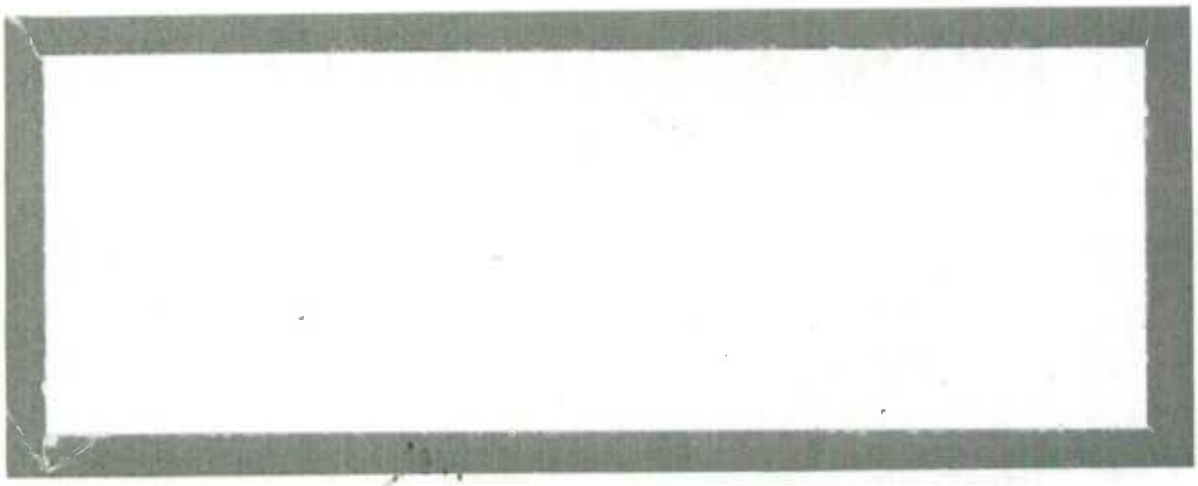
Statistics Canada Statistique Canada



Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes entreprises

WORKING PAPER NO. BSMD-91-006E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-91-006E

DIRECTION DE LA METHODOLOGIE

AN INTEGRATED APPROACH TO DATA EDITING, ERROR CORRECTION AND IMPUTATION

DETAILED REPORT OF THE SIMULATION STUDY

by

Bob Downer and Jean-Marie Berthelot

April 1991

AN INTEGRATED APPROACH TO DATA EDITING, ERROR CORRECTION AND IMPUTATION

- - - - -

DETAILED REPORT OF THE SIMULATION STUDY

Bob Downer
J.-M. Berthelot
BSMD
April 1991

TABLES OF CONTENTS

ABSTRACT

In this study, the use of the generalized data collection and data capture (DC2) function and of the generalized edit and imputation system (GEIS) in a production environment was simulated. Certain concepts were developed independently within the two functions. To investigate the DC2-GEIS compatibility and to further enhance their development, these concepts were applied to data from the Annual Retail Trade Survey.

The main objective of the study was to investigate the use of the following integrated error detection and correction strategy: DC2 identifies suspicious units and follows-up only the most influential of these units to correct and validate data. Subsequently, any remaining problems are corrected automatically within GEIS.

The simulation confirmed the compatibility of the DC2 and GEIS functions and produced promising results. Follow-up of some of the most influential units has a beneficial effect on the estimates. It is also clear that follow-up of all suspicious respondents is not required. It was found that, excluding follow-ups for total non-response, out of scope and suspicious small cell units, a follow-up rate of thirty-three percent of units which fail one or more edits was sufficient. At this rate, the overall simulation estimates were within one percent of the production estimates for major variables. Analysis at the dissemination cell level for these variables revealed that there was very little increase in the precision of the estimate when even more units were followed-up.

This study has been valuable in examining some new ideas. More importantly, it shows that an excessive number of follow-ups leads to unnecessary resource consumption. A limited number of follow-ups of "influential units" is sufficient to ensure acceptable data quality.

SOMMAIRE

Dans cette étude, l'utilisation en production de la fonction générique de collecte et de saisie (DC2) et du système générale de vérification et d'imputation (GEIS) a été simulée. Certains concepts ont été développés indépendamment à l'intérieur des deux fonctions. Pour évaluer la pertinence de ces concepts et la compatibilité entre DC2 et GEIS, ces concepts ont été appliqués aux données de l'Enquête annuelle du commerce de détail.

L'objectif principal de cette étude était d'évaluer l'utilisation de la stratégie mise de l'avant pour la détection et la correction des erreurs. Dans cette stratégie, DC2 identifie des données suspectes et effectue un suivi uniquement auprès de celles qui ont un impact significatif, dans le but de corriger et de valider les données. Par la suite, toute incohérence est corrigée automatiquement par GEIS.

La simulation a confirmé la compatibilité entre DC2 et GEIS et a produit des résultats prometteurs. Le suivi de certaines unités influencielles a un effet bénéfique sur les estimations. La simulation démontre aussi qu'il n'est pas nécessaire de suivre toutes les unités suspectes. Excluant les suivis pour la non-réponse totale, les unités hors enquêtes et les données suspectes dans des petites cellules, un taux de suivi de trente-trois pour-cent a produit, lors de la simulation, des estimations qui sont à moins de 1 pour-cent des estimations obtenues en production pour les variables principales. Des analyses au niveau des cellules de diffusion pour ces variables ont montré qu'il y avait peu d'amélioration de la précision des estimations quand un plus grand nombre d'unités étaient suivies.

Cette étude a permis d'évaluer de nouvelles idées concernant la vérification et la correction des données. Encore plus important, elle a démontré qu'un nombre excessif de suivis entraine une dépense superflux de ressources. Un nombre limité de suivis auprès des "unités influencielles" est suffisant pour assurer un niveau de qualité acceptable.

1 PREFACE

Statistics Canada has undertaken a major endeavor called the Business Survey Redesign Project (BSRP) within which a new Central Frame Database is being designed.

In order for economic surveys to adapt to this new environment and take full advantage of the new facilities available to them through the Business Register, some modifications are required to their production processes.

As part of the Business Redesign Project, there are approximately 200 surveys that may need partial or total redesign. Previously, this would have required large investment of development resources to produce customized survey systems. However, due to recent technological advances, mainly in the area of microcomputers, tools are now available that enable a more general approach to development and production systems.

This gave rise to the creation of the Generalized Survey Function Development (GSFD) team as part of the overall BSRP.

Within the context of the BSRP objectives, the main goal of the GSFD team is to develop generalized tools that will be capable of being adapted easily to the majority of business and social surveys which will undergo redesign in the future. The systems to implement these functions will be based on a limited set of standardized methodological approaches designed to improve timeliness, reduce respondent burden and minimize resources in the production process. In addition, these systems will be flexible enough so that new processing modules can be added as different or new methodological principles are introduced while maintaining or improving data quality. Finally, the systems will be portable across various system architectures and sites contributing to more cost effective methodologies and operations.

