WP. 2
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE

CONFERENCE OF EUROPEAN STATISTICIANS

Work Session on Statistical Data Editing
(Ottawa, Canada, 16-18 May 2005)

Topic (i): Editing administrative data and combined data sources

USE OF ADMINISTRATIVE DATA IN STATISTICS CANADA'S ANNUAL SURVEY OF MANUFACTURES

Invited Paper

Submitted by Statistics Canada[1]

I. INTRODUCTION

1. In an ongoing effort to reduce respondent burden, to reduce the cost of survey programs and to improve data quality, Statistics Canada has been working to increase the use of administrative data in its survey programs. One survey that makes extensive use of administrative data is the Annual Survey of Manufactures (ASM). The ASM is an annual survey covering all manufacturing establishments as defined by the North American Industrial Classification System (NAICS). Information collected by the ASM includes financial variables (revenues, expenses and salaries for example) and commodity variables (manufacturing inputs and outputs for example). Due to certain agreements with principal data users, the ASM is required to produce micro-data for a large portion of its target population. This pseudo-census is achieved by extensive use of tax data. This paper describes the use of the tax data in the ASM, with particular attention to the Edit and Imputation stage. Section 2 provides a quick overview of the tax data used by ASM. The ASM is described in section 3 and the use of tax data in the ASM is presented in section 4. Section 5 provides a summary of the work done to evaluate the impact of the use of tax data on the quality of the survey estimates. Conclusions and future work are given in section 6.

II. TAX DATA

2. Statistic Canada has long recognized the advantages of an increased use of administrative data for Statistics Canada's Business Survey Program in terms of reducing respondent burden and potential cost savings. In response, extensive programs have been put in place to provide administrative data to Statistics Canada surveys. Two of these programs, covering incorporated and non-incorporated businesses, are administered by the Tax Data Division of Statistics Canada. The tax data is collected by the Canadian Revenue Agency (CRA) and Statistics Canada has access to the tax files for statistical purposes through a data sharing agreement with the CRA. In this section, the two programs responsible for producing the required data are described.

A. T1 (Non-incorporated) Tax Data

3. While detailed data for incorporated businesses are available through the T2 tax program, for the non-incorporated businesses only estimates of totals for specific variables at the industry by province level are available through tax data. These estimates are produced based on a simple random sample of

[1] Prepared by Steve Matthews, Canada Post and Wesley Yung, Statistics Canada

non-incorporated businesses, for which several variables are obtained by transcribing data from paper tax forms or from electronically filed tax forms. The set of variables available for the non-incorporated businesses is not nearly as extensive as the set for incorporated businesses and the data are available only for the sampled units. The data for non-incorporated businesses are subjected to an edit and imputation process, and the estimates are produced using a calibration estimator that benchmarks the estimates to the total revenues for the T1 population. For more information on Statistics Canada's T1 tax program, see Hutchinson, Jocelyn and Cooray (2004).

B. T2 (Incorporated) Tax Data

4. Incorporated businesses in Canada are now required by law to provide the information contained in their financial statements (T2 tax form) in the Generalized Index of Financial Information (GIFI) format. The GIFI format is an extensive list of financial statement items where each item has a unique code. CRA captures the tax files from all incorporated businesses in Canada and makes this data available to Statistics Canada. This data source provides a very detailed breakdown of the expenses, revenues and inventories for each business (approximately 700 variables divided into 7 sections) but only eight fields are mandatory (the 7 section totals and Net Income/Loss). CRA collects the data throughout the year and provides the data to Statistics Canada on a monthly basis where several verification processes are performed. These processes include editing the data to ensure that the data are clean and that the data balance; a generic to detail allocation where totals are allocated to the detailed level necessary; and transforming the data to a standard format to facilitate tax data use by business surveys. Once a year, annual processes, such as imputation, are executed. For more details on the processing of GIFI data at Statistics Canada, see Hamel and Belcher (2002).

III. THE ANNUAL SURVEY OF MANUFACTURES

5. The Annual Survey of Manufactures is administered as part of Statistics Canada's Unified Enterprise Survey Program (UES). The UES has united separate business surveys into a single master survey program with the goals of collecting more industry and commodity detail at a provincial level, of having survey questionnaires with a consistent look, structure and content, of reducing respondent burden by avoiding overlap in survey questionnaires and of employing similar survey methodologies. Although the ASM uses the same sampling frame as the UES, it does differ in some aspects, such as the survey methodology, due to the nature of the details collected and the importance of its estimates. The estimates from the ASM are used fairly widely since the Manufacturing industry represents a significant portion of the business activity in Canada; roughly 17% of the Gross Domestic Product. In this section, some details of the ASM methodology will be presented.

