Statistics Canada / Statistique Canada

Methodology Branch

Time Series Research and Analysis Division

Direction de la méthodologie

Division de la recherche et de l'analyse des chroniques

Canada

WORKING PAPER TSRA- 85-042

TIME SERIES RESEARCH & ANALYSIS DIVISION

METHODOLOGY BRANCH

CAHIER DE TRAVAIL TSRA-85-042

DIVISION DE RECHERCHE ET
ANALYSE DES SERIES
CHRONOLOGIQUES
DIRECTION DE LA
METHODOLOGIE

SENSITIVITY OF SMALL AREA ESTIMATORS TO MISCLASSIFICATION AND CONCEPTUAL DIFFERENCES OF VARIABLES

by

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

This is a preliminary version. Do not quote without authors' permission.

Comments are welcome.

SENSITIVITY OF SMALL AREA ESTIMATORS*

TO MISCLASSIFICATION AND CONCEPTUAL DIFFERENCES OF VARIABLES

by

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

*Invited paper presented at the International Symposium of Small Area Statistics, Ottawa, May 22-24, 1985.

1.0. Introduction

In recent years there has been an increasing demand for statistical information at subprovincial levels. Most of the surveys at Statistics Canada were designed to produce reliable estimates at the province level but not at a lower level of disaggregation. To answer the need for small area statistics without increasing the cost and response burden associated with the larger sample size necessary for producing such statistics through a survey, attention was directed towards the possible use of administrative records. This paper discusses the problems encountered with the reconciliation of data from two administrative files at Statistics Canada and Revenue Canada considered as possible sources for producing small area statistics, related to Business Income and Wages and Salaries. It also investigates the impact of conceptual differences and discrepancies in industrial classification between the two files, on a set of small area estimators. This latter analysis is complemented by a simulation study carried out on unincorporated business data in the Accommodation industry from Nova Scotia and Ontario to produce Wages and Salaries estimates at the Census Division level.

In 1971, Statistics Canada (STC) was given access to Income Tax data for the purpose of statistical analysis through the Statistics Act. Tax data for unincorporated businesses (T1) (the scope of this present study) have been transcribed by Statistics Canada since 1973 on a sample basis to produce a file known as the COMBINED-MASTER. It contains a 25% sample of the unincorporated universe with Gross Business Income (GBI) between $25,000 and $500,000 and a 100% sample for tax filers with Gross Business Income over $500,000.

The second administrative file considered in this study is created at Revenue Canada by transcribing certain tax items for all unincorporated businesses with income over $25,000. The original purpose of this file called COMSCREEN was to serve as a tool in Revenue Canada's auditing procedure. The two files contain a number of economic variables which are comparable in concept. These are Sales (known as GBI at STC), Capital Cost Allowance, (Depreciation at STC), Net Profit, Filer's share of Net Profit for filers that are involved in partnership. The COMBINED-MASTER has a number of additional economic variables which are not transcribed on the COMSCREEN file. These are Wages and Salaries, (W&S) Inventories, and Assets. Thus one file (COMSCREEN) is more complete in coverage but contains less information, while the second file (COMBINED-MASTER) has more variables of interest but only on a sample basis.

In a previous study (1984), the authors investigated the possibility of producing Wages and Salaries statistics for small business in small areas using the two files. They found that a strong linear relationship existed between Wages and Salaries (available on the COMBINED-MASTER on a sample basis) and Gross Business Income (Sales) (which was present on COMSCREEN for all T1 tax filers) at the major division industrial breakdown by province. It was necessary to introduce a square root of GBI transformation to the data to make the residuals of the regression homoscedastic. This transformation led to a ratio-type estimation. Applying the ratio obtained from the sample file to Sales on the universe file produced improved estimates of Wages and Salaries at the small area level when compared to the ones obtained by simply blowing up the sample at the small area level. This estimation technique is otherwise known as the ratio-synthetic estimation (SYN/R).

Other small area estimators included in that study were the post-stratified estimator (POST), the count-synthetic estimator (SYN/C) popularized by Gonzales (1973) and two estimators developed by Särndal (1981, 1983) to correct for the bias introduced through synthetic estimation called regression-count (REG/C) and regression-ratio (REG/R).(1) A simulation study carried out on small business data from Nova Scotia indicated that in terms of mean square error (MSE) and especially in small domains, the ratio-synthetic estimator was the most efficient followed by the regression ratio. In terms of bias of the estimates, the regression-ratio estimator performed best, i.e. it produced estimates with minimum bias.

After the completion of the simulation and the analysis of the results, there were several areas that needed further investigation, such as:

a) Is it possible to cut down on the bias introduced by the ratio-synthetic estimator without considerably sacrificing efficiency as was the case with the regression-ratio estimator?

b) Does the ranking of the estimator change when moving from a smaller province (Nova Scotia) to a larger province (Ontario)?

