

11-617

no.89-08

c.2

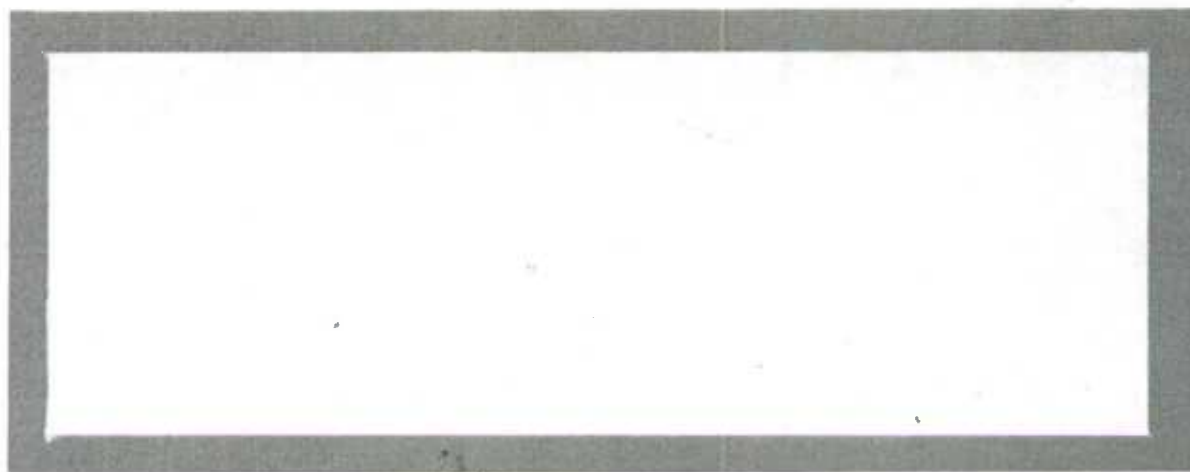
Statistics
Canada

Statistique
Canada

STATISTICS STATISTIQUE
CANADA CANADA

1997

LIBRARY
BIBLIOTHÈQUE



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes
entreprises

Canada

WORKING PAPER NO. BSMD-89-008E

CAHIER DE TRAVAIL NO. BSMD-89-008E

METHODOLOGY BRANCH

DIRECTION DE LA MÉTHODOLOGIE

REDESIGN OF THE SURVEY OF THE CLOTHING INDUSTRIES
1987 SURVEY REFERENCE YEAR

by

Douglas Yeo
May 29, 1989

REDESIGN OF THE SURVEY OF THE CLOTHING INDUSTRIES

1987 SURVEY REFERENCE YEAR

Douglas Yeo,
Business Survey Methods
August 29, 1988
Updated May 29, 1989

TABLE OF CONTENTS

	Page
A. Introduction	1
B. Background	1
C. Methodology	2
D. Recommendations for Future Improvements	3
E. Details of the Redesign	4
E.1 Preliminary Studies	4
E.2 Decisions	5
E.3 Frame Creation	6
E.4 Stratification	7
E.5 Sample Size Determination and Allocation	8
E.6 Sample Selection	8
E.7 Final Frame Creation	9
E.8 Sample Data Base Creation	9
E.9 Estimation	10
Appendix I: Definitions	11
Appendix II: Processing of the Annual Survey of Manufactures - QUIPS and CESE	12
Appendix III: Selection of the Take-All Establishments - Hidiroglou's Method	13
Appendix IV: Stratification of the Take-Some Establishments - the Cumulative Square Root Technique	14
Appendix V: Sample Size Determination and Allocation	15
Appendix VI: Estimation	17

REDESIGN OF THE SURVEY OF THE CLOTHING INDUSTRIES 1987 SURVEY REFERENCE YEAR

A. Introduction

This paper describes the redesign of the Survey of the Clothing Industries (SOCI) that was introduced in the 1987 reference year. This annual sample survey has been conducted by Statistics Canada for several years, on a cost-recovery basis for the Textile and Clothing Board of Canada. It is an adjunct to the Annual Survey of Manufactures, providing advance estimates for clothing commodities.

Sections B, C, and D provide an overview of the redesign in terms of background information, a summary of the methodology, and recommendations for future improvements. Section E gives a detailed review and description of each aspect of the redesign. The appendices include definitions and acronyms, and further subject matter and methodological detail.

B. Background

The Annual Survey of Manufactures covers all establishments in the manufacturing sector. Estimates of commodity inputs and shipments, as well as various financial variables, are produced annually, about 18 months after the end of the survey reference year. The Textile and Clothing Board of Canada desires earlier annual estimates of domestic shipments for certain clothing commodities. A sample of clothing manufacturers is collected and processed, producing estimates roughly 6 months after the end of the survey reference year. These estimates of Canadian clothing production aid the Board in recommending policies on the importation of foreign clothing goods. The Board needs estimates of commodity outputs, in terms of the number of items shipped (shipment quantities) and, to a lesser extent, the value of items shipped (shipment values).

SOCI needed to be redesigned for four reasons. First, the sample was last updated in 1984 and no longer accurately represented the frame. Second, small establishments were not being asked for shipment quantities; rather, these data were being imputed from shipment values. Third, little documentation existed, and the Board wanted a fully documented system. Fourth, the Business Survey Redesign Project entailed changes to the methodology of the entire Annual Survey of Manufactures, and thus to SOCI.