A. Target Population

6. The target population for the ASM consists of all incorporated and non-incorporated businesses in Canada who are involved in the manufacturing industry. A survey frame is produced on an annual basis using Statistics Canada's Business Register (BR), which contains all businesses in Canada. All businesses on the BR have a statistical structure consisting of four levels in the following order: the enterprise, the company, the establishment and the location. An enterprise can consist of one or many companies and a company can consist of one or many establishments and so on. For the majority of businesses in Canada, the four levels coincide and the business is known as a simple business. If a business is not simple, then it is considered a complex business. In addition to the statistical structure, a number of administrative variables are available on the BR including the industry classification based on NAICS, geographic classification and an estimate of annual revenues.

7. The target population of the ASM is highly skewed with relatively few businesses representing a large share of the total revenue and expenses. In order to reduce survey costs and respondent burden on the smaller businesses in the target population, a large number of units are assigned to a 'take-none' portion where no data collection is performed. Based on frame information, units in the take-none portion represent less than 10% of the overall economic activity. Estimation of the contribution from the

take-none portion is done using tax data and is discussed in section 3.5. Once the take-none units are identified, the remaining units make up the survey portion.

B. Sample Design

8. The sample design within the survey portion is a stratified simple random sample with take-all strata and take-some strata. All complex units, i.e. units that are not defined as simple, are placed in the take-all strata and are sampled with probability 1. The remaining units, i.e. the simple units, are then stratified according to industrial classification (at the 3-digit NAICS level), geographic classification (provinces and territories) and a size measure based on annual revenues from the BR. Stratum boundaries and sample allocation are determined by applying the Lavallée-Hidiroglou (1988) method within each industry by province cell for pre-specified target CVs. Note that the Lavallée-Hidiroglou algorithm identifies additional take-all units.

C. Data Collection

9. The ASM questionnaire consists of two parts; a financial portion and a commodity portion. The financial portion is quite detailed and asks for a detailed breakdown of revenues, expenses and inventories. These data items are referred to as financial variables and are similar to information available from tax data. The commodity portion of the questionnaire asks for details such as the amount of each type of goods manufactured and those used as a manufacturing input. These items are referred to as commodity variables and will not be discussed further in this paper.

10. The ASM is a mail-out/mail-back survey with telephone follow-up. The mail-out occurs between November of the reference year and March of the year following the reference year depending on the fiscal year end of the selected business. Follow-up is managed by a score function to ensure efficient follow-up procedures. For more on the ASM score function, see Philips (2003).

D. Edit and Imputation

11. As with most large business surveys, edits are applied to the data at various stages of the processing system to ensure accuracy and coherence. In terms of imputation, the ASM is different than most business surveys due to the need for complete financial data for all units in the survey population. As the collection budget for the ASM is limited, imputation methods using T2 tax data extensively are used to produce this pseudo-census. In addition to the use of T2 data in the survey portion, it is used in conjunction with T1 data to estimate the contribution of the take-none portion. Take-none estimation is described in the following section and the use of T2 data in the survey portion is described in section 4.

E. Estimation

12. Once imputation is complete, micro-data is available for all units in the survey portion and estimation for the survey portion consists simply of aggregating the micro-data to obtain estimates. For the take-none portion, only Total Revenues is estimated and this estimate is based on T1 and T2 tax data as both incorporated and non-incorporated businesses make up the take-none portion. The contribution from the incorporated businesses will be obtained from the T2 tax data and non-incorporated portion will come from the T1 tax program. For the incorporated businesses, revenue values will come from their GIFI forms and will simply be added up. For the non-incorporated businesses, a sample of their T1 tax records will be selected, processed and a weighted estimate, benchmarked to Total Revenues from the T1 population, will be produced.