It will be possible for surveys undergoing redesign to select appropriate methods, systems, operations and performance measures from available generalized options. These selections will then be assembled into an efficient production process with very little development effort.

To accommodate the development process most effectively, production functions have been grouped into four main modules. Some of the functions that will be developed are:

1. Generalized Sampling Module

   Components within this module include: sample size determination, allocation, initial selection, sample maintenance and estimation.

2. Generalized Data Collection/Capture/Preliminary Edit Module (DC2)

   Components include: standardized specification input, a standard question bank, generation of materials (specifically generation of personalized questionnaires), simultaneous generation of data capture programmes, and efficient operational strategies.

3. Generalized Edit and Imputation Module

   Components within this module include: specification of edits, edit analysis, application of edits, error localization, and various options for imputation.

4. Generalized Tabulation Module

Components within this module will probably be derived from existing software packages.

The use of these generalized modules will lead to a reduction in development time and resource requirements for future redesigns, efficiency gains in production and a reduction in person year utilization in the production process.

2 INTRODUCTION

The collection, correction, capture, and verification of data is a very time consuming and resource intensive activity. Verification and correction of data constitutes a significant percentage of these resources. Hence attempts to minimize these costs without serious impact on the quality of the data are being researched as part of the GSFD project.

In the development of the generalized survey functions, two modules have been designed for verification and correction of the survey data: the data collection and capture (DC2) and the generalized edit and imputation system (GEIS).

2.1 DATA COLLECTION AND DATA CAPTURE (DC2)

The generalized function for collection and capture covers the required steps for collection, conversion (to a computer readable format) and verification of survey data. It intends to ensure:

(i) a minimum validity of collected information (e.g. that proper characters are numeric),

(ii) that the information captured corresponds with that given by the respondent,

(iii) that only those respondents who have a significant impact on the survey estimates are followed-up, and

(iv) that respondent burden is minimized.

2.2 GENERALIZED EDIT AND IMPUTATION SYSTEM (GEIS)

After the data has been processed by DC2, the Generalized Edit and Imputation System identifies and corrects inconsistencies within a questionnaire. GEIS is an automatic system for continuous numeric data which assumes that the most influential suspicious units have already been followed-up and that data for them has been corrected. Most of the remaining problems in the data are then assumed to be of very small impact. These problems may result from respondents being unable or unwilling to supply complete or correct data. GEIS identifies inconsistent data by the application of a set of linear edits to the data. Imputation methods correct the identified inconsistencies thus producing a clean, consistent dataset.

2.3 OBJECTIVE OF THE SIMULATION

The long range objective is an integrated DC2-GEIS function with one correction and verification strategy. At present GEIS is at a fairly advanced stage of development and it was thought that application of the system to economic survey data would be very valuable. A first simulation

involving DC2 and GEIS and the Annual Retail Trade Survey data was run in 1989 and is described in BSMD (1990). To further investigate the compatibility of the two functions and examine some new ideas, a second simulation was run on the same data.

The objective of this report is to describe this second DC2-GEIS simulation. The goals of the study were to verify the pertinence of the concepts developed in DC2, simulate the use of GEIS with business survey data, ensure the compatibility of the two general functions, and evaluate the quality of the data obtained.

With cost as a prime motivating factor, a major emphasis in this simulation run was on investigating the potential gain in reducing the follow-up to a limited number of units. In proceeding towards this goal, much more processing was performed within DC2 and GEIS than in the first run and this work should prove to be rewarding for future reference.

The two subsequent sections give a brief description of the data and methodology used in the simulation. In the fifth section, the general procedure is described. Section 6 presents the overall results and section 7 gives summary conclusions.

3 DATA USED

The Annual Retail Trade Survey (ARTS) is a census of retailers who have Total Net Sales and Receipts greater than or equal to 1 million dollars. The population consists of chains and independents retailers. A chain is generally considered to be a larger retailer with at least four business locations in the same trade group.

Each retailer involved in this census must provide information at two levels. Appropriate questionnaires are completed for the establishment level (Questionnaire A) and individual place of business (Questionnaire B). Although the survey's results are not published, they are used by the System of National Accounts.

ARTS was ideal for the simulation study for the following reasons: it is a fairly large survey with heterogeneous small and large units, has a sufficient number of numerical variables, several members of the study team were familiar with the survey, and previous consultations with subject matter representatives had been very positive.

Only the data collected at the establishment level was used in the study. This portion was chosen because it would allow us to restrict our attention to 12 continuous variables (listed in Appendix 1). The raw data, originally provided by the respondents, was re-captured with the co-operation of Headquarters Operations Division. The study was confined to 2053 'A' questionnaires. These questionnaires were from the food products sector (trade group 20) and the motor vehicle equipment/manufacturing industry (trade group 120) as well as all questionnaires from the provinces of Prince Edward Island and Alberta (province codes 11 and 48 respectively). One hundred and ninety six records were later removed as they were considered to be out of scope (Total Net Sales and Receipts less than one million or Non-Operating Revenue greater than Total Net Sales and Receipts) to leave 1857 records in the study.

As well as the 1987 raw data, the final data files from 1985, 1986 and 1987 were also used. The data from 1985 and 1986 provided parameters for the editing and follow-up strategies of DC2. The final 1986 file was also used for estimator imputation in GEIS. The final 1987 data, as released by subject matter, was used as a control comparison to indicate the effectiveness of the integrated DC2-GEIS approach.

4 METHODOLOGY

In a regular production situation data would be initially processed by DC2 and then passed to GEIS.

The DC2 editing strategy is used to establish edit rules, identify suspicious data and determine which questionnaires must be verified by means of a follow-up. Groups of correlated variables establish the edit rules and the tolerance method determines suspicious units. A list of the DC2 edits can be found in Appendix 2(i).

The DC2 editing strategy consisted of a 2-step process. First, units that needed automated follow-up based on specified criteria were identified. For the remaining units, a score function was used to select "influential" units that require follow-up. Only "influential" units that failed the editing process are subject to follow-up.

After the data is passed from DC2, GEIS uses a set of linear edits to identify any remaining inconsistencies within a record. For any inconsistent record, a minimum set of fields is identified for imputation and these fields are then corrected using an appropriate imputation method. For a more detailed description of the steps in the GEIS processing, see Kovar, MacMillan, and Whitridge (1988). GEIS treatment of the data is not conditional on whether or not correction was performed by DC2, that is, overwriting may take place.

The following sub-sections provide more details on the methodology used for the simulation study.

4.1 DC2 METHODOLOGY

Verification in the generalized data collection and capture system is an integrated process which consists of the following parts:

(i) Data analysis, which forms a natural grouping of related variables for the establishment of edit rules,

(ii) Suspicious units are identified through the use of statistical techniques, and

(iii) Follow-up of units which have a significant impact on the estimates is performed.

Follow-up may occur as a result of satisfying criteria for automatic follow-up or through selection by a score function.