The Board desired a new sampling methodology and estimation procedure. A sample size of under 500 establishments was specified. Since large establishments cost twice as much to process as small ones, a target of about 200 large establishments was set. The estimates for each commodity group were to have coefficients of variation (CVs) under 2.5%.

C. Methodology

Before 1987, SOCI used combined ratio estimation. Each establishment in the frame was classified as a take-all, take-some, or birth establishment. Post-stratification was used to separate establishments by production or non-production of particular commodities. After shipment quantities were imputed from shipment values for small establishments, total quantities and values shipped were estimated using domain estimation. Estimates were produced at the 2-digit commodity level, and pro-rated to the 3-digit level using previous-year commodity distributions.

For the 1987 survey it was decided that a special questionnaire would be designed for small establishments. Space was provided for the respondent to report up to 4 commodities, as well as one "all other commodities shipped", since studies showed that very few small establishments produced more than 4 different types of products. A quantity question was added to each commodity line.

Stratified random sample was used, with each establishment first being placed in a stratum based on its "dominant commodity", the commodity with the largest shipment value. Then, the establishments within each dominant commodity stratum were assigned to sub-strata, according to size by shipment value. Typically, each dominant commodity stratum was divided into a take-all substratum and 1 to 3 take-some sub-strata. Strata with very few establishments were designated as entirely take-all. Each take-all - take-some boundary was determined using Hidioglou's Method. To reduce the number of large establishments sampled, for cost reasons, relatively small take-all sub-strata were chosen by using a CV of 10% in the algorithm. The take-some sub-strata boundaries were chosen using the cumulative square root technique.

Births were handled differently than in previous years. To reflect the fact that in the future the Central Frame Data Base will not have access to current-year births, they were not included in the birth frame. Thus, births were defined as previous-year births not processed at the time the master frame was built. Since no commodity information was available for these establishments, they were stratified by industry.

For each dominant commodity stratum with only 1 take-some sub-stratum, the sample size was calculated under simple random sampling. Similarly, the sample size and allocation were determined under stratified random sampling for dominant commodity strata with 2 or 3 take-some sub-strata. A cost-ratio of 2:1 for large versus small establishments was included in the sample allocation formula. The final overall CV was specified as 3% for estimates of shipment values. Since stratification was done by shipment value, it was recognized that the CVs for estimates of shipment quantities would be higher than those for estimates of shipment values. The CVs calculated at this stage were those that would apply when using a simple expansion estimator. However, since a combined ratio estimator was to be used, it was expected that the actual CVs would be lower.

A uniform random number generator was used to initially select the sample. Manual intervention ensured 3 things: that no cell (a particular dominant commodity stratum and size sub-stratum) was undersampled compared to the

originally calculated sample allocation, that every cell had at least 6 sampled establishments, and that every cell had a sampling fraction no less than 10%. The birth sample was chosen by random ordering and systematic selection.

Subject matter officers handled all aspects of data collection from questionnaire design to imputation. Since stratification was done on only one variable, post-stratification and domain estimation were no longer required. Any commodity not reported by an establishment was treated merely as a total commodity shipment of zero. Also, since shipment quantities were asked directly of the respondent, it was no longer necessary to impute them.

Due to many frame changes after mail-out, the frame needed to be rebuilt and reweighted. These frame changes were significant, and reduced the efficiency of the allocation. It is important to note that these changes were largely deaths and births, discovered long after mailout, and would have necessitated frame rebuilding no matter what sampling and estimation scheme was used.

Estimation for births used simple expansion, while a form of ratio estimation was used for all other establishments. Studies had shown that ratio estimation would result in smaller variance than would simple expansion. The separate ratio estimator was rejected in favour of a combined ratio estimator because it is prone to serious bias if sample sizes are small, or if the CV of the auxiliary variable is large. These two problems seemed significant because of the problems involved in building an accurate frame, and because of reductions in data quality resulting from extensive imputation of frame and sample data. As well, ratios did not differ greatly between size sub-strata, and thus it was recommended to use the combined ratio estimator. A combined ratio was computed across size sub-strata for each dominant commodity stratum. Previous-year shipment values and quantities were chosen as auxiliary variables, since they had the highest correlations with the main variables.

Final estimates were produced, for shipment values and quantities, by adding birth and take-some estimates to the take-all totals. These estimates, and their CVs, were produced at the 2-digit commodity level. The Textile and Clothing Board obtained 3-digit commodity estimates by pro-rating to 3-digit commodity proportions from the most recent previous-year data available.

D. Recommendations for Future Improvements

Several aspects of the survey warrant methodological analysis, with a view to further improvements. Three recommendations are presented.

Recommendation 1: Use shipment quantity as the principal stratifying variable, instead of shipment value. Since the client is most interested in estimates of shipment quantities, and not shipment values, quantity data should be used as the stratifying variable. 1987 survey data should be examined to see if sufficient quantity data are available. If this change can be implemented, data quality should improve, and CVs should decrease.

The change would necessitate a new definition of dominant commodity, and would require some minor program changes.