IV. USE OF TAX DATA IN E&I FOR THE SURVEY PORTION

13. As with most types of administrative data, there are certain limitations to the T2 data as pertains to the ASM. Some of the challenges faced by the ASM were the differences in the collection units between tax and ASM and the availability of tax data corresponding to ASM variables. One important

difference between the T2 data and the data collected by the ASM is the collection unit. The CRA receives data at the legal entity level (a unit defined according to legal concepts), while the survey data is generally collected at the statistical entity level as defined by the statistical structure on the BR. For simple businesses, the legal entity and the statistical entity correspond and the use of tax data for these units is straightforward. However, for complex businesses the use of tax data is much more difficult as an additional process is required to allocate the tax data from the legal entity to the statistical entity (which do not correspond in complex businesses). Due to this difficulty, the ASM uses tax data for simple businesses only and the complex businesses, treated as take-all units, will continue to be surveyed.

14. Of the 81 financial variables collected by the ASM, T2 data is able to provide 46 variables that are equivalent in concept to the ASM variable. A number of approaches were proposed for the use of tax data for these variables, including direct tax replacement, and use as auxiliary variables for model-based imputation methods. The first priority was to identify the variables that could be directly tax replaced. Before accepting the data from the tax files at face value, an evaluation study was done to compare the values obtained from the tax data with the data collected via the reference year 2001 ASM survey. Comparisons were done at the micro level according to several criteria to evaluate each variable individually. The analysis included a number of aspects, based on the set of businesses for which we have both reported tax data and reported survey data (approximately 6,000) including:

(a) Correlation Analysis: The correlation between the reported tax and survey data was calculated at several levels to determine if a linear relationship exists between the survey and the tax value. The analysis showed that for some variables, such as Total Expenses, the correlation was very high (approximately 90%) but for others, such as Maintenance and Repair Expenses, it was essentially zero.

(b) Consistency of Details: Since not all businesses incur expenses of a certain type, it is valid to receive zero values for many of the detailed revenues, expenses and inventories. This analysis verified whether or not the same units tended to report zero values for the same details according to the tax data and the survey data. The analysis showed similar results as the correlation analysis. That is, some variables matched very well (98% for Total Expenses) and some not so well (50% for Maintenance and Repair Expenses)

(c) Distribution of Ratios: The ratio of the survey to the tax value was calculated for each unit and for each variable, and histograms of the ratio values were produced to identify potential biases or increased variances that would result from substituting tax data for missing survey values. The results of this analysis showed that the ratios of survey to tax data are not that stable. The percentage of ratios that were in the range of (0.9, 1.1) varied from 60% for Total Expenses to 17% for Maintenance and Repair Expenses.

(d) Population Estimates of Totals: Weighted estimates were produced based on the set of units for which both sources of data were available, using the survey data and the tax data to measure the variable of interest. The relative difference between the tax-based estimate and the survey-based estimate was calculated. The relative differences obtained varied from a low of 0.7% for Total Manufacturing Output to a high of 20% for Amortization and Depreciation Expenses.

15. The results of this analysis can be generalized to conclude that for variables that relate to very general concepts (for example Total Revenues, Total Expenses), direct tax replacement will not have an important impact on the data quality (both in terms of microdata and aggregated estimates). However, for many variables that relate to detailed concepts (Expenses for Advertising, Expenses for Depreciation/Amortization) direct tax replacement may affect the data quality. This could be in terms of a bias or an increased variance. For this reason, based on these analyses a group of 7 variables was identified for which survey data will be replaced with tax data for some businesses. For a more detailed description of the study and recommendations, please refer to Batten and Matthews (2003).

16. Assuming that these 7 variables will be completed by direct substitution of tax values, the challenge remains to produce micro-data for all of the other financial variables. Since these 7 variables correspond to the totals of each financial section, this task consists of estimating the distribution of detailed revenues, expenses and inventories that should be imposed on the record requiring imputation. This is done in the general imputation system using the following methods:

(a) Historical Imputation – In this method, if the business reported their survey data in the previous reference year, the reported distribution from each section is copied to the current year and pro-rated to the totals coming from the tax data. Note that for the majority of records that are non-sampled, historical data will not be available.
(b) Ratio Imputation – This method is used if the business did not report their survey data in the previous reference year (this would be the case for most of the non-sampled records). Using the set of respondents from the current reference year, the ratio between the total and each detail is estimated within imputation groups, and these ratios are applied to the total coming from the tax data. Each section is then pro-rated to ensure that the totals are respected.
(c) Nearest Neighbor Donor Imputation – This method is used if the auxiliary data that is required for Historical or Ratio Imputation is not available. In Nearest Neighbor Donor Imputation, the data from the responding record that is most similar to the record requiring imputation is used to complete the entire record. All of the donor-imputed values are adjusted to reflect the difference in annual revenues of the donor and recipient.