Although the previous study pointed out that there were discrepancies in the industrial classification and in the concepts of variables present on the two files, these discrepancies were ignored in the estimation procedure. The simulation and the analysis on the performance of the estimators assumed that the Revenue Canada SIC codes were identical to those on the Statistics Canada file and that the 'Sales' entry of each

---

(1) The formulas for these estimators are given in Appendix A.

record on COMSCREEN coincided with the Gross Business Income entry of the corresponding record on the COMBINED-MASTER, i.e. that the concept of Sales and Gross Business Income is equivalent. In light of the discrepancies the following further questions arise:

(c) What is the difference in Wages and Salaries when tabulating by RC SIC code versus STC SIC code?

(d) How much extra error is introduced in the estimation procedure by using Revenue Canada SIC codes and concepts?

(e) Does the ranking of the estimators change due to the added discrepancy in coding and concepts?

The objective of the present study is to answer these five questions.

Section 2 introduces two new estimators and compares their performance to other five estimators based on efficiency, bias and coefficient of variation measures using a simulation on data from the Nova Scotia Accommodation industry. Section 3 evaluates the impact of the size of the province on the ranking of estimators by comparing simulation results from Ontario to those from Nova Scotia.

Wages and Salaries are tabulated according to both Statistics Canada and Revenue Canada SIC codes in Section 4 to assess the difference resulting from misclassification. Section 5 presents the results from a simulation on Nova Scotia and Ontario data using Revenue Canada classification and concepts. The estimates are compared to those obtained earlier in a similar simulation based on Statistics Canada SIC codes and concepts. This section also examines the impact of coding discrepancies and conceptual differences on the ranking of the estimators used. The conclusions of the study are given in Section 6.

2.0 An evaluation of small area estimators used to estimate Wages and Salaries for unincorporated businesses.

In order to study the properties of various small area estimators, a simulation was undertaken by the authors in 1984. The simulation mimicked the use of administrative data arising from several sources and their subsequent combination to yield small area estimates. Since the Statistics Canada administrative file had all the required information, it was used as the file for drawing the samples required for the simulation. Five hundred samples of size 429 were selected from the target population of 1,678 unincorporated businesses in Nova Scotia. The small areas of interest were major industrial groupings by Census Division.

The estimators used (direct (DIR), post-stratified (POST), count-synthetic (SYN/C), ratio-synthetic (SYN/R), regression-count (REG/C) and regression-ratio (REG/R)) were evaluated in terms of several criteria; i.e. a) relative percentage efficiency; b) relative percentage bias; c) coefficient of root mean square error, etc. The estimator that produced the lowest root mean square error (RMSE) of the estimates was the ratio-synthetic estimator. However, it suffered from the drawback of introducing a relatively large bias. Sarndal's regression-ratio estimator which came second and which was nearly unbiased, on the other hand, generated estimates with higher MSE. This present study will consider two new versions of the regression ratio estimator designed to improve the RMSE measure without significantly deteriorating the bias.

The first of the two modified regression-ratio estimators (MREG/R(1)) developed by Hidiroglou and Särndal (1984) (see estimator $t_7(ai)$ in Appendix A) gives gradually less weight to the residual correction term as

the realized sample take deviates from the expected sample take. This may introduce a small bias in exchange for a reduced variance contribution when the realized sample take is lower than expected.

Another alternative to the regression ratio estimator was based on the rationale that given the small domains in question, if the realized sample take is lower than expected, the resulting sample size is so small that the correction term is not reliable at all. Thus, the correction should not be used and the estimator defaults to the ratio-synthetic estimator. If the realized sample take is higher than expected, the correction term should enter with the corresponding inverse weight. For the formula see estimator $t_8(ai)$ in Appendix A.

The estimators MREG/R(1) and MREG/R(2) were used together with the other six estimators from the previous study to produce small area Wages and Salaries estimates in the Nova Scotia Accommodation industry through a 250 sample Monte Carlo simulation exercise. The universe in the simulation consisted of the 86 businesses found in this industry on the COMBINED-MASTER file.

The formulas for the three measures used in the calculation of the estimators, i.e. relative efficiency (RE) relative bias (RB) and coefficient of root mean square error (CRMSE) are given in Appendix B. Table 2.1 summarizes the results of the simulation at the province level for the eight estimators according to the three measures.

Table 2.1 Performance of Estimators in Nova Scotia Accommodation - using STC coding and concepts

	Estimators							
Measure	DIR	SYN/C	SYN/R	REG/C	REG/R	POST/C	MREG/R(1)	MREG/R(2)
RE	1.000	.369	.271	.725	.425	.809	.326	.302
RB	.045	.459	.361	.040	.045	.336	.195	.209
CRMSE	1.076	.533	.403	.846	.626	.747	.416	.415

The ranking of the original six estimators remained the same, i.e. SYN/R is still the best in terms of relative efficiency (RE) and RMSE but almost the worst regarding the bias (RE). The two new estimators strike a compromise in performance between SYN/R and REG/R. They improve upon REG/R in terms of RE and CRMSE but at the expense of deteriorating the bias, as expected. MREG/R(2) is more efficient than MREG/R(1), it falls short of the efficiency of SYN/R by only 11% as opposed MREG/R(2) that has an RE measure 20% higher than SYN/R. On the basis of bias, MREG/R(1) outperforms MREG/R(2). It cuts down on the bias introduced by SYN/R by 46% compared to 42% reduction by MREG/R(2). In terms of RMSE, the two new estimators rate the same, their measure is only 3% higher than the best performing SYN/R's as opposed to REG/R that was 55% worse than SYN/R, based on this measure.

