EXPERIENCES IN THE CODING AND SAMPLING OF ADMINISTRATIVE DATA

Michael Colledge, Victor Estevao and Pierre Foy, Statistics Canada

ABSTRACT

Administrative data are a primary source for the construction and maintenance of frames for economic surveys at Statistics Canada. In general, these data are not sufficient to enable full precision industrial classification of all frame units. Given that accurate classification leads to efficient sample designs, the problem is to determine the target precision of industrial coding. The resources required to produce more detailed codes than the administrative data alone provide can, alternatively, be used to select and process larger samples.

The paper addresses this problem in relation to sub-annual surveys, where payroll deduction data provide frames of small units, and annual surveys, for which income tax data are used both as a frame source and to replace direct data collection. The paper presents measurements of the accuracy of administrative data as a source for coding, the precision of industrial coding supported by these data, the quality of the codes assigned manually and by automated routine, and the rates of change of codes. It describes approaches which have been considered in balancing the allocation of resources between classification and sample size.

1. INTRODUCTION

1.1 Problem Statement

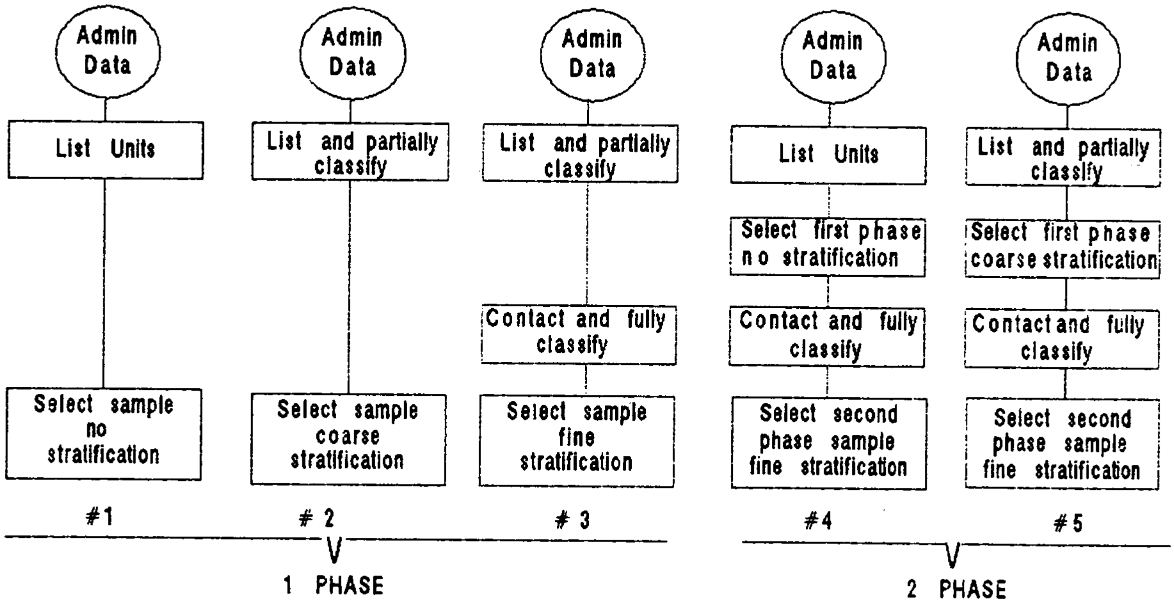
The problem concerns the allocation of resources to the industrial classification of frames for repeated economic surveys. It is described in the context of economic surveys at Statistics Canada but is likely to be relevant to similar survey programs at other agencies.

For survey publication purposes, breakdown of data by industry, by geography and, less frequently , by size is usually required. Also, for efficiency of sampling, classification of units on the survey frame by industry, geography and size is desirable. For many important economic surveys the target population is comprised of a small number of large units and a large number of small, individually insignificant units. To economize on resources, the small units on a survey frame are often defined and classified using administrative data. However, these data are collected to suit the corresponding administrative processes. They usually include sufficient information for precise classification of units by geography and by size, but not necessarily by industry. As industrial classification by industry is desirable for sampling, and essential for publication, this poses the problem: what resources should be expended in further classifying the frame prior to sampling?