17. These methods are applied within imputation classes, which are constructed according to geographic, and industry classifications. In general, we attempt to construct these classes at a detailed level to reduce biases, but in some cases the groups need to be aggregated to produce a suitable number of respondents in each group.

V. IMPACT OF TAX DATA

18. In order to evaluate the impact of using tax data on the quality of the estimates produced, estimates of population totals and corresponding variances were calculated under three different scenarios. The first scenario was under the assumption that tax data is not available and the population total is estimated using weights based on the sample design. The second scenario assumed that tax data is not available, but a pseudo-census is produced using revenue from the BR as an auxiliary variable for imputation purposes. The final scenario assumed that tax data is available for use for replacing non-responding and non-sampled units as described in section 4. For variance estimators, we use the Shao-Steel approach (1999) to take into account the imputation performed. The imputation variance could be considerable under the scenarios where imputation or tax data replacement is used for non-responding and non-sampled units.

19. In the Shao-Steel approach, the traditional view of non-response and sampling is reversed. Traditionally, it is assumed that a sample is drawn from the population, and some of the sampled units do not respond. For the Shao-Steel approach, it is assumed that the units in the population are divided into respondents and non-respondents and then a sample is drawn from this population. This allows one to derive the variance estimators using reversed conditional arguments. For example, under the traditional approach, the expectation with respect to the non-response mechanism is conditional on the realized sample, whereas under the Shao-Steel approach, the expectation and variance with respect to the non-response mechanism is not conditional, but the variance and expectation with respect to the sample is conditional on the non-response mechanism. The two approaches yield similar results with small sampling fractions, however the variance estimators from the Shao-Steel approach can be simpler to derive, can be extended to complex imputation strategies, and are valid regardless of the sampling fractions.

20. We now compare three estimators that could be used to produce population level estimates from the ASM data. The first two would be possible in the absence of tax data, and the third relies on the availability of tax data in the population. Note that under all three scenarios, we assume that the

imputation classes are mutually exclusive and exhaustive subgroups, which consist of aggregates of sampling strata. We assume that within each imputation class there is a uniform response mechanism when deriving the Shao-Steel variance estimates. That is, within each imputation class the probability of a unit responding is the same for all units and is independent of other units. Since the imputation classes are independent of each other in terms of both sampling and imputation, the variance estimates calculated at this level can be aggregated to produce variance estimates for estimates of totals at more aggregated levels.

A. Tax data unavailable, macro estimates only

21. If the tax data were not available, the massive imputation approach would not be feasible, and we would be relegated to a Horvitz-Thompson estimator. Ratio imputation (using an annual revenue estimate from the survey frame as the auxiliary variable *x*) could be used to complete the non-responding records among the sample, but would not produce micro-data for non-sampled records as is currently required. In this case, the estimator of the total is given by

$$\hat{Y}^{(1)} = \sum_h \sum_{i \in s_h} w_{hi} y'_{hi}$$

where (*hi*) represents the *i*-th unit in the *h*-th stratum, s_h is the set of sampled units in the h-th stratum, w_{hi} are the survey weights defined by the sample design, $y'_{hi} = a_{hi} y_{hi} + (1 - a_{hi})\hat{R}_x x_{hi}$, a_{hi} is the response indicator variable defined as $a_{hi} = 1$ for responding units and $a_{hi} = 0$ otherwise and

$$\hat{R}_x = \frac{\sum_h \sum_{i \in s_h} a_{hi} w_{hi} y_{hi}}{\sum_h \sum_{i \in s_h} a_{hi} w_{hi} x_{hi}}.$$

22. The estimated variance for this estimator derived under the Shao-Steel approach is given by $v^{(1)} = v_1^{(1)} + v_2^{(1)}$, where

$$v_1^{(1)} = \sum_h \left(1 - \frac{n_h}{N_h}\right) \frac{N_h^2}{n_h} \frac{\sum_{i \in s_h} (x_{hi} - x_{h\bullet})^2}{(n_h - 1)}$$

with $x_{hi} = y'_{hi} + a_{hi}\hat{c}_{hi}\hat{e}_{hi}$, $\hat{c}_{hi} = \dfrac{\sum_h \sum_{i \in s_h} w_{hi}(1 - a_{hi}) x_{hi}}{\sum_h \sum_{i \in s_h} w_{hi} a_{hi} x_{hi}}$ $x_{h\bullet} = \dfrac{\sum_{i \in s_h} x_{hi}}{n_h}$, $\hat{e}_{hi} = y_{hi} - \hat{R}_x x_{hi}$ and