Recommendation 2: Simplify and improve the computer programs for sampling. The new stratification, allocation, and sample selection system is clumsy, does not produce an optimal allocation, and requires manual intervention. The new Lavallée-Hidiroglou stratification method should be tested to see if the system would be simplified, and if the overall sample size would decrease. It is possible that the cost of this particular work would be paid for several times over by sample size reductions. At the same time, 1987 sample sizes and final CVs should be examined to see if specific commodity groups had unusually high variance and would benefit from a larger sample. The sample selection program could also be simplified and improved, by looking at the various existing packages available. This change would require only minor reprogramming, and would be able to handle all establishments, including births.

Recommendation 3: Re-evaluate the estimation methods, given the change to Harmonized System coding. The Annual Survey of Manufactures is switching to the Harmonized System for commodity coding in 1988. If data quality is reduced significantly, especially year-to-year correlations, it may be necessary to use a different estimator, such as simple expansion. Analysis should be done, if a concordance between new and old coding lists can be produced in time for methodological studies to be completed.

E. Details of the Redesign

E.1 Preliminary Studies

The first step in the analysis of the population of establishments involved the production of frequency tables. It was found that fewer than 8% of establishments had fiscal years ending in February or March. This was true for establishments in both the Integrated Portion (IP) and Non-integrated Portion 1 (NIPl) of the Central Frame Data Base. The breakdown of establishments that had changed from the Full Scale Questionnaire (FSQ) to the Other Characteristics Questionnaire (OCQ) or vice versa was also tabulated. Subject matter officers expressed concerns that these establishments would have more than 4 commodities to report, while there were only 4 write-in lines allowed for this purpose on the OCQ. Analysis showed that 68% of these establishments had reported 4 or fewer commodities in the past (by Industrial Commodity Code, or ICC), and that the 4 largest commodities accounted for 94% of the total value of shipments. As well, reported commodities constituted 4 or fewer Import Control Groups (ICGs) for 98% of these establishments.

Since each establishment could have up to 15 ICGs, it appeared that the survey would require a complex sampling methodology, using 15 stratifying variables. To save a great deal of time and to reduce complexity, the "dominant ICG" approach was taken. For each establishment, the ICG with the greatest aggregate shipment value was calculated. A cross-tabulation of ICG and this dominant ICG showed that on average about 80% of the total shipment value for a particular ICG falls within its dominant ICG. In

other words, most of the shipment value of each ICG falls in the category of "primary product". Only a small portion of each ICG is produced as a secondary product or byproduct. The relationship was very strong for every ICG except 41, a very small ICG that later became entirely take-all, and 99, a catch-all ICG. Because of this strong relationship, a univariate approach, based on dominant ICG, was taken for sample size determination, allocation, and selection.

Another cross-tabulation, this time of Standard Industrial Classification (SIC) and dominant ICG, revealed a strong relationship for most 4-digit SICs. This information was used to sample births, where SIC is the best stratifying variable available.

E.2 Decisions

The Annual Survey of Manufactures sends an FSQ to all IP establishments, and an OCQ to a NIP1 sample. For SOCI, the project team decided to mail OCQs to all NIP1 clothing manufacturers. A sample would then be processed in advance of the rest of the Annual Survey of Manufactures to provide the estimates. NIP2 would be disregarded, and would not be included in the survey frame. The Textile and Clothing Board specified a CV of 2.5% for each dominant ICG stratum, and an overall sample size not greater than 500 establishments, of which not more than 250 could be in IP.

A special OCQ would be designed, with 4 write-in commodity lines and 1 "all other" line. Since shipment quantity is the key variable required by the Board, a quantity question would be added for each commodity line. "Value of shipments" would also be asked by commodity. Although the collection of financial data in the NIP portion contravenes Business Survey Redesign Project guidelines, and will require a special exemption in future years, it was considered desirable since the respondent would typically use shipment values to calculate quantities of commodities shipped. Inclusion of a shipment values question would also improve the editing of shipment quantity data and increase data quality.

It was decided that the mail-out would include establishments with late fiscal year-ends. However, NIP1 establishments with fiscal year-ends in February or March would be followed up only once or twice. FSQ mail-out would take place in early January, followed by OCQ mail-out at the end of January.

A number of establishments ended up with ICG 99, the catch-all ICG, as their dominant ICG. However, subject matter officers found the problem of recoding these establishments to more precise ICGs intractable. Thus it was decided to retain ICG 99 as a legitimate dominant ICG for sample selection and estimation, as had been done in the past.

E.3 Frame Creation

To accommodate the upcoming Central Frame Data Base environment, the frame was based on the previous-year universe. Current-year births were not included in the frame to reflect the fact that in the future the Central Frame Data Base will not have access to these births. Previous-year survey files were produced, by combining the principal statistics and commodity files, reduced as follows:

- clothing manufacturers, found in SIC Major Group 24,
- records not in error, and
- in-scope SICs, made up of all SICs in Major Group 24 except contractors, the glove industry, and the fur good industry.

The data were further reduced and transformed using the following steps:

- Step 1. Keep only IP and NIP1 establishments;
- Step 2. Keep only those questionnaire lines that contain commodity data.
- Step 3. Transform ICCs to ICGs, by matching to a concordance list. This transformation also caused all out-of-scope ICCs to be dropped;
- Step 4. Aggregate shipment values to ICG totals within each establishment;
- Step 5. Calculate the dominant ICG for each establishment by ordering by aggregated shipment value and taking the largest value;
- Step 6. Add all shipment values, that is, shipment values for all in-scope ICC lines, for an establishment to the dominant ICG. At this point the file contained 1 line per establishment;
- Step 7. Keep all establishments with non-zero total shipment value.