Five possible solutions to this problem are illustrated in Figure 1. With option # 1 no attempt is made to classify units on the frame by industry. Classification costs are minimal as only sampled units are classified. However, because stratification by industry is impossible, the sample has to be large and the sampling costs are correspondingly large. Under option # 2 classification data are available from the administrative source are used for stratification, hence, somewhat increasing classification costs and reducing sample sizes. Under option # 3 all units are fully classified prior to sampling, by direct contact of the unit wherever the administrative data are not sufficient. High classification costs are incurred but sample sizes and costs are correspondingly minimal.

Options # 4, 5 based on two-phase designs pro-

Figure 1.--Design Options


Admin Data
List Units
Select sample no stratification
#1
Admin Data
List and partially classify
Select sample coarse stratification
# 2
Admin Data
List and partially classify
Contact and fully classify
Select sample fine stratification
# 3
1 PHASE
Admin Data
List Units
Select first phase no stratification
Contact and fully classify
Select second phase sample fine stratification
#4
Admin Data
List and partially classify
Select first phase coarse stratification
Contact and fully classify
Select second phase sample fine stratification
#5
2 PHASE

vide a compromise between the extreme options # 1, 3. Option # 4 is an extension of # 1 in which a first phase sample is selected without stratification and the sampled units are then fully classified to provide the frame for the (second phase) survey sample. Likewise, option # 5 is an extension of # 3.

Conceptually, selection of the most appropriate option can be considered in terms of the choice of optimum sampling parameters for a two-phase design. Unfortunately the established optimization procedure(e.g., as given by Cochran (1977, chapter 12)) applies to a situation in which the goal is to minimize the overall variance of an estimate, at a given cost, using stratification to improve sampling efficiency. In the context of economic surveys under discussion here, the objective is somewhat different, namely to produce estimates of uniform minimum coefficient of variation within each industry, at a given cost. There is no readily available theoretical solution.

The factors which have to be taken into account in determining the most appropriate option are:

(a) the precision of classification which is available from the administrative source, i.e., the number of significant digits of the Standard Industrial Classification (SIC, Statistics Canada, 1980), which can be assigned;

(b) the classification response error rate, i.e., the errors which respondents to the administrative process make in completing the industrial descriptive items on the administrative forms which are used for classification;

(c) the classification coding error rate, i.e., the errors in assigning classification codes to industrial descriptions;

(d) the costs of classification, using administrative data, and by direct contact;

(e) the rate of change of classification over time, i.e., the propensity of units to move from one industry to another; this determines the frequency with which classification codes have to be validated and, hence, the maintenance costs;

(f) the potential gains in efficiency resulting from classification, i.e. the reductions in sample sizes and costs, which can result from stratification by industry; and

(g) the constraints on sample design imposed by the need for high (even 100%) sampling rates in certain industries with small numbers of units, in order to get samples of adequate size.

1.2 Content of Paper

This paper addresses some aspects of the problem in the context of two particular administrative data sources used for frame maintenance, namely payroll deduction data and income tax data.

Section 2 contains a brief overview of the program of economic surveys at Statistics Canada, its objectives and its use of administrative data. Sections 3 and 5 summarize the results of a number of studies concerning payroll deduction data and income tax data, respectively. The topics covered are, first, the inherent accuracy of administrative data as a source for assigning industrial classification codes; secondly, the precision of classification codes available from data; and, thirdly, the achieved quality of coding, both manual and automated. Given these measures of precision and quality of administrative data, Sections 4 and 6 describe some approaches for determining the appropriate allocation of resources between classification and sampling, and for improving classification procedures. The concluding remarks in Section 7 are followed by the list of references.