$v_2^{(1)} = N\hat{p}(1 - \hat{p})\tilde{s}_d^2$ with

$$\hat{p} = \frac{\sum_h \sum_{i \in s_h} w_{hi} a_{hi}}{\sum_h N_h}$$

and

$$\tilde{s}_d^2 = \frac{\sum_h \sum_{i \in s_h} w_{hi} a_{hi} \hat{e}_{hi}^2 (\hat{c}_{hi} + 1)^2}{\sum_h \sum_{i \in s_h} w_{hi} a_{hi}}.$$

B. Tax data unavailable, pseudo-census

23. In order to satisfy the requirement for a pseudo-census, in the absence of tax data, an estimator could be used where the non-sample and non-responding units could be imputed using ratio imputation (with the annual estimated revenues, x_{hi}, as the auxiliary variable). In this case, we would produce a full population dataset so no weighting would be required and the estimator of the total would be

$$\hat{Y}^{(2)} = \sum_h \sum_{i \in s_h} d_{hi} y'_{hi} + \sum_h \sum_{i \in U_h \setminus s_h} (1 - d_{hi}) y^*_{hi},$$

with y'_{hi} as defined above, $y^*_{hi} = \hat{R}_x x_{hi}$, d_{hi} is the survey portion indicator variable defined as d_{hi}=1 for units in the survey portion and 0 otherwise, U_h is the set of population units in the *h*-th stratum and $U_h \setminus s_h$ is the set of non-sampled units in the *h*-th stratum. Since we proceed with the assumption of a census with non-response, the variance estimator includes only one term, representing the non-response variance. The estimated variance for this estimator derived under the Shao-Steel approach is given by $v^{(2)} = N\hat{p}(1-\hat{p})\tilde{s}_d^2$, with

$$\hat{p} = \frac{\sum_h \sum_{i \in U_h} a_{hi}}{\sum_h N_h}, \quad \tilde{s}_d^2 = \frac{\sum_h \sum_{i \in U_h} a_{hi} \hat{e}_{hi}^2 (\hat{c}_{hi} + 1)^2}{\sum_h \sum_{i \in U_h} a_{hi}}, \quad \hat{e}_{hi} = y_{hi} - \hat{R}_x x_{hi}$$

and

$$\hat{c}_{hi} = \frac{\sum_h \sum_{i \in U_h} w_{hi} (1 - a_{hi}) x_{hi}}{\sum_h \sum_{i \in U_h} w_{hi} a_{hi} x_{hi}}.$$

C. Tax data available, pseudo-census

24. The methodology that will be used in production would directly replace the tax data for non-respondents and out-of-sample units. Clearly this would not be possible without the availability of the tax data for all population units, and this satisfies the requirement to produce data for a pseudo-census. For the variables that are directly tax replaced, the estimate of the total is given by

$$\hat{Y}^{(3TR)} = \sum_h \sum_{i \in s_h} d_{hi} y'_{hi} + \sum_h \sum_{i \in U_h \setminus s_h} (1 - d_{hi}) z_{hi}$$

with $y'_{hi} = a_{hi} y_{hi} + (1 - a_{hi}) z_{hi}$, a_{hi} and d_{hi} as defined above and z_{hi} is the corresponding tax data value for the (*hi*)-th unit.

25. The estimated variance for this estimator derived under the Shao-Steel approach is given by $v^{(3TR)} = N\hat{p}(1-\hat{p})\tilde{s}_d^2$ with $\hat{p} = \frac{\sum_h \sum_{i \in U_h} a_{hi}}{\sum_h N_h}$, $\tilde{s}_d^2 = \frac{\sum_h \sum_{i \in U_h} a_{hi} \hat{e}_{hi}^2}{\sum_h \sum_{i \in U_h} a_{hi}}$, and $\hat{e}_{hi} = y_{hi} - \hat{R}_z z_{hi}$.