Data for year Y-2 were reduced separately, but in a nearly identical manner. Step 2 was slightly different: since CESE had been run for year Y-2 estimated commodity data was available. Therefore these data were used, rather than the raw data.

Previous-year data were not complete at the time the frame was being built, and would not be for a number of months afterwards. Just slightly more than half the establishments, or 549, had previous-year commodity data at this point, necessitating the use of the commodity data from year Y-2 for the remainder of the establishments. The reduced previous-year file created by step 7 was compared with the list of all clothing manufacturers. Any establishments without commodity data were matched to the reduced year Y-2 file, once again after step 7, thus obtaining the commodity data from year Y-2. These data were, in effect, imputed for 468 establishments. The resulting 1017 establishments made up the "regular" frame, excluding any previous-year births without commodity data, which could not, of course, be imputed from year Y-2.

The birth frame included all previous-year establishments with no commodity data and a record serial number, the establishment identifier, starting with 56 or 86. In all, 147 establishments made up the birth frame.

E.4 Stratification

From this point, all computation was done using SAS (Statistical Analysis System) on the microcomputer, reducing turn-around time, computer time, and especially costs.

The population was stratified to help obtain a reasonable sample size with minimum variance. Each establishment was first placed in a stratum based on its "dominant commodity", the commodity with the largest shipment value. Ideally, quantity data should be used to determine the dominant commodity, but were not available for all establishments. Then, the establishments within each dominant commodity stratum were assigned to sub-strata, according to size by shipment value. Typically, each dominant commodity stratum was divided into a take-all substratum and 1 to 3 take-some sub-strata. However, dominant ICG strata 41, 43, and 44 were so small that they were designated as entirely take-all.

Hidiroglou's Method was used to determine the take-all/take-some boundaries. A CV of 10% was used in the algorithm to control the number of IP take-all establishments. (See Appendix III for a detailed look at Hidiroglou's Method, and the reasons behind the use of a 10% CV at this first stage of stratification.) At this time, subject matter officers needed a list of take-all establishments for contact purposes, in preparation for the FSQ mail-out. Copies of the list of 163 IP take-all establishments were provided, on paper and on a file uploaded to the mainframe from the microcomputer and transferred to an operating system disk file.

Initially, the take-some establishments of each dominant ICG stratum were stratified into 2 and 3 sub-strata using the cumulative square root technique. (See Appendix IV for a description.) Tabulations were produced, both unweighted and weighted by shipment value, to aid in determining the ideal number of take-some sub-strata for each dominant ICG stratum. At this point it was decided that:

- dominant ICG strata 45, 48, 49, and 99 would have 1 take-some sub-stratum,
- dominant ICG strata 32, 40, 46, 47, and 50 would have 2 take-some sub-strata, and
- dominant ICG strata 37, 39, and 42 would have 3 take-some sub-strata.

Each establishment was placed in its proper cell. All establishments with complex structures - known as combined reporters and artificial splits - were automatically placed in the take-all sub-stratum. Tabulations of the final breakdown of the frame by cell were produced.

Due to the strong SIC - dominant ICG relationship, the birth population was stratified by SIC. The 3-digit level was chosen because some 4-digit strata had as few as 1 or 2 establishments.

E.5 Sample Size Determination and Allocation

(For further details, see Appendix V.) Establishments in take-all sub-strata were subject to 100% sampling, as their name implies. For those dominant ICG strata with only 1 take-some sub-stratum, the sample size was determined under simple random sampling. The sample size and allocation were determined under stratified random sampling for dominant ICG strata with 2 or 3 take-some sub-strata. A cost-ratio of 2:1 for FSQs versus OCQs was included in the allocation formulae. The first sample size was determined using a CV of 2.5%. When the overall sample size proved to be a little high, totalling 458 establishments, a new sample size of 411 was determined using a CV of 3%. Note that this CV applied only if simple expansion was to be used later on. However, the estimator of choice was a combined ratio estimator, which would, it was hoped, further reduce the final CV.

For births, a sampling fraction of 1/3 determined the sample size. The sample was proportionally allocated across 3-digit SICs.

E.6 Sample Selection

A sample of size 410 across all cells was chosen by cell using a uniform random number generator and a CV of 3%. Each establishment was assigned a random number between 0 and 1. The establishment was added to the sample only if the random number was less than the sampling fraction calculated in the original allocation. 25 more establishments were randomly added to the sample to satisfy the following constraints:

- maximum weight of 10 in any cell,
- minimum sample size of 6 in any cell, and
- sample size in any cell no less than that calculated in the original allocation. This constraint was necessary because the sample selection mechanism undersampled in some cells.

Births were ordered by 3-digit SIC and the last 2 digits of the record serial number. Then a systematic sample was selected. An integer between 1 and 3 was selected randomly as the starting point of the sampling process. The selected establishment and every third one thereafter were selected in the sample. 47 of the 147 births were so chosen, resulting in a total sample size of $435 + 47 = 482$, of which 222 were IP. After final consultations with subject matter officers, a sample of 217 IP, 215 NIP1, and 44 birth establishments was taken, for a total of 476.