2. GENERAL BACKGROUND

The program of economic surveys at Statistics Canada is presently subject to extensive review and overhaul as part of the Business Survey Redesign Project (Cain et al., 1984). The objectives of the project, succintly expressed, are:

(a) to rationalize, integrate and streamline the procedures for provision of frame data and for acquisition, sampling and use of income tax data;

(b) to develop and utilize generic systems and procedures for aspects of the survey process;

(c) to review and redevelop all economic surveys in accordance with these new concepts and procedures.

At the core of the project strategy is the notion of a central frame data base divided into two parts, the so-called "integrated" and the "non-integrated" portions. The integrated portion will contain all large units. These units will be maintained using the full range of administrative sources and special purpose contacts to establish economic structures and related frame data. The units will be fully classified, linked to the various sources, and tracked through time.

The non-integrated portion will provide coverage of the balance of the target population. It will be based on two, unlinked administrative sources-- payroll deduction data and income tax data--both from Revenue Canada. It will thus provide two alternative frames of small units. The frames for sub-annual surveys will be derived from the payroll deduction data, and the frames for annual surveys from income tax data. In general, annual surveys will collect detailed structural information and will make maximal use of income tax data in order to reduce respondent burden, whereas, sub-annual surveys will focus on measures of change rather than level and on timeliness of the survey operations rather than costly collection of detailed items. Colledge (1987) provides more details.

In the context of this paper, the most significant features of the project strategy are, first, that frames of small units will be derived from administrative sources, and, secondly, that the publication requirements for annual and sub-annual surveys will require industrial classification at 4-digit, and at 2- or 3-digit levels, respectively, of the 1980 SIC.

3. PAYROLL DEDUCTION DATA: EVALUATION OF INDUSTRIAL CLASSFICATION

3.1 Source of Data

Payroll Deduction (PD) data is supplied by Revenue Canada in two forms:

(a) a machine readable file with identification information on all PD account holders, but no data on the nature of the holders' activities;

(b) a PD-20 questionnaire sent to each PD account applicant requesting various basic items of information, including business activity.

The administrative unit is the PD account holder. There are about 200,000 new accounts registered every year and the total number of active accounts at any given point in time is about 800,000. The classification of each account is based on PD-20 data.

3.2 Studies of Data Quality

The following results refer to the qualities of PD-20 data and processing procedures as the basis for assigning standard industrial classification codes. The results have been extracted from three studies which had broader objectives, and from a fourth study which focussed directly on this topic.

The Statistics Canada Business Register has been a major vehicle for provision of frame data to economic surveys for the last fifteen years. The Business Register Master File has been systematically created from information reported on the PD-20 questionnaires. The industrial classification of small units having a single PD account is based in most cases upon PD-20 data received when these accounts were first opened, and, to a lesser extent, upon subsequent updates, e.g., feedback from surveys and special purpose frame data enquiries. Thus, the quality of industrial classification of these units provides a general indication of the utility of PD-20 data as a source for SIC coding.

A study by Estevao et al. (1983) found that 21% of the units were incorrectly coded at the 3-digit level of the 1970 SIC (Statistics Canada,1970) and 11 % of the units were incorrectly coded at an industry group level. A repeat study by Estevao and Tremblay (1985) two years later indicated there had been little change. About 21% units were incorrectly coded at 3-digit level and 12% were incorrectly coded at the industry group level.

A third study by Beckstead et al.(1985) measured the time lag from the opening of a PD account with Revenue Canada to the creation of a corresponding record on the Business Register Master File available for survey purposes. The median value was 6 months, of which 2 and 4 months, respectively, could be attributed to waiting for, and processing, the PD-20 questionnaire.

These studies give a general impression of the quality which results from using PD-20 data as the main source for industrial coding. They do not, however, make any distinction between codes assigned from PD-20 data and codes assigned from other sources. More importantly, they do not break down the coding errors into groups corresponding to initial errors in coding, and to genuine changes in industrial activity over time.

A fourth study by Estevao and Tremblay (1986a) focussed specifically on PD-20 data and procedures for industrial coding. The study was conducted on a sample of 800 selected from the 134,000 PD-20 questionnaires processed during a twelve month period from May 1984 to April 1985 inclusive. The principal findings were as follows.