(i) Classification of Variables in DC2

In this initial stage, natural groupings of variables (i.e. those which are correlated) are formed in order to establish a limited set of edit rules. To analyze the variables, frequency tables were considered in order to establish the presence or absence of the variables for the full census population in 1985 and 1986. Next, graphical analysis establishes relationships between the variables (for example, total opening inventory of 1986 versus total closing inventory of 1985 should resemble the function $y = x$). Histograms for the variables verified that the typical distribution of economic variables was present (i.e. asymmetric with a long right-hand tail).

Correlation matrices were then produced. The Pearson coefficient was used because it is easy to calculate and interpret but we note that it is influenced by extreme values. A log transformation was applied to the data so that the working distribution would be approximately normal and thus symmetric. The Pearson coefficient could then be more appropriately used.

The above process determined related variables within a questionnaire for variables of the same year and the method then established links between variables of consecutive years (1985 and 1986 for the study). The combined results lead to a set of edits for pairs of variables. A more detailed description of the editing strategy can be found in Bilocq and Berthelot (1990).

The pairs of variables identified by the grouping method are then processed through the tolerance edit to identify suspicious units. The following section describes how suspicious units are identified.

(ii) Detection of Suspicious Units

The tolerance edit method is a bivariate method which has been effectively used in the detection of suspicious units (Hidiroglou and Berthelot, 1986). The method models links between two variables by analyzing their ratio. The method identifies units whose relationship differs from the corresponding overall trend of other units. A questionnaire is considered to be suspicious if at least one of its variables is suspect. The edit rules were based on result from the procedure described above in section (i). The trend (ratio) can either be defined between variables of the same year or between historical values of the same variable.

Graphical analysis revealed that the distribution of the study's variables was similar between provinces but differed according to trade group. Hence tolerance bounds were established according to trade group for the full population. If there were few non-zero values for a given trade group for a particular variable, a global tolerance boundary at the Canada level was calculated.

These boundaries were then applied to the 1987 records. Each of the variables was implied in more than one edit. The score function, (see section (iii) below) requires a single error flag for a variable. A hierarchy of priorities was developed to produce this single error flag.

First, the edits were applied to the 1987 records. For each edit, a flag was produced which indicated whether the variable passed or failed the edit or if it was not applied. If a historical edit was applied for the variable, the result was retained as the final error flag. This priority was established because historical links had been shown to be superior to links within a record. If a historical edit was not applied then an intra-record edit would be considered. Finally, if no edit could be applied (for example if there was a birth and no value was present in the previous survey) then the variable was deemed to be in error if it was a frequently reported variable. Hence the resulting file had only one error flag (indicating whether suspicious or not) for every variable of each record.

(iii) Description of the Score Function

The purpose of the score function is to select suspicious units which have a significant impact on the estimates. In this study, these units were identified for follow-up in addition to those satisfying automatic follow-up criteria. The highest score function values are considered to be most influential and these units are followed-up. The score function used emphasizes the

A measure of the relative discrepancy explained by follow-up was produced for each of the twelve variables. It gives a measure of the efficiency of following-up a given number of units and is defined as follows:

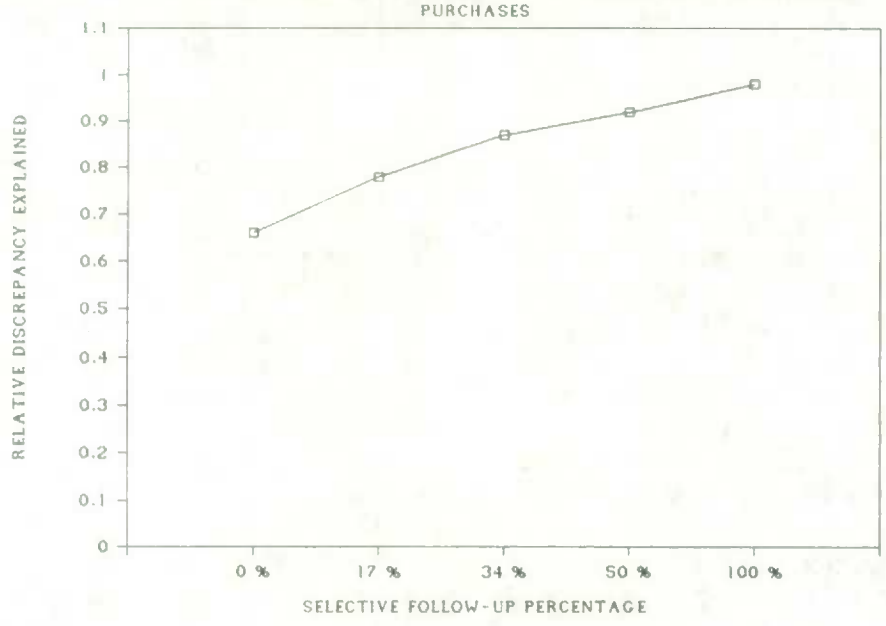
$$Rel\_discp = \frac{\sum_{i=1}^{K} |DG_i - R_i|}{\sum_{i=1}^{E} |F_i - R_i|}$$

where:

DG_i is the value of the variable after the DC2-GEIS simulation
R_i is the reported 1987 value
F_i is the 1987 final production value
K is the number of follow-ups
E is the number of DC2 errors

Figure 3 presents the trend in the relative discrepancy explained over the five files for the variable Purchases. Although the value increases as expected over all five files, we have already explained eighty-nine percent of the discrepancy with thirty-four percent of the selective follow-ups. Other variables showed a similar pattern.

FIGURE 3: RELATIVE DISCREP. EXPLAINED



Finally, we note that a large number of follow ups at the capture stage will lead to more overwrites of DC2 by GEIS. An overwrite is considered to be a change of at least one variable by GEIS for a followed-up unit. If too many follow-ups are performed, the confirmation or correction resulting from a costly follow-up is more likely to be changed by an imputation. This pattern is illustrated in Table 1. With only 84 records overwrites of DC2 in 326 GEIS imputations, the thirty-four percent selective follow-up rate appears to be an appropriate choice from this perspective as well.

TABLE 1:NUMBER OF RE-CONTACTS AND IMPUTATIONS FOR THE 5 FILES

PERCENTAGE SELECTIVE FOLLOW-UP	NUMBER OF FOLLOW-UPS	NUMBER OF IMPUTATIONS	NUMBER OF OVERWRITES OF DC2 BY GEIS
0%	388	407	53
17%	486	360	70
34%	584	326	84
50%	677	291	106
100%	967	207	155 (*)

* In this study, an overwrite occurred when GEIS changed at least one variable for a re-contacted unit. As can be seen above, when 100% of the errors were followed-up, approximately 3 out of every 4 imputations was an overwrite. Further analysis showed that these overwrites generally involved the variables Net Sales, Total Opening Inventory, Total Closing Inventory, Purchases and Salaries. This suggests that the GEIS edits may have been too restrictive for these variables.