It is recommended that this systematic sampling method be used for all establishments, and not only the births. It produces a good random sample, with much better control over the sample size than with the uniform random number generator. A program with a random number generator could be used if it can provide strict sample size control.

E.7 Final Frame Creation

Once the preliminary file from the Annual Survey of Manufactures was available for the previous year, it was possible to build the final frame. This file was reduced on the mainframe by keeping only good lines from IP and NIP1 establishments. Once this reduced file was downloaded to the microcomputer, ICC lines were matched to ICGs. Thus only in-scope ICC lines were kept. Some manual data corrections were made at this point. Any shipment quantities of zero were manually imputed by subject matter officers.

The final frame was then built, with one data line per establishment, by reconciling this new frame with the original frame. The original ICG file, with one data line per ICG, was kept for use in combined ratio estimation. Mismatches - establishments on the new frame but not on the old one, or vice versa - were printed out for analysis by subject matter officers. The number of deaths, 46, and late additions, 68, found in this manner was significant, and reduced the efficiency of the allocation. Higher CVs may have resulted from this considerable frame change. The deaths were considered to be frame changes, since they were found outside the sampling process, and were removed from both the frame and the sample.

Shipment quantities were converted to units, from dozens, for example, and a unit price check was done to find further inconsistencies. Analysis by subject matter officers uncovered several errors. The number of establishments that had moved to a different stratum was also checked, and found to be relatively small. After all corrections were made, there were 1039 establishments in the final frame, excluding births.

The final birth frame was created in much the same manner. 36 establishments made up the birth frame, an extreme drop from the 147 establishments on the original birth frame. Although this drop was not too surprising, because "births" often turn out to be merely out-of-scope, it caused a great reduction in the efficiency of the birth sample, since sampling and mail-out had been completed several months earlier.

E.8 Sample Data Base Creation

The SOCI sample data file was reduced, on the mainframe, by keeping only good lines. Once this reduced file was downloaded to the microcomputer, ICC lines were matched to ICGs, keeping only in-scope ICC lines.

The data were checked for accuracy and some manual corrections were made. After shipment quantities were converted to units, a unit price check was performed. The final sample was checked against the original sample and against the final frame. All discrepancies were reported to subject matter officers. Deaths were given a value of zero for shipment values and quantities, but were kept in the frame. After all corrections were made, there were 428 establishments in the regular sample and 14 in the birth sample.

E.9 Estimation

(For further details, see Appendix VI.) Much analysis was done to determine the best estimator. It was determined that a ratio estimator would have a smaller variance than would simple expansion. Correlation studies also showed that shipment values and quantities had the highest correlations with previous-year shipment values and quantities, respectively. Thus previous-year shipment values and quantities were chosen as auxiliary variables.

Simulation studies were also done, to confirm the best type of estimation. The combined ratio estimator, using previous-year shipment values and quantities as auxiliary variables, was eventually chosen as the best estimator. The separate ratio estimator was considered, but dropped for several reasons.

At this point, all that remained was the creation of all input files for the estimation program. An establishment by ICG matrix was created for shipment values and quantities. Shipment value totals for any ICGs not reported by an establishment are not considered missing values but true shipment value totals of zero. Thus these ICGs were given a value of zero, a step necessary for the calculation of estimates and variance. The take-some and birth establishments were extracted from the original survey data matrix. Both a take-some matrix and final cell totals were calculated using previous-year data, to be used by the combined ratio estimator. Survey take-all totals were calculated, to be used in the creation of the final estimates.

Now that the input files were ready, estimates could be calculated for the birth stratum using simple expansion. The strata were ignored, due to their small size. Then, after reweighting the establishments to take into account the changed frame and sample sizes, combined ratio estimation or simple expansion was used to provide estimates for the take-some sub-strata. Simple expansion was used in those cells with no previous-year data. These estimates and their calculated variances were added to the birth estimates and take-all totals to produce overall estimates, at the 2-digit ICG level. 3-digit ICG estimates were obtained by the Textile and Clothing Board by using year Y-2 3-digit ICG proportions within each 2-digit ICG. Proportions from year Y-2 were used because previous-year final figures were not yet complete.

Some adjustments were made to the data, especially in swimwear, and then the final estimates were produced. Shipment values showed an overall increase of about 5%, while shipment quantities increased by over 6%. CVs for shipment values were quite reasonable for most ICGs. As expected, shipment quantity CVs were higher, and were very high in ICGs 42 and 47. ICG 99 had very high CVs, due mainly to the inherent variation found in a "catch-all" category. In future years, an increase in the minimum sampling fraction, from 10% to 15%, should reduce the CVs in the more variable cells. This increase would raise the sample size by only 10 to 20 establishments. ICGs 32 and 37 may need an even greater increase.

APPENDIX I: DEFINITIONS

Births: establishments new to the frame; as opposed to deaths.

Business Survey Redesign Project: An extensive project involving the creation of, and the redesign of business surveys to operate using, the Central Frame Data Base.

Central Frame Data Base: new master frame for the Business Survey Redesign Project.

CESE: Commodity Estimation for Small Establishments; program used to estimate commodity data for small establishments in the Annual Survey of Manufactures; see Appendix II for a complete description.

CV: Coefficient of Variation; $CV = \text{Standard Deviation (X)} / X$.

Dominant ICG: The major ICG for an establishment; the ICG with the greatest aggregate shipment value.