For about 30% of the questionnaires, the PD account holder failed to provide a description of the business activity. For the remaining records, in terms of the level of industrial precision supported by the PD-20 data. 93% of the units could be coded to major industry (2-digit) level of the 1980 SIC; 88% could be assigned a 3-digit code.

In terms of the accuracy of the PD-20 data, the respondent error rates were 4% and 7%, respectively, at 2-and 3-digit levels.

In terms of the quality of the clerical coding procedures, three clerks had error rates of 13%, 18% and 19%, for an overall average of 17%. In defining the clerical coding error rate, the assignment of a code of more precision than the data could genuinely support was included as an error. Thus, the traditional practice of making an intelligent guess at the last one or two digits of the code, in the face of uncertainty, was penalized.

The quality of automated coding was also assessed using the same data. It was found that 51% of the units were not assigned a code by the automated routine. Of the remaining 49%, 20% were mis-coded, compared with a 10% clerical rate for same records. Subsequent to the study, improved performances have been achieved with enhanced routines.

A count was made of PD accounts for which the account holders had remained unchanged and in active econonomic production, yet had altered their industrial activities. It was found that, over a period of 8-19 months, less than 1% of these PD account holders had changed industrial classification.

In summary it was concluded that the assignment of SIC codes from PD-20 data involves a response error at major industry group level of 4%, a clerical coding error of about 17%, and a change of activity of less than 1% per annum among the accounts which remain active with no change of PD ownership. Hence the SIC error rate of about 20% on the Business Register is substantially due to errors in initial coding, and to a lesser extent, errors in response, rather than to changes of activity.

4. PAYROLL DEDUCTION DATA: USE FOR SUB-ANNUAL SURVEYS

4.1 Introductory Remark

As previously noted, PD data are a basic source in the construction of frames for sub-annual surveys.In this context, the derivation and reliability of industrial classification based on PD-20 data is a major concern. The results of the studies summarized in Section 3 were instrumental in formulating the following revised procedures for processing and using PD-20 data.

4.2 Industrial Classification: Initial Assignment

The PD-20 questionnaire will continue to provide the basic source of information for initial assignment of industrial codes for sub-annual survey frame units.

Specification of the requirements for industrial classification are being made more precise. In particular, it is being taken into account that sub-annual surveys require only 3-digit precision of the 1980 SIC.

Where a precise (3-digit) code is not available from the PD-20 questionnaire, clerks will check with the PD account holder. They will not guess. Quality control procedures will be more stringent.

Automated routines are being improved as regards both percentage of records coded and error rates. In view of their potential for encapsulating classification intelligence, for standardizing code assignments, and for reducing clerical work loads, considerably increased use of automated routines is envisaged. In the face of uncertainty, clerical intervention will be involved.

Processing will be faster as a result of minimizing classification requirements, extensive use of automated coding and improved procedures for questionnaire handling and follow-up of non-respondents.

4.3 Industrial Classification: Review

As noted in Section 3, the rate of change of industrial activity of PD account holders is very low. However, bearing in mind that changes in account holder do occur, and that the initial classification may be inaccurate even after the introduction of new procedures, review of classification is required. The following procedures are being implemented.

Negotiations are taking place with Revenue Canada regarding the possibility of sending a PD-20 questionnaire to a business whenever there appears to be a change of account holder.

A program of periodic review on a rotational basis will be introduced along the lines of the US Bureau of Labor Statistics "Refiling Program" (Hostetter 1983).

The review program will be blended with an "initial contact" policy whereby all new entrants to sub-annual survey samples will be contacted to explain the purpose of the survey and to validate the classification information. New entrants may be new PD account holders or existing holders who have entered the sample under the sub-annual sample rotation policy. In the event of classification errors being detected, appropriate measures have to be taken, when updating the frame, to avoid rendering the sample unrepresentative for future survey occasions.

4.4 Classification and Sample Design Options

Of the five options for classification and sample design outlined in Section 1, the following three were considered in some detail.