Due to the above results and with budgetary constraint as a prime motivating factor, we recommend following-up approximately one-third of the remaining records in error (after automatic follow-ups have been determined). The gain in quality of the estimates in going beyond this point could be an unnecessary expense.

6.2 FURTHER RESULTS

As a result of recommending the 34% selective follow-up rate, the results described below will be for this rate only. Similar results have been obtained for the other four rates and this work is available on request.

For selected variables, Table 2 shows the percentage of records changed by DC2 and GEIS, the simulation and production totals and overall relative pseudo-bias in percent for the file of thirty-four percent of selective follow-ups. For example, DC2 follows-up changed the variable net sales 18.7 percent of the time and GEIS changed it just 4.8 percent of the time giving an over-estimate of 0.28 percent. Slight over-estimates were also obtained for Total Closing Inventory, Total Purchases, and Total Salaries (0.34, 0.33, and 0.78 percent respectively) while Total Net Sales and Receipts was under-estimated by 0.18 percent. Across all variables, there was an even split of over-estimates and under-estimates for this file.

TABLE 2

PERCENTAGE DISCREPANCY ('PSEUDO-BIAS')
34% OF SELECTIVE FOLLOW-UPS
FOR SELECTED VARIABLES

VARIABLE	DC2 CHANGES (% RECORDS)	GEIS CHANGES (% RECORDS)	DC2-GEIS TOTAL (X 1000)	PRODUCTION TOTAL (X 1000)	PSEUDO-BIAS (%)
NET SALES	18.7	4.8	8,530,953	8,515,601	0.28
TOTAL NET SALES AND RECEIPTS	17.3	0.2	8,999,518	9,015,594	-0.18
TOTAL CLOSING INVENTORIES	15.8	3.4	1,480,133	1,475,179	0.34
TOTAL PURCHASES	18.3	8.3	6,870,403	6,847,548	0.33
TOTAL SALARIES	15.3	3.3	944,708	937,354	0.78

Figure 4 below is a frequency histogram of the pseudo-bias at the cell level for the variable Net Sales. The pseudo-bias class with midpoint at zero had the greatest frequency for all variables. This result, a pattern common to all variables, is very encouraging. For Net Sales, we notice that almost all cells have a simulation total within 1 percent of the production total (very little variability in the distribution). The pseudo-bias is actually within 4 percent for all cells. There are slightly more cell over-estimates but this was not a general pattern with other variables. The high frequency for the zero class can be explained by the fact that some of the automatic follow-ups are suspicious units in small cells. For the cell size distribution, see Appendix 4. Since these follow-ups have the reported data replaced by the final released data, they will have a zero pseudo-bias by design.

FIGURE 4: PSEUDO-BIAIS BY CELL

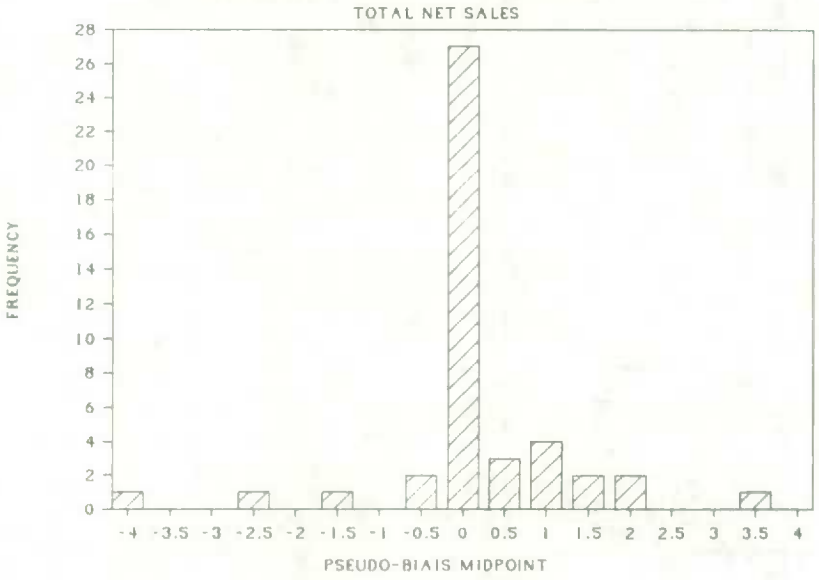


Table 3 which follows gives the overall modification results of the study by variable for the entire sample. The percentage changed in the production environment for 1987 is shown in the second column. It represents a combination of changes made in 1987 at the collection and capture stage by Headquarters Operations Division as well as changes resulting from subject matter analysis of the data. The percentages changed by the simulation, DC2, and by GEIS are also presented. Except for Receipts for Food Services which was reported to be zero 97 percent of the time, the DC2-GEIS simulation changed less records. For example, total purchases was changed 29.7 percent of the time by the 1987 method while it was changed only 25.9 percent of the time by the simulation. For a frequency count of non-zero reported values by variable, see appendix 5.

At the record level, (when we consider a modified record to be one with at least one changed variable), the simulation changed 108 less records (46.5 percent to 40.7 percent). Note that the columns depicting changes by DC2 and GEIS individually may not add exactly to the percentage for the simulation overall because of a small number of records where GEIS over-wrote the value of DC2.

TABLE 3

MODIFICATION RESULTS
(34% OF NON-AUTOMATIC FOLLOW-UPS)
IN PERCENT OF TOTAL RECORDS
BY VARIABLE

VARIABLE	MODIFIED BY 1987 METHOD	MODIFIED BY DC2-GEIS SIMULATION	MODIFIED BY DC2 ONLY	MODIFIED BY GEIS (INCLUDES OVERRIDES OF DC2)
Net Sales	26.6	22.9	18.7	4.8
Gross Commissions	2.0	1.6	1.3	0.3
Receipts from Repairs	11.0	9.1	7.1	2.4
Receipts from Rentals	2.9	2.5	2.2	0.3
Receipts from Food Services	0.5	0.8	0.4	0.3
Other Operating Revenue	4.0	2.9	2.5	0.6
Total Net Sales and Receipts	23.2	17.4	17.3	0.2
Non-Operating Revenue	5.7	5.2	5.1	0.1
Total Opening Inventory	23.6	20.4	17.3	3.4
Total Closing Inventory	19.5	18.2	15.8	2.5
Total Purchases	29.7	25.9	18.3	8.3
Total Salaries	18.9	18.1	15.3	3.3
Record level (at least one variable changed)	46.5	40.7	25.5	18.6

17 were erroneous units in small cells, and 126 potential out of scope records remained in the sample after follow-up was made. Five fractions (zero, seventeen, thirty-four, fifty, and one-hundred percent) of the remaining 579 suspicious units determine the rest of the follow-ups for each of the files. This fraction was approximately the same for all production cells contained in the sample. For example, the file with the fifty percent selective follow-up rate (thirty-six percent overall), the 289 units with the most influence on the estimates were selected for follow-up by the score function and these units had the 1987 reported data replaced by the 1987 final production values.

