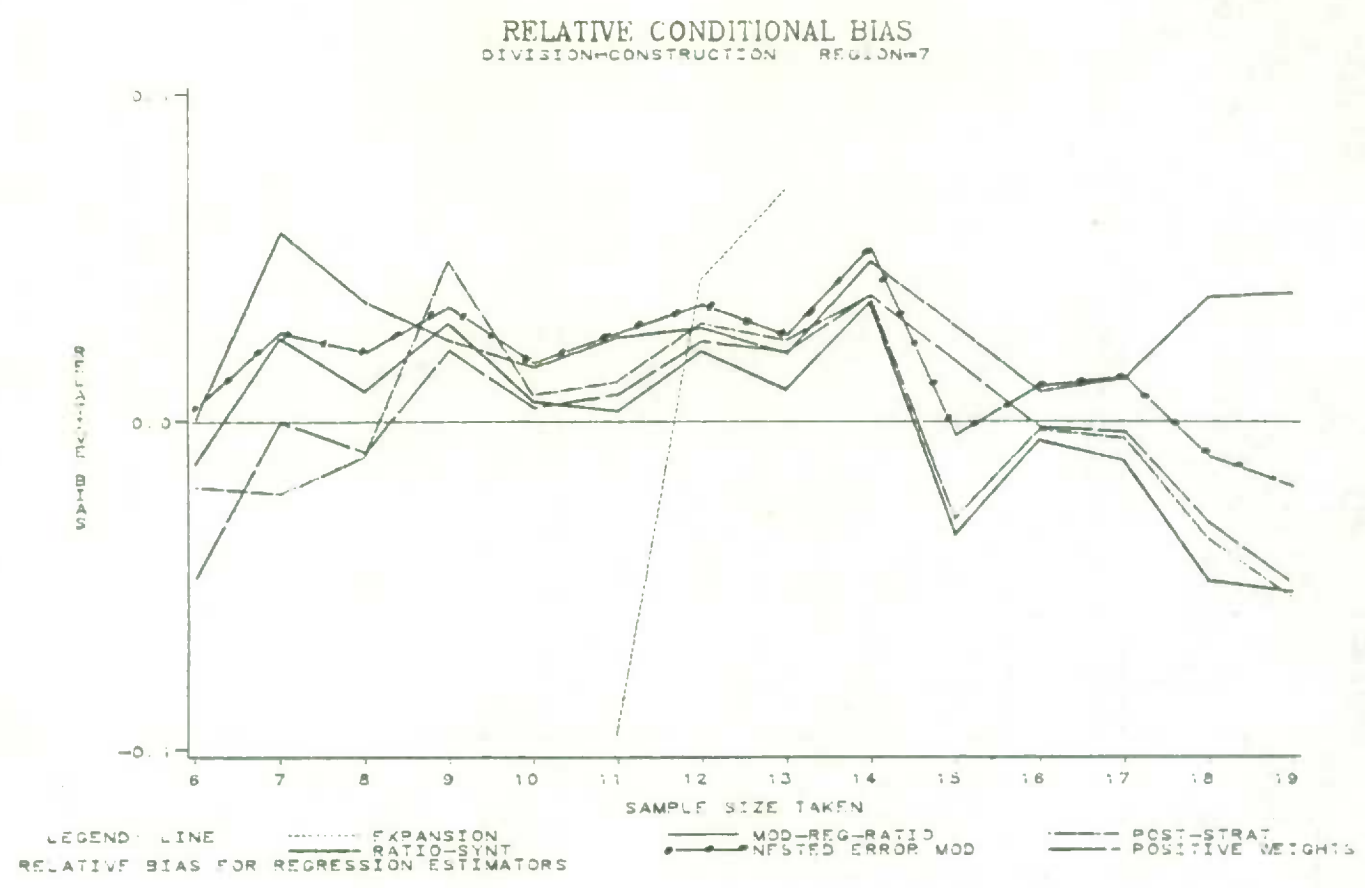
SQUARE ROOT MSE FOR REGRESSION ESTIMATORS