In the final analysis, it can be said that the two new estimators introduced are a definite improvement over REG/R. They are only slightly worse than SYN/R regarding root mean square error and efficiency at the same time they substantially reduce the bias associated with SYN/R. In overall performance MREG/R(2) is preferable to MREG/R(1) to some extent. However, it is also somewhat more biased.

3.0 The effect of the size of the province on the performance of the estimators

In order to assess the impact of sample size on the optimality of estimators, a Monte Carlo simulation was carried out on data from Ontario. Because of the costs associated with working on data from such a large province, the simulation exercise was restricted to the Accommodation industry only. This yielded an 1874 record universe taken off the COMBINED-MASTER file from which two hundred and fifty 25% samples were drawn for the simulation study.

Table 3.1 presents the performance measure of the eight estimators at the province level according to the same three criteria that were applied to the Nova Scotia data.

Table 3.1 Performance of Estimators in Ontario Accommodation - using STC coding and concepts

	Estimators							
Measure	DIR	SYN/C	SYN/R	REG/C	REG/R	POST/C	MREG/R(1)	MREG/R(2)
RE	1.000	.841	.352	.770	.471	.810	.413	.394
RB	.031	.333	.169	.022	.015	.064	.051	.090
CRMSE	.571	.340	.174	.434	.265	.458	.206	.202

As expected, the size of the relative bias and root mean square error is smaller in Ontario than in Nova Scotia for each and every estimator.

In terms of efficiency the estimators do not improve on the performance of the direct estimator as much in Ontario as in Nova Scotia, but it has to be borne in mind that the percentage root mean square error of the estimates obtained from direct sample blow-up are much smaller in Ontario than in Nova Scotia. (.571 versus 1.076). These results are not

surprising given the difference in sample size between the two provinces. What is of more interest to us is whether the larger sample size affected the ranking of the estimators and in which direction?

Table 3.2 shows how the ranking of the estimators changes when moving from a large province to a small province.

Table 3.2 Comparison of ranking of estimators in Nova Scotia and Ontario

		Rank of Estimators							
Measures	Prov.	DIR	SYN/C	SYN/R	REG/C	REG/R	POST/C	MREG/R(1)	MREG/R(2)
RE	N.S.	8	4	1	6	5	7	3	2
	Ont.	8	7	1	5	4	6	3	2
RB	N.S.	2-3	8	7	1	2-3	6	4	5
	Ont.	3	8	7	2	1	5	4	6
CRMSE	N.S.	8	4	1	7	5	6	2-3	2-3
	Ont.	8	5	1	6	4	7	2-3	2-3

It is evident that the three top performing estimators; namely the ratio- synthetic and the two modified regression ratio estimators continued to stay at first, second and third place in terms of CRMSE and RE even in Ontario. The increase in the sample size did not effect the low ranking of these estimators according to the relative bias measure either. Similarly, the direct blow-up estimator remained in the last place when moving to a larger province. The most significant change occurred in the performance of the count synthetic estimator(SYN/C). While in the smaller province, the domain count combined with the average W&S was relatively adequate in describing the variation in Wages and Salaries among estimators, reflected by a rank 4 RE assigned to the SYN/C estimator. In

Ontario this type of estimation procedure only outperformed the direct sample blow up in terms of efficiency as shown by a seventh placing.

Only minor reshuffling of ranks took place in terms of bias among the first second and third rank estimators and also among the fourth, fifth and sixth rank ones. Based on root mean square error, the count synthetic and regression-ratio estimators switched places in ranking and so did the sixth and seventh ranking REG/C and POST/C estimators.

It can be concluded that although the performance of some of the estimators was affected by the size of the province, the ratio-synthetic estimator still proved to be the most successful followed by the two modified regression-count estimators.

4.0 Difference in Wages and Salaries resulting from discrepancy in industrial classification between STC and RC.

In the preceding analysis it was assumed that there were no differences in SIC coding and in concepts used by the two agencies. Before introducing this added source of error into the estimation procedure, it is worthwhile to examine the magnitude of discrepancy resulting from tabulating Wages and Salaries according to RC SIC codes instead of STC SIC codes.

For this purpose the COMBINED-MASTER file was matched against the COMSCREEN file using the tax filer's social insurance number as the matching key to create what will be referred to as the MATCHED file. Only single businesses were included to avoid mismatches between the two files. Whenever a match occurred, a new record was created containing all the information from the STC file as well as the industrial classification and the sales entry from the RC file. Information on this MATCHED file was then tabulated according to both STC and RC SIC codes. Table 4.1 indicates the number of businesses and the total wage bill paid out in the major division industrial grouping as coded by STC and RC.

The ratio of the Wages and Salaries total according to STC and RC classification is used as the measure of the discrepancy. In eight out of the seventeen industry groupings, total Wages & Salaries according to RC coding is less than according to STC coding. The ratio among these eight industries ranges from .622 for Real Estate to .988 for Transportation i.e. Revenue Canada figures fall short of STC figures by 37% in the worst case to 1% as the closest agreement. In the remaining industries, RC total wages exceed STC total wages anywhere from 4% (Business Service) to 212% (Finance and Insurance).