26. For variables that are imputed via ratio imputation based on a variable directly replaced by tax, the estimator of the total is given by

$$\hat{Y}^{(3RI)} = \sum_h \sum_i d_{hi} y'_{hi} + \sum_h \sum_i (1 - d_{hi}) y^*_{hi},$$

with y'_{hi} and d_{hi} as defined above, $y^*_{hi} = \hat{R}_z z_{hi}$ and

$$\hat{R}_z = \frac{\sum_h \sum_{i \in s_h} a_{hi} w_{hi} y_{hi}}{\sum_h \sum_{i \in s_h} a_{hi} w_{hi} z_{hi}}.$$

27. The estimated variance for this estimator derived under the Shao-Steel approach is given by

$v^{(3RI)} = N\hat{p}(1-\hat{p})\tilde{s}_d^2$ with $\hat{p} = \frac{\sum_h \sum_{i \in U_h} a_{hi}}{\sum_h N_h}$, $\tilde{s}_d^2 = \frac{\sum_h \sum_{i \in U_h} a_{hi} \hat{e}_{hi}^2 (\hat{c}_{hi}+1)^2}{\sum_h \sum_{i \in U_h} a_{hi}}$, $\hat{e}_{hi} = y_{hi} - \hat{R}_z z_{hi}$ and

$$\hat{c}_{hi} = \frac{\sum_h \sum_{i \in U_h} w_{hi}(1-a_{hi}) z_{hi}}{\sum_h \sum_{i \in U_h} w_{hi} a_{hi} z_{hi}}.$$

28. Using this approach, the estimates of totals and their corresponding variances were calculated for all possible domains at the All Manufacturing, Province, 3-digit NAICS, and 3-digit NAICS by Province level. The population used for the study was from reference year 2002, the latest data available. This population was reduced by excluding the imputation classes with fewer than 5 responding units from the total manufacturing population. This was done to avoid dealing with variance estimates that were based on very few observations and extremely sensitive to outliers. As well, we assume that all complex units responded so that we can measure the variance associated with the treatment of non-responding and non-sampled simple structured units only. Two variables were selected to provide examples of the quality achieved for tax-replaced variables and ratio imputed variables. For the tax replace variable, Total Expenses is used, and for the ratio-imputed variables, Total Energy Expenses is used. Table 1 provides a summary of the relative differences between the estimates of totals produced under each methodology, relative to the estimates produced by the Horvitz-Thompson estimator. The values given in the table are the median values of the absolute relative differences for the estimates at each level.

Table 1 - Median Value of Relative Differences between Estimates of Total Revenue at each level

	Tax data unavailable, pseudo-census		***Tax data available, pseudo-census***	
Level	**Total Expenses**	**Total Energy Expenses**	**Total Expenses**	**Total Energy Expenses**
All Manufacturing	1.81%	1.18%	0.53%	0.80%
PROVINCE	0.73%	0.47%	0.77%	0.64%
NAICS3	1.42%	0.92%	1.05%	1.50 %
NAICS3 x PROVINCE	0.03%	0.02%	1.32%	1.20%

29. It is clear from these results that the estimates from the different estimators are very similar. We know that the Horvitz-Thompson and Ratio Imputed estimators are unbiased, and the estimates for Total Expenses are in line with these, so it appears that the direct use of tax data has not introduced an appreciable bias to the estimates of the Total Expenses variable. This is not unexpected since the variables to be tax-replaced were selected based on a similar criterion. In fact, the relative difference Total Expenses at the "All manufacturing" level is smaller for the Tax Data Available/Pseudo-Census scenario (difference of 0.5%) than the Tax Data Unavailable/Pseudo-Census scenario (difference of 1.8%).

Table 2 – Comparison of CV's at each level

	Median CV of Horvitz Thompson		**Relative Efficiency of *Tax data unavailable, pseudo-census***		**Relative Efficiency of *Tax data available, pseudo-census***	
Level	**Total Expenses**	**Total Energy Expenses**	**Total Expenses**	**Total Energy Expenses**	**Total Expenses**	**Total Energy Expenses**
All Manufacturing	0.28%	0.31%	0.96	1.23	0.46	1.41
PROVINCE	0.64%	0.66%	0.97	1.02	0.41	0.97
NAICS3	1.09%	1.48%	1.13	1.11	0.41	1.15
NAICS3 x PROVINCE	1.46%	1.80%	1.00	1.00	0.46	1.01

30. Table 2 gives the median of the estimated co-efficients of variation (CV) at each domain level from the Horvitz-Thompson weighted estimator, and the median of the relative efficiency of the alternative estimators for the estimates at each level. The relative efficiency is calculated as the CV of the estimator divided by the CV of the Horvitz-Thompson estimator. A median relative efficiency less than one is an indicator that the variance of the estimator is generally lower than that of the Horvitz-Thompson estimator for the estimates at that level.