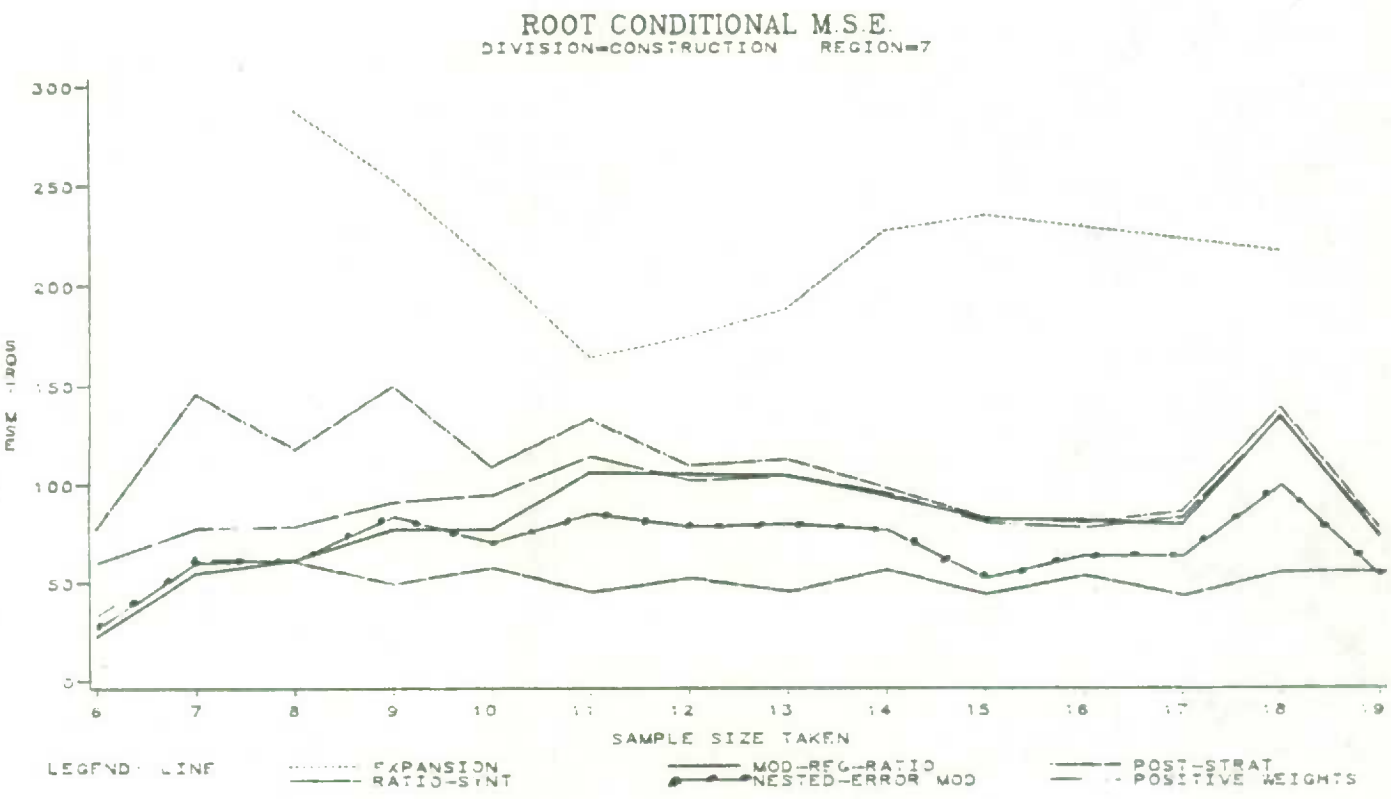
RELATIVE BIAS FOR REGRESSION ESTIMATORS