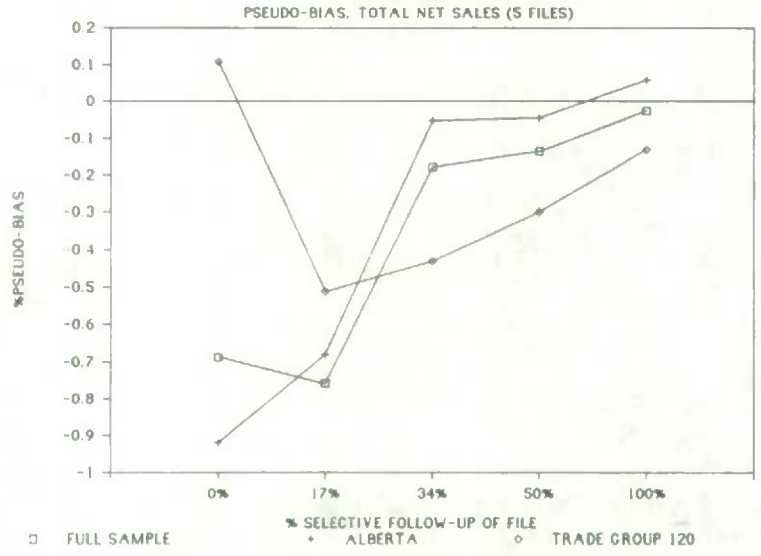
As described earlier, comparison of the results for the five rates was important in discovering the effect on the estimates of following up more records. Our main criterion in the analysis was the comparison of the simulation estimate obtained at a given level of aggregation with that of the 1987 final data. We assumed that the final production value is correct. For a given variable we define:

$$\% \text{ PSEUDO-BIAS} = \frac{\text{DC2-GEIS TOTAL - 1987 PRODUCTION TOTAL}}{\text{1987 PRODUCTION TOTAL}} * 100$$

Hence a negative pseudo-bias indicates an under-estimate of the production total while a positive value indicates an over-estimate.

For all selective follow-up rates, the results are encouraging with pseudo-bias of low magnitude for the frequently reported variables. For example, in Figure 1 we illustrate the overall pseudo-bias at three different levels of aggregation for the variable Total Net Sales and Receipts. Note that the pseudo-bias is less than an absolute value of 1 percent for any rate at any level. Considering the full sample for all files, we obtained a slight under-estimate but this pattern was not common to all variables. Since half of the records in the sample are from Alberta, the results for this province understandably follow the same pattern as the full sample. For trade group 120, we begin with only a 0.1 percent over-estimation for the zero percent file (only automatic follow-up performed) but by chance we under-estimate by 0.5 percent after the first 98 of the selective follow-ups (in the seventeen percent file).

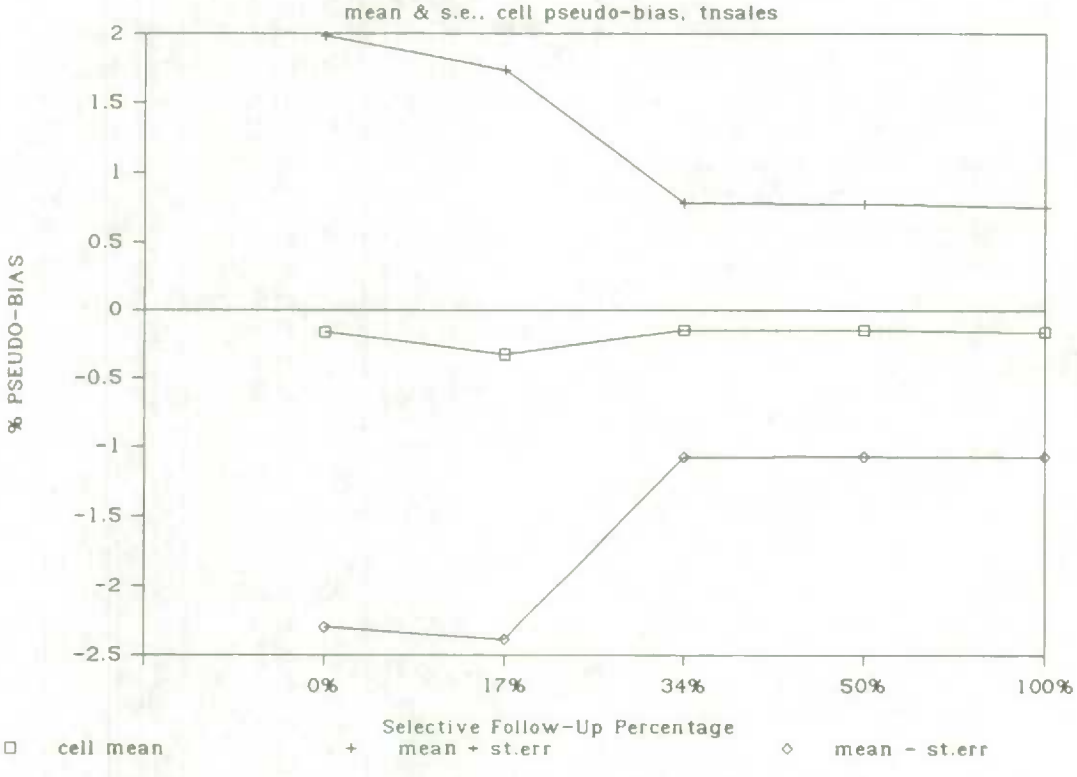
FIGURE 1



The critical result from Figure 1 is the following. The reduction in absolute magnitude of the pseudo-bias in going from seventeen to thirty-four percent of selective follow-ups (486 to 584 records followed-up) is much greater than in going from thirty-four to fifty percent (584 to 678 records). With 584 records followed-up with the thirty-four percent rate, the pseudo-bias for the full sample has already been reduced to -0.178 percent and only improves to -0.136 percent for the fifty percent rate. These 98 follow-ups improved the estimate but not substantially. A similar pattern was observed for other variables. As expected, the pseudo-bias was the lowest when all errors had been followed up (one hundred percent file). The pseudo-bias isn't zero with one hundred percent of the flagged errors followed-up because some additional errors can slip through the edits. Subsequent to data processing, error may have been identified by subject matter analysis. Future budget constraints may prevent this rate of follow-up from being a logical alternative.

Further comparison of the five rates was required at a lower level than the global level described above in order to make appropriate conclusions. As a result, the pseudo-bias was considered at the production cell level (province by trade group). Figure 2 shows the mean and standard error of the pseudo-bias at the cell level, for Total Net Sales and Receipts. Observations here support the conclusions from Figure 1. We see a significant decrease in the standard error when we go from seventeen to thirty-four percent of the selective follow-ups.

Figure 2



The mean cell pseudo-bias is actually closer to zero for the thirty-four percent file than the fifty percent file and the standard error is approximately 0.92 percent for both. No clear reasons have been identified to cause this situation. It is however believed that the way the score function is defined overestimates the impact of partial non-responses and may contribute to this effect. Once again this was a pattern common among the other variables.

could correct all errors of a given questionnaire. It was recognized that this assumption may be somewhat unrealistic but was necessary for the simulation. The 1987 original data of a questionnaire flagged for follow-up was replaced by its 1987 final data.