FSQ: Full-Scale Questionnaire; the detailed questionnaire sent to IP establishments; includes financial information; replaces the Annual Survey of Manufactures "long form"; as opposed to the OCQ.

ICC: Industrial Commodity Code.

ICG: Import Control Group; clothing commodity breakdowns used by the Textile and Clothing Board; groups of ICCs.

IP: Integrated Portion of the Central Frame Data Base; large, important establishments and multi-establishments; fully profiled and linked; surveyed by FSQ; as opposed to the NIP.

NIP: Non-Integrated Portion of the Central Frame Data Base; smaller, generally single establishments; not linked, not fully profiled; surveyed by OCQ and the use of tax data; as opposed to the IP. NIP1 establishments are smaller than those in IP but > 250 k; NIP2 establishments are < 250 K.

OCQ: Other Characteristics Questionnaire; the short questionnaire sent to NIP1 establishments; cannot request financial information; replaces the Annual Survey of Manufactures "short form"; as opposed to the FSQ.

SIC: Standard Industrial Classification; for example, SIC Major Group 24 is the clothing industry.

SOCI: The Survey of the Clothing Industries.

Subject Matter: Industry Division and the Textile and Clothing Board.

Take-all: establishments subject to 100% sampling; may include large, extremely important establishments, or all those in a very small stratum;

Take-some: establishments subject to a sampling rate of less than 100%.

APPENDIX II: PROCESSING OF THE ANNUAL SURVEY OF MANUFACTURES -
QUIPS AND CESE

QUIPS stands for the Questionnaire Information Processing System. It is the generalized edit system used to process the Annual Survey of Manufactures. The data file it produces contains all survey and tax data collected for establishments in the Annual Survey of Manufactures. It also contains estimated data from the Short Form Estimation System, CESE, and data that have been estimated by other, usually manual, means. The QUIPS Data File has a variable record length with 1 establishment being represented on only 1 record. The Standard Data File, much simpler to use because of its fixed record length, can be derived from it. A QUIPS Data File exists on disk and tape for the current Census year and the previous year, Y-1, and on tape only for all data prior to Y-1.

QUIPS is used to edit data on both the OCQ, mainly form numbers 31 and 3011, and the FSQ. Line numbers on the FSQ and OCQ always correspond to specific questions. For example, line 1.9 is the reporting year on both the FSQ and the OCQ. The FSQ surveys commodities on lines 8.1 (8.1.1 - 8.1.99), 8.6 and 9.1 - 9.6; commodities are surveyed on line 13 (13.1 - 13.4) on the OCQ.

The majority of NIP establishments are surveyed using tax data. However, tax records contain only financial data; commodity information is absent. To ensure commodity data are obtained for all records, an OCQ is sent to each NIP1 establishment on a 3- or 4-year cycle. The Rotation Program is run every year to determine which NIP1 establishments need to be mailed an OCQ in order to obtain fresh commodity data. For all other NIP1 establishments, and all NIP2 establishments tax records are used to obtain financial data, and CESE is run to estimate commodity data. CESE estimates shipment quantities, acquires historical commodity information for tax-return and non-response establishments, and generally improves data quality.

CESE creates a new line number for each NIP establishment: line 20.1 for all NIP establishments surveyed by OCQ, and line 21.1 for all NIP establishments using only tax records. Lines 20.1 and 21.1 do not appear on the OCQ; in this sense they are "imaginary". The 2 lines are created in the Historical Acquisition Module. Commodity data are transferred from line 13 directly to line 20.1 for all OCQ respondents. For OCQ non-respondents, CESE acquires their historical commodity data, transforms it, and then enters it in line 20.1. Tax record establishments have commodity data entered in line 21.1 in a similar manner.

CESE is run at a very late stage of processing, generally after an industry has been closed. When the 1987 sample was being selected, a new version of CESE was being tested on 1985 data. 1986 data were not yet completely processed, and CESE had not run. Thus 1986 commodity data were read from line 13, reducing slightly the total number of establishments with commodity data (since no tax record establishments had had commodity data acquired historically yet) but not otherwise affecting the quality of the data.

APPENDIX III: SELECTION OF THE TAKE-ALL ESTABLISHMENTS - HIDIROGLOU'S METHOD

"It is desirable to stratify highly skewed populations on the basis of the size of the units...put a certain number of large units into a take-all stratum and sample those with certainty...There are several advantages in stratifying a highly skewed population for the given method. For a fixed coefficient of variation, the overall sample size associated with this procedure will invariably be lower than the sample size associated with no stratification...confidence intervals are essentially based on populations that are less skewed...this type of stratification guards against overestimation of population characteristics when highly skewed distributions are sampled."

Hidiroglou's Method determines cutoff rules for stratifying a population into a take-all and take-some universe. Approximate cutoff rules can be calculated in terms of the required CV for the overall sample in an iterative fashion. The first approximation is:

$$\text{lim1} = \mu + \sqrt{(c^2 * Y^2 / N + S^2)},$$

where μ = the population mean,
 c = the desired coefficient of variation,
 Y = total shipment value
 N = the number of units in the population, and
 S = the population standard deviation.

Subsequent approximations are:

$$\text{lim}(j+1) = \mu + \sqrt{((N-t-1) * c^2 * Y^2 / (N-t)^2 + S^2)},$$

where μ = the mean of all units below the j th cutoff,
 t = the number of take-all units in the j th approximation and
 S = the standard deviation of all units below the j th cutoff.