31. We first note that the estimated CV's are relatively low for the Horvitz-Thompson estimator at each level. This is to be expected as the sample is composed so heavily of large take-all units, which don't contribute to the variance. For the Tax Data Unavailable/Pseudo-Census scenario we note that the estimated efficiencies are approximately one for Total Expenses and are slightly higher than one for Total Energy Expenses. This can be interpreted to mean that the relationship between Total Expenses and the auxiliary variable is strong enough that the extensive imputation yields variability that is similar to the sampling variability in the Horvitz-Thompson estimator. On the other hand, we notice a slight loss in efficiency for the estimates for Total Energy Expenses (the linear relationship with the auxiliary variable is not as strong). For the Tax Data Available/Pseudo-Census scenario we notice an important efficiency gain for Total Expenses. The CVs of this estimator are notably decreased from those of the Horvitz-Thompson estimator. The results for Total Energy Expenses show a loss of efficiency compared to the Horvitz-Thompson estimator, particularly at the "All manufacturing" level.

VI. CONCLUSIONS

32. This paper has given some background on the T1 and T2 tax programs at Statistics Canada and the use of tax data by the Annual Survey of Manufactures. It has also outlined and presented a comparison of three methodologies that could be used to produce population level estimates from the ASM. These methodologies vary in terms of the amount of imputation that is involved and the extent to which they make use of auxiliary data that is available. According to our results, the three estimators yield similar point estimates. We have demonstrated that the use of the Horvitz-Thompson estimator could produce estimates with a smaller CV, but there would be many challenges associated with producing a pseudo-census dataset that yields the same population level estimates. This pseudo-census dataset is required in order to provide the analytic capabilities that data users need. Thus, in order to make efficient use of the data (i.e. select a methodology that yields a small CV), the imputation approach that makes use of tax data is the preferred option. The availability of the tax data allows us to produce estimates with a CV that is lower than we could otherwise achieve through imputation based on auxiliary information such as annual revenues. In particular, to produce estimates with the same level of quality, we would need a larger sample size if we did not have the tax data available.

A. Future Work

33. Future work on this project will be concentrated in three main areas. First of all, it may be possible to expand the direct use of the tax data to more variables by applying some editing to the tax data to improve its quality. This may be possible through fully automated edits, or through editing systems that require subject-matter input, but it is suspected for some variables that by removing outlying observations, the data may be more suitable for use in the survey programs.

34. Another area that may be explored is the use of tax data in less direct ways. For example, the direct use of the tax data (as described in this paper for the ASM) relies on a relatively solid linear relationship between the tax data and the survey data, with no intercept term. While this is reasonable to expect, what we have observed in our analysis is that this is not always the case for a variety of reasons. It may also be possible to reap the benefits of the available tax data through modeling of the other variables. For example, models with an intercept, or multiple regression models may be fitted to try to find models that would lead to efficiency gains for the variables that are not currently tax replaced.

35. Finally, quality indicators that reflect the variance associated with imputation need to be developed. Given the large fraction of records that are completed via imputation, it is important to select a suitable quality indicator; one which conveys the increased variance and the potential for bias. This challenge is not specific to the Annual Survey of Manufactures and will be explored in the near future.

References

Batten, D. and Matthews, S. (2003). Direct Use of Tax Data for the Annual Survey of Manufactures, Variable Selection Study, Internal Statistics Canada Document.

Hamel, N. and Belcher, R. (2002). Edit and Imputation of the General Index of Financial Information, Contributed Paper to the Work Session on Statistical Data Editing, Helsinki, May, 2002.

Hutchinson, P., Jocelyn, W. and Cooray, L. (2004). Design of the T1 TEP/UES Tax Sample TY 2003, Internal Statistics Canada Document.

Lavallée, P. and Hidiroglou, M. (1988). On the Stratification of Skewed Populations. *Survey Methodology*, **14**, 33-43.

Philips, R. (2003). The Theory and Applications of the Score Function for Determining the Priority of Followup in the Annual Survey of Manufactures, *Proceedings of the Survey Methods Section*, Statistical Society of Canada.

Shao, J. and Steel, P. (1999), Variance Estimation for Survey Data with Composite Imputation and Non-Negligible Sampling Fractions, *Journal of the American Statistical Association*, **94**, 254-265.