Potential follow-ups were categorized as either automatic or selective. The automatic follow-ups were all total non-responses, suspicious units in small production cells and out of scope observations. A record was defined to be total non-response if the five most important variables (Total Net Sales, Salaries, Purchases, Net Sales, Total Closing Inventory, and Total Opening Inventory) were all equal to zero. The second criterion for an 'automatic' follow-up occurred if the record was identified as being suspicious by the editing process and the production cell into which it was classified contained less than 10 units. The final criterion for an automatic follow-up occurred when the record was identified as potentially being out of scope.

With these records aside, the remaining units in error may be followed-up according to their score function value. The score function emphasizes the difference between the 1987 reported value and the final value from 1986. A high score function value indicates a high potential impact on the estimates (variable totals) and hence follow-up is desirable. Different upper percentiles of the score function distribution were used as cut-off values to create 5 data files with different selective follow-up rates.

The selective follow-up percentages considered were zero, seventeen, thirty-four, fifty and one-hundred. When these percentages are converted to actual numbers of follow-ups out of 1858 records, they correspond respectively to 388, 486, 584, 677 and 967 (automatic follow-ups included). Zero percent of selective follow-ups was considered to show the results when only the automatic follow-ups were performed. One hundred percent was considered to show the results when all errors detected were followed up while fifty was the half-way point. The seventeen and thirty-four percent rates were chosen as possible alternatives to either zero percent or fifty percent (of the errors that remain after automatic follow-up). The important premise here is that the cost of extra follow-ups may be unnecessary if there is only a small potential gain with respect to accuracy of the estimates.

5.2 GEIS

The five files from DC2 were used as input into GEIS and each file was passed through the components of edit application, error localization, and imputation in exactly the same manner. After the files from DC2 were loaded into separate ORACLE tables, they could be processed by the GEIS system.

Prior to application of the edits, a pre-processor was used to identify the cases in which Total Net Sales and Receipts was reported as positive but its components (Gross Commissions, Other Operating Revenue, Receipts from Food Services, Receipts from Repairs, Receipts from Rentals and Net Sales) were all equal to zero . These variables were set to a missing value to ensure that the error localization module would later identify these fields for imputation. Other such cases may have already been identified and corrected by a DC2 follow-up. The number of records identified by the pre-processor was 29, 23, 18, 12, and 0 respectively for the zero, seventeen, thirty-four, fifty, and one hundred percent selective follow-up rates.

The edits for GEIS were derived from several different sources. Some of the edits were in place for the 1987 survey, others were developed from examining the questionnaire or were formulated in discussions with subject matter personnel. The edits just described formed a set of edits which were common to all records. In addition to this set, pairs of edits were formed

from the ratio edits in DC2. Three edit groups were formed according to trade group. The edits were analyzed to identify inconsistency and redundancy and determine the minimal set of edits for each edit group.

After application of the edits to each of the five files, error localization was then performed to identify fields for imputation. Subject matter personnel were contacted to discuss the reliability of the survey variables. After this consultation, it was decided that a higher weight would be placed on the most reliable variable, Total Net Sales and Receipts. It was given a weight of 3 while the other 11 variables were each given the default weight of 1. With this weight, Total Net Sales would then be less likely to be flagged for imputation and, in fact, it was imputed only twice for each of the 5 files.

Donor imputation was then performed in two stages. Initially, only records with the same standard industrial classification at the 4 digit level (SIC4) were used to impute the recipients at that level. This matching ensured that an imputed record had the same type of retail business. If a donor could not be found at the SIC4 level donor imputation was performed at the trade group level (nature of retail business preserved). In contrast to the first run of the simulation, if a donor was found at the SIC4 level, the imputed record was not available as a donor at the trade group level. Although this situation is fairly unlikely, this concept of imputed values being used as donors at subsequent stages was discussed and considered to be undesirable.

Post-imputation edits in GEIS ensure that the newly imputed records do in fact satisfy the originally established edit rules. The original edits still remain in the post-imputation edit groups but changes or additions may be made. In our study, EDIT1 of Appendix 2(ii) was present in all pre-imputation edit groups but was relaxed in the post-imputation edit groups . This relaxation consisted of two inequalities which defined a range within plus or minus 5 percent of Total Net Sales and Receipts. This action was necessary since no appropriate donors would be found so that a record would pass the pre-imputation edit. For later analysis it was necessary to re-establish the equality so that the sum of the components would be equal to Total Net Sales and Receipts.

If donor imputation was unsuccessful at both levels, estimator imputation was used. In general, the DIFTREND estimator was used. It calculates a trend based on the reported values from the current and previous surveys. There were a few cases in which this estimator was unsuccessful. This situation occurred when a record was not present in the previous survey or the mean of that variable was zero in the previous survey. In these remaining cases, either the ratio estimate based on the current survey or, when it could not be applied, an estimate based on the mean of the field in the current survey was used.

6 RESULTS

This section presents the results of the simulation study. First, we provide a comparison of results from each of the five selective follow-up rates and offer our subsequent conclusions and recommendations. The second part of this section gives a more detailed description of the results for the recommended rate of selective follow-up.

6.1 COMPARISON OF THE 5 SELECTIVE FOLLOW-UP RATES

The 5 files which corresponded to the 5 selective follow-up rates were analyzed and compared. With out of scope records removed, there were 1858 records remaining in the sample. 967 of these records failed at least one of the DC2 edit rules. Of these 967 records, 388 satisfied automatic follow-up criteria (Appendix 3). 245 of these 388 records were total non-response,

absolute discrepancy between the reported value of a variable in the current year (in our case 1987) and the final released value of the previous year (1986 for this study). This discrepancy is weighted by dividing it by the variable's cell estimate from the previous year. Each quantity in the summation is multiplied by the corresponding DC2 error flag for the variable (i.e. no contribution to the score value from a variable not in error).

Thus, a record's score value is:

$$\sum_{i=1}^{12} \frac{| Y_{(87)i} - Y_{(86)i} | E_i}{TOT_{(86)i}}$$

Where: $Y_{(87)i}$ is the reported value of variable i in 1987
$Y_{(86)i}$ is the value of variable i in the final 1986 released data
E_i is the error flag for variable i (equals 1 if error, 0 if not)
$TOT_{(86)i}$ is the cell total for variable i from final 1986 released data

Selected Upper percentiles of the distribution of this score constituted the selective follow-ups considered in this study.

4.2 GEIS METHODOLOGY

The functions of the Generalized Edit and Imputation System (GEIS) are performed by a set of modules, each of which performs a sub-task of the major processes: edit application, error localization, and imputation.