TABLE 4.1 CANADA - Discrepancy in Wages and Salaries due to coding differences at the major division SIC level

Major Division	No. of Units STC	No. of Units RC	STC W&S	RC W&S	W&S RC / W&S STC
1. Agriculture	170	237	3,031	3,270	1.079
2. Fishing	8	13	159	177	1.113
3. Logging & Forestry	685	630	19,007	17,521	.922
4. Mining	61	62	1,255	1,139	.907
5. Manufacturing	2,127	1,294	47,451	29,964	.631
6. Construction	8,495	8,041	136,607	131,420	.962
7. Transportation	2,806	3,033	40,317	39,811	.988
8. Communication	149	213	3,166	4,139	1.307
9. Wholesale	1,023	636	11,382	7,618	.669
10.Retail	12,603	13,492	158,835	179,881	1.132
11.Finance & insurance	7	14	48	149	3.122
12 Real Estate	169	108	2,676	1,666	.622
13.Business Service	337	344	6,900	7,153	1.037
15.Educational Service	68	106	1,609	1,913	1.189
16.Health & Social	254	329	6,947	8,257	1.189
17.Accommodation	4,039	3,936	84,957	83,105	.978
18.Other Services	3,588	3,822	61,003	63,780	1.046

Table 4.1 suggests that in certain industries such as Manufacturing, Communication, Wholesale, Finance and Insurance, Real Estate, Educational Service and Health and Social Service, some recoding of SIC on the Revenue Canada file needs to be carried out before it can be used in any tabultion or estimation procedure. These results basically agree with a previous analysis by the authors (1984) that measured the coding differences in terms of tabulated Gross Business Income.

5.0 Impact of misclassification and conceptual differences on the estimation of Wages and Salaries in small areas.

Most of the estimators discussed in this paper make use of the administrative information contained on the Revenue Canada file. The count-synthetic and regression-count estimators obtain the universe count of businesses per small area from the COMSCREEN while the ratio-synthetic and the regression-ratio type estimators use the universe total of Gross Business Income per small area as it is present on the COMSCREEN file. Those small area estimators which are based on counts will introduce an error into the estimates originating from the miscoding of industrial classification present on the COMSCREEN file. The estimates from the ratio-type estimators will not only be subject to misclassification errors, but in addition, they will be influenced by the replacement of the Gross Business Income concept with the Revenue Canada Sales concept in the estimation procedure. The estimator that is expected to be most affected by erroneous information on the COMSCREEN universe file is the ratio-synthetic estimator because it relies most extensively on COMSCREEN data. Therefore, the analysis to follow will concentrate on the effect of the discrepancies between the two files on the estimates produced by the ratio-synthetic estimator. The results will represent the maximum error introduced into the estimates by any estimator due to the discrepancies between the files.

5.1 Analysis of errors introduced to the ratio-synthetic estimates due to coding and conceptual differences, at the Canada level.

For the purposes of this analysis, it was assumed that the small business records available on the MATCHED file constitute the unincorporated universe. This simplication was not expected to effect the

findings at the Canada, province or industry breakdown level. The ratio of W&S and GBI at the provincial level in each industry was applied to all the Census Division GBI totals within that province and industry to produce Wages and Salaries estimates at the Census Division industry grouping level. This estimation was carried out twice; first the ratios were applied to GBI using STC SIC codes then they were applied to Sales using RC SIC codes. The resulting two sets of estimates were then compared to the true Wages and Salaries total at the Census Division per industry grouping level to produce two sets of errors at that level. The errors were calculated according to the following formulas:

$$\text{error (ai)} = \frac{\sum_{a=1}^{A} \sum_{k=1}^{N_{ai}} W\&S_{aik}}{\sum_{a=1}^{A} \sum_{k=1}^{N_{ai}} GBI_{aik}} X_{ai.} - W\&S_{ai.} \qquad (4.1)$$

A - number of Census Divisions
N_{ai} - number of units in domain ai

where in the first estimation (STC)

$X_{ai.}$ is total GBI in Census Division a industry grouping i (STC SIC)

in the second estimation (RC)

$X_{ai.}$ total sales in Census Division a industry grouping i (RC SIC)

These errors can be summarized at the Canada industry grouping level through several statistics, such as the mean absolute percentage error:

$$MAPE_i = \sum_{a=1}^{A} \left| \frac{error(ai)}{W\&S_{ai}} \right| / A \qquad (4.2)$$

and the total absolute percentage error

$$TAPE_i = \frac{\sum_{a=1}^{A} |error(ai)|}{W\&S_{.i.}} \qquad (4.3)$$

While the MAPE statistic gives equal weight to each Census Division regardless of size, the TAPE statistic takes size into consideration to some extent.

Table 5.1 compares the errors from the two sets of estimates at the Canada Major Division industry grouping level based on the above two statistics.