The cutoffs tended to stabilize after 3 or 4 iterations. To be safe, 8 iterations were performed in the computer program. The desired CV for the entire sample was 2.5%, but calculating the cutoffs using this level would have resulted in very large take-all sub-strata. Since no more than 200 IP establishments were wanted in the sample, due to cost constraints, a CV of 10% was used for this first stage of stratification, resulting in reasonably-sized take-all sub-strata with 163 IP establishments. Later stages of stratification achieved an expected CV of approximately 3%.

Hidiroglou's Method is designed for 1 take-all and 1 take-some sub-stratum, drawing a simple random sample from the take-some sub-stratum. However, the cutoff values obtained remain fairly good with a stratified sample of the take-some units, consisting of 2 or 3 sub-strata. As well, Hidiroglou's Method also deals with the simplest case, involving simple expansion. If a ratio estimator is used, the CV should be further improved.

Hidiroglou, M.A.(1986), "The Construction of a Self Representing Stratum of Large Units in Survey Design," The American Statistician, 40, 1, 27-31.

APPENDIX IV: STRATIFICATION OF THE TAKE-SOME ESTABLISHMENTS -
THE CUMULATIVE SQUARE ROOT TECHNIQUE

Over half of the dominant ICG strata contained enough take-some establishments to warrant further stratification. They were broken down into both 2 and 3 sub-strata, and then a decision was made on the best number of sub-strata for each dominant ICG stratum.

The take-some portion of each dominant ICG stratum was stratified again using the cumulative square root technique, also known as Dalenius' Rule. All take-some establishments were first ordered by shipment value within each dominant ICG stratum. Within each dominant ICG stratum the cumulative total of the square root of the shipment value was tabulated.

To obtain 2 take-some sub-strata, the cutoff was placed at $1/2$ the cumulative total. Likewise, the cutoffs were placed at $1/3$ and $2/3$ the cumulative total to obtain 3 take-some sub-strata. Thus $1/2$ or $1/3$ of the cumulative square root of the shipment value lay in each sub-stratum after stratification.

Based on the size of each dominant ICG stratum, and the results of the 2-and 3- sub-strata breakdowns, the best number of take-some sub-strata for each dominant ICG stratum was determined.

APPENDIX V: SAMPLE SIZE DETERMINATION AND ALLOCATION

1 Take-some Sub-Stratum

When only 1 take-some sub-stratum existed, the sample size was determined under simple random sampling as follows:

$$\text{VarX} = (N^2 S^2 / n) (1 - f) = (N^2 S^2 / n) (1 - n/N) = N^2 S^2 / n - NS^2.$$

Thus $\text{VarX} + NS^2 = N^2 S^2 / n,$

and $n = N^2 S^2 / (\text{VarX} + NS^2) = N^2 S^2 / (C^2 X^2 + NS^2).$

VarX is the variance of X, total previous-year shipment values, used to estimate Y, total current-year shipment values. 4 dominant ICG strata had only 1 take-some sub-stratum, and used this method for sample size determination.

C is the coefficient of variation, that is, $C = \text{VarX} / X$. Note that C is the overall CV, as compared to the CV of the take-some sample. The relationship between the two is as follows:

$$C^2 X^2 = \text{Var}_{\text{overall}} = \text{Var}_{\text{TA}} + \text{Var}_{\text{TS}} = 0 + \text{Var}_{\text{TS}} \text{ and}$$

$$C^2 (X_{\text{TA}} + X_{\text{TS}})^2 = \text{Var}_{\text{TS}} = C_{\text{TS}}^2 X_{\text{TS}}^2.$$

Thus $C(X_{\text{TA}} + X_{\text{TS}}) = C_{\text{TS}} X_{\text{TS}},$

$$C[(k)X + (1-k)X] = C_{\text{TS}}(1-k)X,$$

$$CX = C_{\text{TS}}(1-k)X.$$

Thus $C = C_{\text{TS}}(1-k), \text{ or}$

$$C_{\text{TS}} = C/(1-k).$$

where k = proportion of take-alls in the population, by shipment value.

For example, if 50% of the total shipment value is take-all, then

$$C = 2.5\% \text{ and}$$

$$C_{\text{TS}} = C/(1-k) = 2.5\%/(1-0.5) = 2.5\%/0.5 = 5\%.$$

The presence of a take-all sub-stratum allows the CV of the take-some sample to be much higher than the overall CV. This is one of the purposes of the take-all sub-stratum. Thus it is possible to reduce the sampling fraction in the take-some sub-stratum.

2 or 3 Take-some Sub-Strata

When more than 1 take-some sub-stratum existed, the sample size and allocation were determined under stratified random sampling as follows:

$$n = \frac{[\sum_h W_h S_h \sqrt{c_h}] [\sum_h W_h S_h / \sqrt{c_h}]}{\text{VarXbar} + (1/N) (\sum_h W_h S_h^2)}$$

where Σ_h = Summation operator, summing over all sub-strata h ,
 W_h = N_h/N , the weight,
 VarXbar = $C^2 X_{\text{bar}}^2 = c^2 X^2 / N^2$, and
 c_h = the average cost per establishment in sub-stratum h ,
 calculated using the cost-ratio of 2:1 for FSQ:OCQ, and
 calculating the proportion of IP:NIP1 in each substratum.
 $= (2*IP + NIP1) / \text{TOTAL}$
 $= (2*IP + \text{TOTAL} - IP) / \text{TOTAL}$
 $= (IP + \text{TOTAL}) / \text{TOTAL}$
 $= IP/\text{TOTAL} + 1,$

TOTAL is the total number of establishments in sub-stratum h , divided into IP and NIP1 establishments.