(i) Edit application

Edit application determines whether a data record contains incorrect, missing, inconsistent or outlying responses. The general strategy of GEIS is to produce an acceptable or 'clean' record. The GEIS system requires that all data values be positive, continuous and numeric. In addition, the edits must be linear. The edit groups are formed and are applied together to sections of a questionnaire or to subsets of the population. Each edit group defines a feasible region so that records which fall inside the region are acceptable and records which fall outside are unacceptable. The specification of edits is very important since it drives the identification of the fields to be imputed. The quality of the imputed data will only be as good as the quality of the edits. Hence proper analysis is required to ensure that the edit groups reflect the relationships between the variables of the questionnaire. Since GEIS assumes that the edits are linear and the data is positive, analysis of the edits can take place using linear programming techniques. After application of the edits to the data, the user can assess the impact of the edits on the data. Reports are generated which indicate the number of records passed, missed, or failed for each edit as well as overall counts at the record level.

(ii) Error localization

This process determines which fields of a record should be imputed. When a record fails one or more edits, there might be several combinations of fields that could be imputed so that the record would pass the set of edits. GEIS finds all those combinations which will minimize the weighted number of fields to be changed. These weights can be incorporated to reflect the reliability of a given field. Fields which need imputation are flagged.

(iii) Imputation

A chosen imputation method supplies valid values to fields of a record which have been identified to be imputed by error localization.

Donor imputation is one of several imputation options available in GEIS. This imputation strategy replaces the invalid and missing values determined by error localization with the values from a similar, clean record (a donor) from the current data file. This 'nearest neighbor' or 'hot deck' approach imputes all relevant fields at the same time, thus preserving as much of the underlying data structure as possible. The similarity of records is based on variables which are chosen as matching fields. These matching fields are a subset of the reported fields and are either generated by the system or are user-specified. A number of nearest neighbors are found, the closest one is used to supply the missing fields and the imputed record is re-redited. If the edits are not satisfied, then the next closest neighbor is tried until the record is successfully imputed or the supply of nearest neighbors is exhausted. The edits used to re-edit after imputation can be a relaxed set of the original edits.

Other imputation methods are available in GEIS. These alternative methods are deterministic in nature and use a pre-determined method to estimate the missing or invalid values. GEIS currently offers a selection of six estimators, three of which were used in the study and are defined as follows:

(a) DIFTREND (trend estimate based on reported values in the current and previous surveys)

$$\hat{Y} = \frac{\overline{Y}_t}{\overline{Y}_{t-1}} Y_{t-1}$$

(b) CURRATIO (ratio estimate based on the current survey, correlated auxiliary variable is from within questionnaire)

$$\hat{Y} = \frac{\overline{Y}_t}{\overline{X}_t} X_t$$

(c) CURMEAN (mean based on the current survey)

$$\hat{Y} = \overline{Y}_t$$

5 PROCEDURE

This second run of the DC2-GEIS Simulation study expanded on the ideas of the first run. The initial study demonstrated the compatibility between DC2 and GEIS. The focus now shifted to efficiency , essentially on the DC2 follow-up strategy and how the changes made would affect the DC2-GEIS simulation estimates.

5.1 DC2

After a questionnaire has been identified as being suspicious, (at least one error detected in DC2 edits), a follow-up is possible. As described in the report on the first run of the simulation, BSMD (1990), a follow-up was simulated in the following way. It was assumed that one follow-up

DC2 follow-ups were analyzed in terms of whether identified errors were followed-up as well as whether or not the follow-up constituted a confirmation (original value verified) or a correction (original value changed). The results are presented in Table 4. With a total of 587 follow-ups out of 1858 records, one can observe that very few errors were not followed-up for any variable. For example, 325 of 376 errors for Total Closing Inventory were followed-up. For variables frequently reported as zero, we expected most of the follow-ups to be confirmations. The table shows that this was indeed the case.

At the record level one may note that 463 of the 584 follow-ups had at least one correction performed. Overall we can conclude that the follow-up procedure is working well with the editing strategy. The variable considered most reliable by subject matter, Total Net Sales and Receipts, was corrected on 83 percent of the follow-ups (300 of 362) and thus the follow-ups would seem to have been worth the cost.

TABLE 4

BREAKDOWN OF FOLLOW-UPS
34% SELECTIVE FOLLOW-UP
(BY VARIABLE)

VARIABLE	NUMBER IN ERROR	IF IN ERROR, NUMBER FOLLOWED-UP	NUMBER OF CONFIRMATIONS	NUMBER OF CORRECTIONS
Net Sales	464	377	63	314
Gross Commissions	261	259	242	17
Receipts from Repairs	341	295	195	100
Receipts from Rentals				
Receipts from Food Services	271	262	231	31
Other Operating Revenue	252	252	246	6
Total Net Sales and Receipts	298	288	248	40
Non-Operating Revenue	410	362	62	300
Total Opening Inventory	341	341	253	88
Total Closing Inventory	461	351	50	301
Total Purchases	376	325	47	278
Total Salaries	507	371	55	316
Record level (correction is at least one variable changed)	403	335	59	276
	967	587	124	463

The GEIS imputation results were consistent with expectations. The file with the least amount of data corrected by DC2 (the file with zero percent selective follow-up) had the most fields identified for imputation in each of the data groups and as the number of follow-ups increased, the number of fields to impute decreased. For each of the 5 files, approximately 92 percent of

the records with fields to impute were imputed by donor imputation at the SIC4 level and approximately half of those remaining found donors at the trade group level. For the file with 34 percent of selective follow-ups by DC2, imputation results are presented in Appendix 6. Only 22 of 326 records remain to be imputed after donor imputation at the SIC4 level and 9 of 22 remained for estimator imputation.

7 CONCLUSIONS

This simulation study has confirmed the new concepts developed for DC2 and verified the compatibility of the DC2 and GEIS. The simulation study has been very worthwhile for investigating new ideas for the two generalized functions. The issues of survey budget, respondent burden, record consistency, and data quality have all been addressed by DC2 or GEIS.

Although the simulation was run for only the Annual Retail Trade Survey, we believe it is representative of business surveys in general. The results obtained are encouraging. It is clear that follow-up of at least some of the most influential units is advantageous with respect to the estimates obtained. Errors were usually investigated through these follow-ups. More specifically, with prespecified criteria automatically determining some follow-ups, follow-up of about one-third of the remaining errors brought the overall simulation estimates within 0.1 percent of the production estimates for the frequently reported variables. At the smaller production cell level, the standard error of this discrepancy was not significantly reduced when more than one-third of the remaining errors were followed-up.

The study has shown that following-up a limited number of units is sufficient to ensure acceptable data quality. If this approach is used in production, valuable resources can be saved without significantly affecting the data quality.

ACKNOWLEDGEMENT

This report is based on a study initiated and supported by the Generalized Survey Function Development Project(GSFD). The team consisted of J.M. Berthelot, P. Whitridge, M. Latouche, S. Perron, J. Morabito and B. Downer from Business Surveys Methods Division.