TABLE 5.1 CANADA - Comparison of ratio estimation errors using STC and RC coding - breakdown by SIC

	(1)	(2)	(3)	(4)	(5)	(6)=(5)-(4)
Major Division	MAPE STC	MAPE RC	MAPE RC / MAPE STC	TAPE STC	TAPE RC	TAPE RC / TAPE STC
1. Agriculture	2.658	3.114	1.171	.388	.582	.194
2. Fishing	.459	000	000	.674	0	-.573
3. Logging & forestry	1.298	1.346	1.037	.282	.334	.052
4. Mining	4.724	1.552	.329	.368	.370	.002
5. Manufacturing	.625	1.267	2.027	.189	.413	.224
6. Construction	.315	.325	1.032	.123	.147	.023
7. Transportation	.729	.966	1.325	.264	.329	.065
8. Communication	2.132	3.622	1.698	.249	.531	.282
9. Wholesale	2.595	1.408	.543	.359	.456	.097
10.Retail	.307	.422	1.374	.155	.185	.030
12.Real Estate	7.371	1.791	.243	.459	.464	.005
13.Business Service	10.664	10.834	1.016	.346	.436	.090
15.Educational Service	4.684	1.547	.330	.316	.590	.274
16.Health and Social	1.478	1.742	1.179	.336	.423	.087
17.Accommodation	.449	.464	1.033	.121	.132	.011
18.Other Services	1.088	2.625	2.413	.202	.251	.049

Table 5.1 indicates that based on STC codes and concepts alone, only a few industries show acceptable quality estimates at the small area level as attested by the entries in column 1. Based on the MAPE measure estimates in only Fishing, Construction, Retail and Accommodation industry are reliable at the Census Division level.*

Judging by column 2 (MAPE using RC codes and concepts) the errors increase substantially in most industries when the complication of incompatible codes and concepts is added to the estimation procedure. The only industries in which the combined effect of estimation and coding on the estimates is relatively small is again Fishing, Construction, Retail and Accommodation. According to column 3, misclassification and conceptual differences increased the error by a maximum of 37% in these four industries. In a few industries such as Mining, Wholesale, Real Estate and Educational Services, the estimates actually got closer to the true value when RC codes were used but the error was still around 150% ($MAPE_{RC}$) - indicating that the estimates are still unsuitable for publication.

According to the TAPE. statistic, the use of Revenue Canada SIC codes and concepts introduced some extra error in the estimates for practically al industries. These errors range from an extra .5% of the total wage bill in the Real Estate industry to an extra 28.2% of the total Wages and Salaries in the Communication industry.

*However, the previous study by the authors (1984) indicated that reliable estimates can be produced at Census Division level by the ratio-type estimator by treating industries other than Construction, Retail and Accommodation as one common industry grouping. These four new industry estimates at the Census Division level can then be added up to yield Census Division data.

TABLE 5.2 CANADA - Comparison of ratio estimation errors using STC and RC Coding - Breakdown by Province

Province	No. of Units	TOTAL W&S (STC)	MAPE STC	MAPE RC	MAPE RC / MAPE STC	TAPE RC - TAPE STC
Alberta	2338	33673	1.233	1.348	1.094	.044
British Columbia	3799	58399	3.086	1.194	.388	.035
Manitoba	1526	20902	1.000	1.250	1.250	.048
New Brunswick	988	13939	1.128	2.683	2.550	.091
Newfoundland	560	7106	1.233	1.493	1.211	.081
Nova Scotia	1235	19451	1.367	2.125	1.550	.064
Northwest Territories	15	266	0.425	0.482	1.134	.054
Ontario	14523	232912	1.074	1.232	1.147	.050
P.E.I.	164	2358	1.017	1.185	1.164	.029
Quebec	10045	178298	2.398	2.636	1.099	.057
Saskatchewan	1368	17625	1.831	1.098	.546	.044
Yukon	25	348	.159	.258	1.622	.130

Because of the presence of industries in which the estimation procedure produces extremely large errors, the MAPE statistics indicates unacceptable quality estimates when summing up Census Divisions by provinces (Table 5.2) with the exception of estimates in Yukon and the Northwest Territories.* The deterioration in mean absolute percentage error due to using Revenue Canada coding, ranges from 9% in Alberta to 155% in New Brunswick. In British Columbia and Saskatchewan the estimates move closer to the value where estimation is based on RC industrial classification and concepts, but the errors are still too large to allow for publication.

*(In these latter two provinces there are no businesses operating in the industries characterized by high error estimates.)

The preceding analysis gave a good insight into the expected size of errors in small areas all across Canada produced by the ratio-synthetic estimator when using two sets of classification codes and concepts. It did not take into consideration, however, the effect of sampling nor did it deal with the performance of the other 7 estimators subjected to RC coding. To obtain this type of information, a simulation study was carried out on small business data from the Accommodation industry in Nova Scotia and Ontario.

5.2. Analysis of a simulation study on Nova Scotia and Ontario data to assess the effect of Revenue Canada coding on the performance of eight estimators.