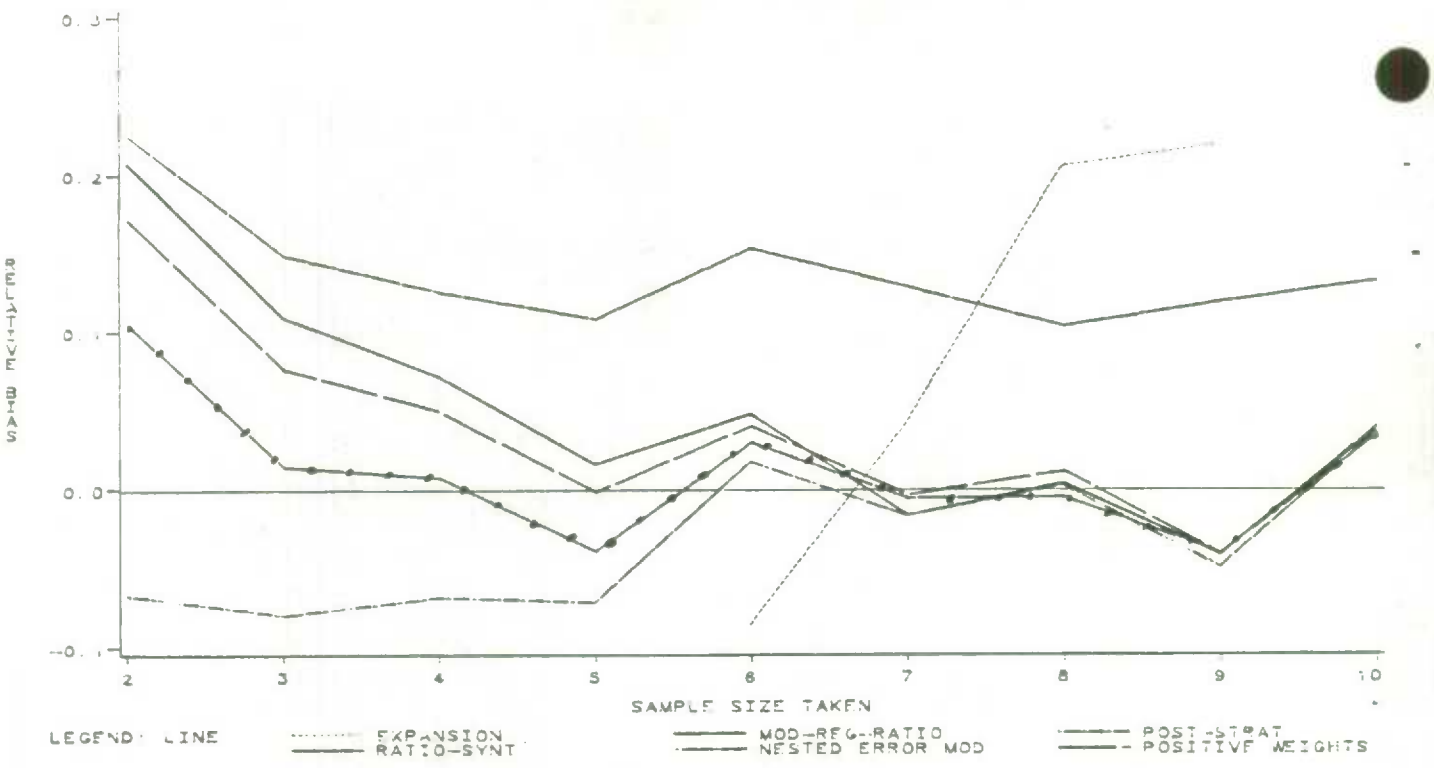
SQUARE ROOT MSE FOR REGRESSION ESTIMATORS