(a) Use the classification data available from PD-20 questionnaires for stratification of the survey samples. Collect the more detailed industrial classification required from sample units as a survey data item, and produce domain estimates. The advantage of this option is that classification costs are minimal. The overwhelming disadvantage is that survey samples have to be much larger, in view of the large number of unclassified PD accounts.

(b) Supplement the PD-20 data by research into other data sources and by contact of the PD account holders, as necessary, to obtain and maintain classification of the required precision for all units on the frame. Under this option classification costs are high but sample sizes are minimum. This is the approach currently used.

(c) Adopt a two-phase sampling option as used, for example, in the Business Survey Program at the U.S. Bureau of the Census (Konschnik et al., 1985). Under this approach a first phase sample of PD accounts would be selected for classification. This sample would then serve as the frame for (second phase) sampling by the sub-annual surveys.

Two phase sampling in this context has been studied by Foy and Corriveau (1986). It is a compromise between the other two options, and, in fact, includes them both on special cases. Its merit is the reduction of the costs of classification assignment and maintenance, as only a sample of PD accounts are classified. The disadvantages are threefold. First, there is a requirement for larger second phase sample sizes to offset the increase in variance resulting from the use of a first phase sample instead of a fully coded universe. Secondly, there is the somewhat increased complexity of the estimation procedures. Thirdly, there is the difficulty of defining a first phase sample which is large enough to meet the needs of all the sub-annual surveys, yet small enough to produce a net saving in resources.

No objective criteria for an optimal solution could be found, and the choice between options was thus made on a pragmatic basis. The first option above was dismissed because several sub-annual surveys are industry-specific and it would be inefficient for each of them to sample unclassified units. As regards the choice between the remaining two options it was ultimately decided that the potential savings in classification costs were not sufficiently large to justify a change from the current, single phase approach, bearing in mind all the attendant redevelopment costs and risks.

5. INCOME TAX DATA: EVALUATION OF INDUSTRIAL CLASSIFICATION

5.1 Source of Data

There are essentially two types of income tax returns in Canada, filed by individuals and by corporations, on T1 and T2 forms, respectively. For economic statistics, all T2 returns, and those T1 returns indicating self-employed income from an economic activity, are of interest. In total there are about 1,000,000 such returns and an annual turnover rate of 20% or more.

The Assessing Unit at Revenue Canada captures data items relevant to tax collection from every return and makes them available to Statistics Canada in machine readable form. These items do not include, however, all the information relating to economic activity which is available on the T1 and T2 forms and on the financial statements accompanying them. In particular, the business descriptive information and other items, such as expenses which might indicate the nature of the economic activity, are not captured. As noted in Section 2, Statistics Canada makes use of income tax data to replace direct data collection by annual surveys. Thus, the agency is involved in sampling the T1 and T2 returns and capturing the required additional data items.

For tabulation purposes and for efficiency of sampling, classification by size, province and industry is required. Size and provincial classifications can be based on data items available in machine readable form, whereas industrial classification must be obtained by examination of the original tax return or facsimile. Revenue Canada is also interested in industrial classification for audit and for tax modelling purposes.

5.2 Studies of Data Quality

Estevao and Tremblay (1986b) conducted a study of the data and procedures used in assigning industrial classification to returns from T2 tax filers. The target population for the study was a subset of all active, incorporated tax filers with economic activity in 1984. Filers reporting less than $25,000 gross revenue were excluded as they are not in scope for the annual survey program. Filers reporting multiple economic activities, as indicated by separate sets of financial statements, or reporting assets of $10 million or more, were excluded as being special cases. From this target population, a sample of about 37,000 tax returns had previously been selected as part of the tax data acquisition program. A sub-sample of about 700 returns was drawn from this sample and an analysis was carried out similar to that for the study of PD-20 questionnaires described in section 3.2.

The principal findings can be summarized as follows:

(a) in terms of the level of industrial classification supported by the data, 98% and 74% could be coded to 3- and 4-digit levels, respectively, of the 1980 SIC;

(b) as regards the quality of the reported T2 data, the response error rates were 6% and 14% at 3- and 4-digit levels, respectively, of the 1980 SIC; and

(c) as regards the quality of clerical coding, the error rates by three clerks were 14%, 21% and 24%, giving an average error rate of 20%.