The file used as the universe file in this simulation study contained all the records from the MATCHED file from Nova Scotia and Ontario that belonged to the Accommodation Industry either according to Statistics Canada or Revenue Canada SIC coding. This selection procedure yielded 89 records in Nova Scotia and 1902 records in Ontario from which 250 25% samples were drawn. In the estimation phase whenever the information originated from the universe file, RC coding and concept were applied. On the other hand, the calculations concerning sample data were all based on STC codes and concepts. Thus the only estimator that was not expected to be effected by Revenue Canada coding was the direct blow-up estimator that relied strictly on the Statistics Canada sample data. The rest of the estimators all used universe file information either in form of Business counts or auxiliary variable (Sales) totals. Tables 5.3 and 5.4 summarize the performance of the estimators in Nova Scotia and Ontario according to the same three measures as before.

TABLE 5.3 Performance of Estimators in Nova Scotia Accommodation - using RC SIC Coding and Concepts

	Estimators							
Measure	DIR	SYN/C	SYN/R	REG/C	REG/R	POST/C	MREG/R(1)	MREG/R(2)
RE	1.000	.403	.340	.730	.449	.769	.365	.353
RB	.045	.429	.411	.103	.104	.386	.252	.256
CRMSE	1.076	.504	.443	.856	.643	.734	.440	.440

TABLE 5.4 Performance of Estimators in Ontario Accommodation - using RC SIC Coding and Concepts

	Estimators							
Measure	DIR	SYN/C	SYN/R	REG/C	REG/R	POST/C	MREG/R(1)	MREG/(2)
RE	1.000	.898	.350	.781	.512	.800	.455	.433
RB	.031	.359	.221	.084	.081	.099	.116	.149
CRMSE	.571	.366	.225	.450	.296	.454	.246	.247

To put the information from Tables 5.3 and 5.4 into perspective, it is necessary to compare them to the corresponding entries in Tables 2.1 and 3.1. The impact of RC coding on the errors of estimation is measured by the ratio of the RC table entries and STC table entries as shown in Table 5.5.

TABLE 5.5 Performance of estimators using RC coding relative to the performance of estimators using STC coding

		Estimators							
Measure	Prov.	DIR	SYN/C	SYN/R	REG/C	REG/R	POST/C	MREG/R(1)	MREG/R(2)
RE RC	N.S.	1.000	1.092	1.255	1.007	1.056	.950	1.119	1.169
RE STC	Ont.	1.000	1.067	.994	1.014	1.087	.988	1.102	1.099
RB RC	N.S.	1.000	.935	1.138	2.575	2.311	1.149	1.292	1.225
RB STC	Ont.	1.000	1.078	1.308	3.818	5.400	1.547	2.274	1.655
CRMSE RC	N.S.	1.000	.946	1.099	1.012	1.027	.983	1.058	1.060
CRMSE STC	Ont.	1.000	1.076	1.293	1.037	1.117	.991	1.194	1.223

As expected, the direct estimator is the only one that was not affected by RC classification. In terms of root mean square error, the ratio-synthetic estimates changed most in both provinces followed by MREG(2) and MREG(1). The least affected estimates are the ones produced by the regression count estimators. Errors actually become smaller when universe counts in the post-stratified estimation were obtained using RC classification. Concerning bias,
the originally nearly unbiased regression-count and regression-ratio estimates deteriorated the most. The bias associated with the ratio-synthetic increased only slightly.

In terms of root mean square error, the impact of RC coding was significantly larger in Ontario than in Nova Scotia. For example, the estimates produced by the ratio-synthetic estimator showed 29% larger errors in Ontario than previously as opposed to only 10% larger errors in Nova Scotia. (The ratio of RC and STC CRMSE-s is 1.293 and 1.099 in Ontario and Nova Scotia

respectively). These figures agree remarkably well with the ratio of MAPE measures in the Accommodation industry calculated in Section 5.1. According to those calculations, the MAPE statistic in Ontario Accommodation when using RC codes and concepts was .229 yielding a MAPE ratio of 1.283. In Nova Scotia the corresponding statistic was .402 yielding a ratio of 1.084.

This type of agreement is relatively close even at the Census Division level. To illustrate this point, Table 5.6 lists the SYN/R CRMSE measures from the simulations together with the absolute percentage error as obtained from the MATCHED file using the two sets of codes and concepts in the Accommodation industry for 16 Census Divisions in Nova Scotia.

TABLE 5.6 - NOVA SCOTIA - Absolute percentage error per Census Division compared to coefficient of RMSE for ratio-synthetic estimator according to STC and RC coding

Census Div.	No.of units	|error| W&S STC	SYN/R C of RMSE STC	|error| W&S RC	SYN/R C OF RMSE RC	error RC / error STC	C of RMSE RC / C OF RMSE STC
01	1	.289	.310	.237	.313	.817	1.009
04	1	.120	.164	.074	.171	.618	1.042
18	2	2.796	2.789	2.640	2.789	.944	1.000
03	3	.065	.129	.231	.225	3.539	1.744
07	3	.174	.203	.208	.204	1.195	1.004
14	3	.327	.342	.355	.341	1.084	.996
06	4	.048	.119	.087	.119	1.821	1.000
10	4	.413	.427	.355	.428	.860	1.002
15	4	.599	.608	.533	.608	.890	1.000
02	5	.227	.253	.168	.243	.741	.960
05	5	.103	.151	.058	.151	.562	1.000
08	5	.259	.229	.479	.465	1.852	2.031
12	5	.274	.293	.379	.365	1.383	1.245
11	8	.027	.113	.246	.238	9.153	2.106
17	12	.092	.142	.324	.310	3.505	2.183
09	21	.067	.129	.014	.113	.212	.875
Mean		.368	.403	.399	.443		
Ratio of means						1.084	1.099

There is a strong positive correlation between the absolute error and CRMSE statistics for both the STC and RC estimates. Similarly, the ratios (column 5 and 6) show a relatively strong relationship. In general, the CRMSE measure slightly overestimates the absolute percentage error statistic.