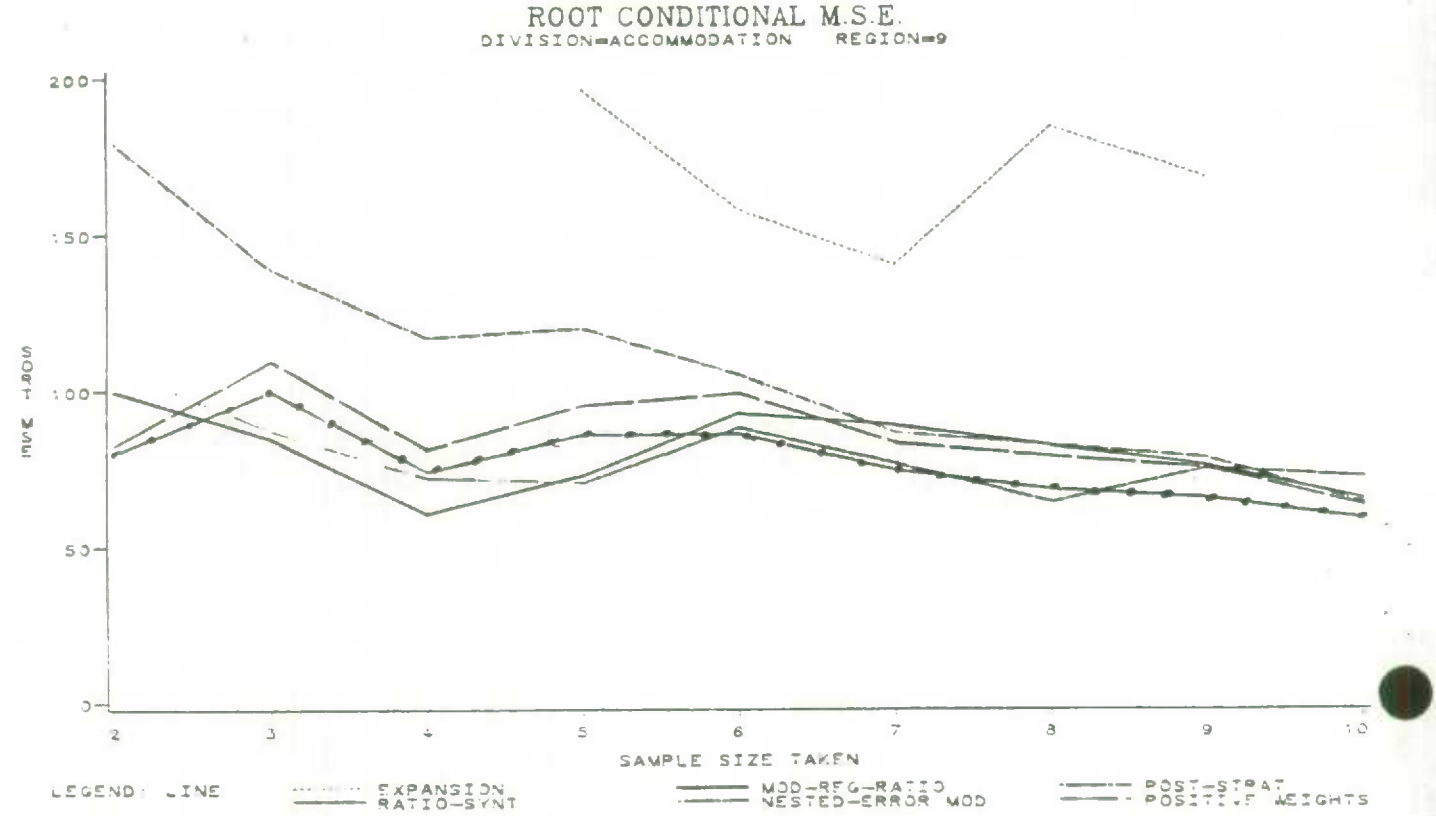
RELATIVE BIAS FOR REGRESSION ESTIMATORS



SQUARE ROOT MSE FOR REGRESSION ESTIMATORS



RELATIVE BIAS FOR REGRESSION ESTIMATORS



SQUARE ROOT MSE FOR REGRESSION ESTIMATORS

APPENDIX B[1]

The Nested-Error Model for Small Areas.

For the data that was dealt with, Wages and Salaries and Gross Business Income, the following simple linear nested error model was used.

$$X_{ai}^{-\frac{1}{2}} Y_{ai} = \beta X_{ai}^{\frac{1}{2}} + U_{ai}$$

where $U_{ai} = v_a + e_{ai}$

$v_a \sim$ NID $(0, \sigma_{vv})$

$e_{ai} \sim$ NID $(0, \sigma_{ee})$

Y_{ai} = Wages and Salaries for the i-th element within the a-th small area,

X_{ai} = Gross Business Income for the i-th element within the a-th small area.