For the same sample, in cases where a full precision 4-digit 1980 SIC code could not be assigned based on tax return data, Estevao and Surman (1987b) investigated the costs and benefits of supplementary procedures involving the use of information such as trade indices, telephone and city directories, etc. It was found that, of the 27% of tax returns for which the tax data was inadequate, about 40% could be assigned a complete 4-digit classification, at an average time expenditure of 10 minutes per return. The accuracy of these assignments was not objectively measured but was felt to be no worse than those full precision assignments made on the basis of tax data alone.

A similar study was carried out on T1 tax returns by Estevao and Surman (1987a). The target population was all active, unincorporated tax filers reporting gross business income of $10,000 or more. Certain categories of tax filers, such as limited partnerships, were excluded from the study. From this target population, a sample of about 64,000 tax returns had previously been selected as part of the tax acquisition program. A sub-sample of about 700 returns was drawn. The principal findings are as follows:

(a) in terms of the level of industrial classification supported by the data, 92% and 50% could be coded to 3- and 4-digit levels, respectively, of the 1980 SIC;

(b) as regards the quality of the reported T2 data, the response error rates were 20% and 22% at 3- and 4-digit levels, respectively, of the 1980 SIC;

(c) as regards the quality of clerical coding, the error rates by three clerks were 21%, 27% and 18% giving an average error rate of 25%.

These results indicate that T2 returns provide a better basis for detailed and reliable coding than T1 returns. The difference between the two may be explicable, however, not as a fundamental difference in the ways in which T1 and T2 returns are designed and completed, but rather in terms of the different distribution of T1 and T2 returns by size of business. This hypothesis has not yet been put to the test.

6. INCOME TAX DATA: USAGE FOR ANNUAL SURVEYS

6.1 Introductory Remarks

As previously noted, income tax data are used to provide both frames and data for the small units in scope for annual surveys. In this context, an important issue is the derivation and reliability of industrial classification based on tax returns. (The coincidence, or lack of it, between tax and survey concepts and data item values is a related issue, but is not discussed here). The formulation of revised procedures for processing and using income tax data for classification and sampling purposes has been along the same lines as for the PD-20 data, though the recommendations themselves are somewhat different.

6.2 Industrial Classification Revised Procedures

Quality control of classification procedures will be more stringent. In the face of uncertainty, clerks will undertake research into other sources and, if this fails to produce a classification of adequate precision, will initiate direct contact of the tax filer. There will be no guessing.

Automated and computer-assisted classification routines will be introduced.

A systematic scheme will be introduced for reviewing classification values on a rotational basis. It will be blended with the information feedback from annual surveys.

6.3 Classification and Sample Design Options

Since annual survey data are to be published at 4-digit 1980 SIC precision, sampling at this level would be very efficient. However, as reported in Section 5.2, income tax data alone do not support classification at the 4-digit level, though they do allow assignment of a 2-digit classification for 98% and 92% of the T2 and T1 returns, respectively. In these circumstances the design options considered in some detail were:

(a) sample without classification;

(b) classify all tax returns at 2 digit level, stratify and sample;

(c) draw a first phase sample without classification, and classify the sampled units at the 2-digit level as the basis for drawing a second base sample;

(d) classify all tax returns at the 2-digit level, stratify and draw a first phase sample, then fully classify the sampled units as the basis for drawing a second phase sample.

The first two approaches are single phase designs which differ in the precision of the classification assigned to the universe of tax filers. The second two approaches are two-phase designs which also differ in the precision of classification.

As in the case of PD data (Section 4), no objective criteria for choosing between options have been established, and the problem is further complicated by the possibilities of data and cost sharing with Revenue Canada. The pragmatic solution adopted is to build a parameterized 2-phase sampling system which can provide samples under any one of the above approaches. The optimum sample sizes will be determined empirically. Foy (1987) provides more details.