The entries in Table 5.6 are sorted according to the size of the Census Divisions. It can be observed that in general, the absolute percentage error (and correspondingly CRMSE) decreases with increasing sample size, as expected. What is not intuitively obvious, however, is that the extra error caused by RC coding, as measured by the ratio of error statistics increases as the small areas get larger. This phenomenon could originate from the fact that misclassification by RC is not uniformly distributed over all Census Divisions, and that misclassification is more likely to occur, the bigger the small area. Once the small area reaches a certain size though, the proportion of misclassified business units per area will become constant and the observed effect of size on coding discrepancy will disappear. This was the case in Ontario where a tabulation corresponding to Table 5.6 failed to reveal a correlation between size and CRMSE ratio for Census Divisions with more than 20 businesses in them.

5.3 Effect of RC coding on the ranking of estimators

Having examined what happened to estimates produced by the ratio-synthetic estimator, after introducing RC coding and concepts, it is time to assess whether the resulting increase in error had an impact on the ranking of this estimator relative to the others. Tables 5.7 and 5.8 summarize the results in Nova Scotia and Ontario.

TABLE 5.7 NOVA SCOTIA - Ranking of estimators according to three measures

		Estimators							
Measure	Origin of SIC Code	DIR	SYN/C	SYN/R	REG/C	REG/R	POST/C	MREG/ R(1)	MREG/ R(2)
RE	STC	8	4	1	6	5	7	3	2
	RC	8	4	1	6	5	7	3	2
RB	STC	2-3	8	7	1	2-3	6	4	5
	RC	1	8	7	2-3	2-3	6	4	5
C of MSE	STC	8	4	1	7	5	6	2-3	2-3
	RC	8	4	1-2-3	7	5	6	1-2-3	1-2-3

TABLE 5.8 ONTARIO - Ranking of estimators according to three measures

		Estimators							
Measure	Origin of SIC Code	DIR	SYN/C	SYN/R	REG/C	REG/R	POST/C	MREG/R(1)	MREG/ R(2)
RE	STC	8	7	1	5	4	6	3	2
	RC	8	7	1	5	4	6	3	2
RB	STC	3	8	7	2	1	5	4	6
	RC	1	8	7	3	2	4	5	6
C of MSE	STC	8	5	1	6	4	7	2-3	2-3
	RC	8	5	1	6	4	7	2-3	2-3

In terms of efficiency and root mean square error, the ratio-synthetic estimator is still number one or tied for first place with the two modified regression estimators. In effect, the relative ranking of all the other estimators

remained exactly the same even after taking into account the misclassification. A very slight shift occurred in terms of the bias among the first 5 ranking estimators in Ontario and among the first three ranking ones in Nova Scotia.

In concluding it can be stated that the estimators considered are not especially sensitive to misclassification of industry grouping or conceptual differences of the variables used. The introduced error due to these factors is relatively small and although they effect different estimators to different degree, they do not influence the overall ranking of the estimators.

6. CONCLUSIONS

1. The two new modified regression-ratio estimators developed are successful in cutting down the root mean square error of the estimates of the regression-ratio estimator without increasing the bias to unacceptable levels. These two new estimators can be considered as a good alternative to the ratio-synthetic one.

2. Moving from a small province like Nova Scotia to a large one like Ontario reduces the error associated with the estimates, but does not significantly effect the ranking of the estimators.

3. Using Revenue Canada coding and concepts introduces added error in the estimates, but without affecting their publishability. Estimates in industries that were publishable at the small area level under STC coding are still of acceptable quality.

4. The simulation study conducted in the Accommodation industry where the correlation between W&S and GBI is strong, indicates that it is the ratio-synthetic estimator that is most sensitive to the misclassification and concepts introduced by using the Revenue Canada file, with the rest of the estimators following closely. Deterioration in quality is more evident in a larger province than in a smaller province. However, the overall ranking of the estimators does not change due to the added error. The ratio-synthetic estimator still outperforms the others with the two modified regression estimators ranking as close second - suggesting that these latter two estimators could be used to advantage in producing reliable Wages and Salaries estimates that are not prone to excessive levels of bias.

APPENDIX "A"

ESTIMATORS USED

1. Direct Estimator (DIR)

$$t_1(ai) = N/n \sum_{k=1}^{n_{ai}} y_{aik} = N/n\ y_{ai.}$$

where y_{aik} is the Wages and Salaries value of k-th sampled unit in the i-th industry in the a-th Census Division.