$a = 1, \ldots, A$; $i = 1, \ldots, n_a$; (sample)

$i = 1, \ldots, N_a$; (population)

The estimator for the mean of the a-th area is

$$\bar{x}_a^{(1)} = \hat{\beta}\, \bar{X}_{a(p)} + \bar{W}_{a(p)}\, \hat{d}_a\, \overset{\Delta}{U}_{a.},$$

where for the a-th area;

$\bar{X}_{a(p)}$ is the population mean of the x variable,

$\bar{W}_{a(p)}$ is the population mean square roots of X_{ai},

---

[1] The author is thankful to Professor Wayne A. Fuller of Iowa State University for providing this material.

$\overset{\Delta}{U}_{a.}$ is the mean of the residuals in the transformed variables

$\hat{d}_a$ is the dampening factor.

The computations required for $\hat{d}_a$ are next given. For details on the derivation used, see Fuller and Harter (1985).

$$\hat{d}_a = 1 - \hat{H}_a;$$

where $$\hat{H}_a = \min (1, \overset{''}{H}_a),$$

$$\overset{''}{H}_a = (d_e + 2)^{-1} d_e [(1 + \alpha_a \hat{\theta}_a)^{-1} \hat{\theta}_a + 2 \alpha_a (1 + \alpha_a \hat{\theta}_a)^{-3} \hat{\theta}_a^2 (d_e^{-1} + d_v^{-1})],$$

$$\hat{\theta}_a = \min (\overset{''}{\theta}_a, 1),$$

$$\overset{''}{\theta}_a = d_v^{-1} (d_v - 2) \hat{M}_{..}^{-1} n_a^{-1} \hat{\sigma}_{ee},$$

$$\alpha_a = (1 - n_a n_*^{-1})$$

$$d_v = 2 \hat{M}^2 / v(\hat{M}_{..}),$$

d_e = degrees of freedom for $\hat{\sigma}_{ee}$,

$$v(\hat{M}_{..}) = 2 n_*^{-2} [A (A - 1)]^{-1} \sum_{a=1}^{A} n_a^2 (\hat{\sigma}_{vv} + n_a^{-1} \hat{\sigma}_{ee})^2,$$

$$\hat{M}_{..} = n_*^{-1} MSA,$$

$$MSA = [\sum_a n_a (\overset{\Delta}{U}_{a.} - \overset{\Delta}{U}_{..})^2] / (A - 1),$$

$$\overset{\Delta}{U}_{..} = \sum_{a=1}^{A} \sum_{i=1}^{n_a} \hat{U}_{ai} / n,$$

$$\hat{U}_{ai} = X_{ai}^{-\frac{1}{2}} Y_{ai} - \hat{\beta} X_{ai}^{\frac{1}{2}},$$

$$\hat{\sigma}_{ee} = [\sum_a (n_a - 1) - 1]^{-1} \sum_a \sum_i (\hat{U}_{ai} - \overset{\Delta}{U}_{a.})^2,$$

$$\hat{\sigma}_{vv} = [MSA - \hat{\sigma}_{ee}]\, n_*^{-1}$$

$$n_* = (A - 1)^{-1} [\sum_a n_a - (\sum_a n_a)^{-1} \sum_a n_a^2].$$

The estimator for the mean of the a-th area taking into account the finite size of the population is:

$$\overline{x}_a^{(2)} = f_a \overline{Y}_{a(s)} + (1 - f_a) [\hat{\beta} \overline{X}_{a(p-s)} + \hat{d}_a \overline{W}_{a(p)} \overset{\Delta}{U}_{a.}]$$

where $f_a = n_a / N_a$, $\overline{Y}_{a(s)}$ is the sample mean of the Y's and $\overline{X}_{a(p-s)}$ is the mean of the units not in the sample for area a.

Ca 005

STATISTICS CANADA LIBRARY
BIBLIOTHEQUE STATISTIQUE CANADA

1010148859

DATE DUE