7. CONCLUSION

Investigations have confirmed that neither payroll deduction nor income tax sources contain sufficient information to support industrial classification at the level of precision needed for survey estimates. The required precision can be derived and maintained only by direct contact of a substantial proportion of the units. Both sources generate high volumes of new units each year, thus the costs of fully classifying the universes are high. This leads to consideration of two-phase designs in which the first phase sample only is fully classified. Such a design has been adapted for use with income tax data.

A second important finding is that, in the context of industrial coding based on administrative sources, stringent quality control procedures are necessary. In particular, there is a tendency to assign classification codes which are more detailed than the available data can support.

Finally, automated coding routinely shows considerable potential for standardizing code assignments and for reducing clerical resources.

Work planned for the future includes: more precise assessment of the rates of change of industrial classification over time; development of theoretical criteria for establishing the optimal allocation of resources between classification and sampling; enhancement of automated and computer assisted industrial classification software; and introduction of classification review sharing with Revenue Canada.

REFERENCES

BECKSTEAD, D. et al. (1985), "PD Processing Time Lag Study," Business Register Working Paper, Statistics Canada, Ottawa.

CAIN, J. et al., (1984), "Infrastructure Development Objectives, Policy and Strategy," Business Survey Redesign Project Working Paper, Statistics Canada, Ottawa.

COCHRAN, W.G. (1977), Sampling Techniques, Wiley, New York.

COLLEDGE, M.J. (1987), "The Business Survey Redesign Project. Implementation of a New Strategy at Statistics Canada," Proceedings of the Third Annual Research Conference, March 1987, U.S. Bureau of the Census, Baltimore.

ESTEVAO, V., AMBROISE, P., and COLLEDGE, M., (1983), "A Study on the Quality of Certain Fields of the Business Register Master File," Business Register Working Paper, July 1983, Statistics Canada, Ottawa.

ESTEVAO, V., and SURMAN, P., (1987a), "An Evaluation of the Assignment of Standard Industrial Codes from T1 Tax Data," Business Survey Redesign Project Working Paper, January 1987, Statistics Canada, Ottawa.

ESTEVAO, V., and SURMAN, P., (1987b), "A Study on the Use of Research Information to Obtain Complete SIC Codes for Incorporated Businesses," Business Survey Redesign Project Working Paper, March 1987, Statistics Canada, Ottawa.

ESTEVAO, V., and TREMBLAY, J., (1985), "A Report on the Quality of the Data in the BRMF - SARUS Study" - 1984/85.

ESTEVAO, V., and TREMBLAY, J., (1986a), "An Evaluation of the Assignment of Standard Industrial Codes from PD-20 Data," Business Survey Redesign Project Working Paper, May 1986, Statistics Canada, Ottawa.

ESTEVAO, V. and TREMBLAY, J., (1986b), "An Evaluation of the Assignment of Standard Industrial Codes from T2 Tax Data," Business Survey Redesign Project Working Paper, November 1986, Statistics Canada, Ottawa.

FOY, P., (1987), "Two-Phase Sample Design for Estimation from Tax Data for Annual Surveys of Economic Production," Business Survey Redesign Working Paper, September 1987, Statistics Canada, Ottawa.

FOY, P., and CORRIVEAU, P., (1986), "Evaluation préliminaire de l'emploi d'un échantillon maître des comptes PD dans la partie non-intégrée du CFDB," Business Survey Redesign Project Working Paper, March 1986, Statistics Canada, Ottawa.

HOSTETTER, S.C., (1983) "The Verification Method on a Solution to the Industry Coding Problem," Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C.

KONSCHNIK, C., MONSOUR, N., DETLEFSEN, R., (1985), "Constructing and Maintaining Frames and Samples for Business Surveys," Proceedings of the Section on Survey Research Methods, American Statistical Association, Washington, D.C.

STATISTICS CANADA, (1970), "Standard Industrial Classification 1970,"Catalogue 12-501E, Statistics Canada, Ottawa.

STATISTICS CANADA, (1980), "Standard Industrial Classification 1980," Catalogue 12-501E, Statistics Canada, Ottawa.