2. POST-STRATIFIED ESTIMATOR (POST/C)

$$t_2(ai) = N_{ai}/n_{ai}\ y_{ai.}$$

N_{ai} - population domain size

n_{ai} - sample domain size

3. COUNT-SYNTHETIC ESTIMATOR (SYN/C)

$$t_3(ai) = N_{ai}\ \bar{y}_{.i.}$$

where $$\bar{y}_{.i.} = \sum_{a=1}^{A} \sum_{k=1}^{n_{ai}} y_{aik} / \sum_{a=1}^{A} n_{ai}$$

A is the number of Census Divisions

4. RATIO-SYNTHETIC ESTIMATOR (SYN/R)

$$t_4(ai) = \bar{y}_{.i.}/\bar{x}_{.i.}\ X_{ai.}$$

where $x_{.i.}$ - is the overall sample mean of Gross Business Income in i-th industry.

X_{ai} is the population total of Gross Business Income in the ai domain.

5. REGRESSION-COUNT ESTIMATOR (REG/C)

$$t_5(ai) = t_3(ai) + N/n\ n_{ai}\ (\bar{y}_{ai.} - \bar{y}_{.i.})$$

where $\bar{y}_{ai.} = \sum_{k=1}^{n_{ai}} y_{aik}/n_{ai}$

6. REGRESSION-RATIO ESTIMATOR (REG/R)

$$t_6(ai) = t_4(ai) + N/n\ n_{ai}\ (\bar{y}_{ai.} - \bar{y}_{.i.}/\bar{x}_{.i.}\ \bar{x}_{ai.})$$

where $x_{ai.} = \sum_{k=1}^{n_{ai}} x_{aik}/n_{ai}$

7. MODIFIED REGRESSION-RATIO ESTIMATOR (1) MREG/R(1)

$$t_7(ai) = \bar{y}_{.i.}/\bar{x}_{.i.}\ X_{ai.} + D_{ai} \sum_{k=1}^{n_{ai}} (y_{aik} - \bar{y}_{.i.}/\bar{x}_{.i.}\ x_{aik})$$

where $D_{ai} = \begin{cases} N_{ai}/n_{ai} & \text{if } n_{ai}/N_{ai} \geq n/N \\ (N/n)^2\ n_{ai}/N_{ai} & \text{if } n_{ai}/N_{ai} < n/N \end{cases}$

8. MODIFIED REGRESSION - RATIO ESTIMATOR (2) - MREG/R(2)

$t_8(ai)$ same as $t_7(ai)$

where $D_{ai} = \begin{cases} N_{ai}/n_{ai} & \text{if } n_{ai}/N_{ai} \geq n/N \\ o & \text{otherwise} \end{cases}$

APPENDIX B

MEASURES USED FOR ASSESSING PERFORMANCE

1. Relative bias of the m-th estimator:

$$\overline{RB}[t_m(\ i)] = 1/A \sum_{a=1}^{A} \left| \frac{\bar{t}_m(ai) - Y_{ai.}}{Y_{ai.}} \right|$$

$$= 1/A \sum_{a=1}^{A} \frac{|\bar{B}[t_m(ai)]|}{Y_{ai.}}$$

$$m = 1,2,\ldots,8$$

2. Relative efficiency of the m-th estimator:

$$\overline{RE}\ [t_m(\ i)] = 1/A \sum_{a=1}^{A} \left\{ \frac{\overline{MSE}\ [t_m(ai)]}{\overline{MSE}\ [t_1(ai)]} \right\}^{1/2}$$

$$\overline{MSE}\ [t_m(ai)] = \sum_{r=1}^{250} [t_m{}^r(ai) - Y_{ai.}]^2/250$$

$$m = 2,\ 3,\ldots,8$$

3. Coefficient of root mean square error:

$$\overline{CRMSE}\ [t_m(\ i)] = 1/A \sum_{a=1}^{A} \frac{\{\overline{MSE}[t_m(ai)]\}^{1/2}}{Y_{ai.}}$$

$$m=1,2,\ldots,8$$

REFERENCES

Dagum, E.B. Hidiroglou, M.A., Morry, M. (1984)
The Use of Administrative Records to Estimate Wages and Salaries for Small Businesses in Small Areas, 1984 Proceedings of the Economic Statistics Section, American Statistical Association, forthcoming.

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

Hidiroglou, M.A., Morry, M., Dagum,E.B. Rao, J.N.K. and Särndal, C.E. (1984), Evaluation of alternative small area estimators using administrative records, 1984 Proceedings of the Survey Methodology Section, American Statistical Association, forthcoming.

Hidiroglou, M.A. and Särndal, C.E. (1984) Experiments with modified regression estimators for small domains, Technical paper, Statistics Canada.

Särndal, C.E. (1981). Frameworks for inference in survey sampling with applications to small area estimation and adjustment for non-response. Bull. Int. Sta. Ind., 49:1 494-513 (Proceedings, 43rd Session, Buenos Aires).

Särndal, C.E. and Räbäck, G. (1983). Variance Reduction and Unbiasedness for Small Domains Estimators. Statistical Review, 5, 33-40.