The allocation is:

$$n_h = \frac{n W_h S_h / \sqrt{c_h}}{\sum_h W_h S_h / \sqrt{c_h}}$$

Note that the formulas are the same for Xbar and X. 8 dominant ICG strata had more than 1 take-some stratum, and used this method for sample size determination and allocation.

APPENDIX VI: ESTIMATION

The ratio estimator is the best if 2 conditions are satisfied (Cochran, page 158):

1. The relation between y_i and x_i is a straight line through the origin,
2. The variance of y_i about this line is proportional to x_i ,

where Y is the variable to be estimated, in this case shipment values or quantities, and X is an auxiliary variable. X should be highly correlated with Y, and is used to provide additional information, thus, hopefully, reducing the variance.

Studies showed that when previous-year shipment values and quantities were used as X, the above 2 conditions were generally true. More specifically, the ratio estimator is more efficient than simple expansion if (Cochran, page 157):

$$\begin{aligned} \text{corr}(X,Y) &> (1/2) (S_x/\bar{X}) / (S_y/\bar{Y}) \\ &= \text{CV}(x_i) / 2 * \text{CV}(y_i). \end{aligned}$$

Once again, analysis showed that this was the case.

For these reasons, a ratio estimator was chosen for use whenever previous-year data were available. The birth stratum was estimated using simple expansion, due to the obvious lack of previous-year data. Therefore:

$$\begin{aligned} \hat{Y}_{\text{birth}} &= N/n \sum_i y_i, \text{ and} \\ v(\hat{Y}_{\text{birth}}) &= N/n (N-n) s_y^2. \end{aligned}$$

There are 2 basic types of ratio estimators. A separate ratio estimate of each sub-stratum total can be made and then these estimates can be added up. Or, a single combined ratio can be computed across sub-strata. It was eventually decided to use the combined ratio estimator, where the take-some sub-strata in each dominant ICG stratum would be combined. Thus, in each dominant ICG stratum:

$$\hat{Y}_{\text{TS(ICG)}} = (\hat{Y}_{\text{st}}/\hat{X}_{\text{st}}) X$$

where X = true total of the auxiliary variable X,
 $\hat{X}_{\text{st}} = \sum_h N_h \bar{x}_h$
 = estimated total of the auxiliary variable X, using the selected sample over h take-some sub-strata, and
 $\hat{Y}_{\text{st}} = \sum_h N_h \bar{y}_h$
 = estimated total of the current-year variable Y.

Therefore, $\hat{Y}_{TS(overall)}$ is the sum of the 15 $\hat{Y}_{TS(ICG)}$, and

$$\hat{Y} = Y_{TA} + \hat{Y}_{TS(overall)} + \hat{Y}_{birth}.$$

The variance for each dominant ICG stratum is as follows:

$$v(\hat{Y}_{TS(ICG)}) = \sum_h N_h^2/n_h (1-f_h) (s_{yh}^2 + \hat{R}^2 s_{xh}^2 - 2\hat{R} s_{yxh}),$$

where $s_{yxh} = \sum_i (y_{hi} - \bar{y}_h)(x_{hi} - \bar{x}_h) / (n_h - 1)$
 = sample covariance between y_i and x_i in sub-stratum h ,
 and $\hat{R} = \hat{Y} / \hat{X}$.

$$\text{Thus } v(\hat{Y}_{TS(ICG)}) = \sum_h N_h/n_h(N_h - n_h) (s_{yh}^2 + (\hat{Y}/\hat{X})^2 s_{xh}^2 - 2(\hat{Y}/\hat{X}) s_{yxh}).$$

Therefore, $v(\hat{Y}_{TS(overall)})$ is the sum of the 15 $v(\hat{Y}_{TS(ICG)})$, and

$$\begin{aligned} v(\hat{Y}) &= v(Y_{TA}) + v(\hat{Y}_{TS(overall)}) + v(\hat{Y}_{birth}) \\ &= v(\hat{Y}_{TS(overall)}) + v(\hat{Y}_{birth}), \end{aligned}$$

since the take-all sub-stratum has no sample variance, by definition.

The calculations for Y and $v(Y)$ were performed 30 times, once for shipment values for each of the 15 ICGs, and again for shipment quantities.

Note: The separate ratio estimator also appeared to be a good estimator. However, it is much more prone to serious bias if sample sizes are small, or if the CV of the auxiliary variable is large. In 1987, with problems involved in building an accurate original frame, with the running of CESE on 1986 data, and with the manual imputation of many shipment quantities, these two problems seemed to be significant. Therefore a large bias in the estimates might have resulted from the use of the separate ratio estimator. As well, the ratios did not differ too much between take-some sub-strata, and thus it was advisable to use the combined ratio estimator for the 1987 survey.

STATISTICS CANADA LIBRARY
BIBLIOTHEQUE STATISTIQUE CANADA



1010230014

#74419
C.2

008