The team would like to thank J.F. Gosselin, L. Boucher and J. Beauchamp from Head Office Operations Division as well as R. Rasia, M. Rivest, and D. Roeske from the Annual Retail Trade Survey, all of whom made this study possible.

REFERENCES

1. Bilocq, F. and Berthelot J.-M. (1990), "Analysis on Grouping of Variables and on Detection of Questionable Units", Statistics Canada, Business Survey Methods Division, Working Paper No. BSMD-90-005E/F.

2. BSMD (1990), "An Integrated Approach to Data Editing, Error Correction and Imputation, Summary Report of the Simulation Study", Statistics Canada, internal report.

3. Kovar, J.G., MacMillan J.H., and Whitridge, P. (1988), "Overview and Strategy for the Generalized Edit and Imputation System", Statistics Canada, Business Survey Methods Division, Working Paper No. BSMD-88-007E.

4. Hidiroglou, M.A. And Berthelot J.-M. (1986), "Statistical Editing and Imputation for Periodic Business Surveys, "Survey Methodology 12", 73-83

APPENDIX 1

VARIABLES USED IN THE STUDY

VARIABLE NAME	DESCRIPTION
NSALES	Net Sales of new and used goods purchased for resale on own account
GCOMM	Gross Commissions earned for buying and/or selling merchandise on account of others
RREPAIRS	Labour receipts from repairs of automobiles, televisions, appliances, furniture etc.
RRENTALS	Receipts from rentals of televisions, home movies, automobiles, etc.
RFOODSRV	Receipts from food-serving activities
OTHREVEN	Other operating revenue (including rents of premises and other service activities)
TNSALES	Total Net Sales and Receipts (sum of the above variables)
NOPREVEN	All non-operating revenue (including subsidies, interest, dividends etc.)
PURCH	Cost of new and used goods purchased for resale, including purchase value of trade-ins, parts and materials for repair work
TOI	Value of stock on hand for resale (Opening)
TCI	Value of stock on hand for resale (Closing)
SALARIES	Total Salaries, wages, bonuses, commissions and any other payments to employees' earnings, excluding taxable benefits (i.e. includes gross payments before deductions for such items as Income Tax, Unemployment Insurance premiums, Canada and Quebec Pension Plans etc.)

APPENDIX 2

(i) DC2 EDITS FOR THE ANNUAL RETAIL TRADE SURVEY STUDY

EDITS PERFORMED BY TRADE GROUP:

EDIT 1 TREND OF TNSALES 1987 / TNSALES 1986
EDIT 2 TREND OF TNSALES 1987 / NSALES 1987
EDIT 3 TREND OF TNSALES 1987 / SALARIES 1987
EDIT 4 TREND OF NSALES 1987 / NSALES 1986
EDIT 5 TREND OF NSALES 1987 / PURCH 1987
EDIT 8 TREND OF SALARIES 1987 / SALARIES 1986
EDIT 9 TREND OF PURCH 1987 / PURCH 1986
EDIT 10 TREND OF TOI 1987 / TCI 1986
EDIT 11 TREND OF TOI 1987 / TCI 1987
EDIT 12 TREND OF TCI 1987 / TCI 1986

EDITS PERFORMED AT THE CANADA LEVEL:

EDIT 13 TREND OF NOPREVEN 1987 / NOPREVEN 1986
EDIT 14 TREND OF GCOMM 1987 / GCOMM 1986
EDIT 15 TREND OF RREPAIRS 1987 / RREPAIRS 1986
EDIT 16 TREND OF RRENTALS 1987 / RRENTALS 1986
EDIT 17 TREND OF RFOODSRV 1987 / RFOODSRV 1986
EDIT 18 TREND OF OTHREVEN 1987 / OTHREVEN 1986
EDIT 19 TREND OF NOPREVEN 1987 / TNSALES 1987
EDIT 20 TREND OF GCOMM 1987 / TNSALES 1987
EDIT 21 TREND OF RREPAIRS 1987 / TNSALES 1987
EDIT 22 TREND OF RRENTALS 1987 / TNSALES 1987
EDIT 23 TREND OF RFOODSRV 1987 / TNSALES 1987
EDIT 24 TREND OF OTHREVEN 1987 / TNSALES 1987

(ii) GEIS EDITS USED IN THE STUDY

EDITS COMMON TO ALL DATA GROUPS:

EDIT 1A: GCOMM + RREPAIRS + RRENTALS + NSALES + RFOODSRV + OTHREVEN $\geq$ TNSALES-1000

EDIT 1B: GCOMM + RREPAIRS + RRENTALS + NSALES + RFOODSRV + OTHREVEN $\leq$ TNSALES + 1000 (*)

EDIT 5: TOI + PURCH $\leq$ TNSALES + TCI

EDIT 8: NOPREVEN $\leq$ TNSALES

EDIT 15: TOI + TCI $\geq$ 0.5 (**)

EDIT 16: PURCH + SALARIES $\leq$ TNSALES

EDIT 17: TNSALES $\geq$ 1000000

EXAMPLE OF EDITS SPECIFIC TO ONE TRADE GROUP:

TRADE GROUP 20:

PURCH $\geq$ 0.41672*TNSALES
PURCH $\leq$ 1.24066*TNSALES

SALARIES $\geq$ 0.02159*TNSALES
SALARIES $\leq$ 0.53629*TNSALES

TOI $\geq$ 0.32674*TCI
TOI $\leq$ 2.90904*TCI

NSALES $\geq$ 0.67904*TNSALES

* IN POST IMPUTATION EDIT GROUPS:

EDIT1A: GCOMM + RREPAIRS + RRENTALS + NSALES + RFOODSRV + OTHREVEN $\geq$ 0.95*TNSALES

EDIT1B: GCOMM + RREPAIRS + RRENTALS + NSALES + RFOODSRV + OTHREVEN $\leq$ 1.05*TNSALES

** 0.5 WAS USED AS AN ARBITRARY CONSTANT IN THIS EDIT SINCE THE SUM OF TOI AND TCI NEEDED TO BE GREATER THAN 0 BUT GEIS WOULD HAVE AUTOMATICALLY CHANGED > 0 to ≥ 0

APPENDIX 3

DC2 PERFORMANCE MEASUREMENTS
34% SELECTIVE FOLLOW-UPS
(AFTER OUT OF SCOPE REMOVED)

				AUTOMATIC FOLLOW-UPS			
				(A)	(B)	(C)	(D)
TRADE GROUP	PROVINCE	# OF UNITS	NUMBER OF DC2 EDIT FAILURES (SUM OF [A]-[D])	TOTAL NON-RESPONSE	SMALL CELL UNITS IN ERROR	OTHER (*)	UNITS FAILING EDITS LEFT FOR SELECTIVE FOLLOW-UP
20	PEI	1	1	0	1	0	0
20	ALBERTA	7	2	2	0	0	0
20	CANADA	179	87	21	7	16	43
120	PEI	1	1	1	0	0	0
120	ALBERTA	73	39	8	0	5	26
120	CANADA	596	362	81	2	37	242
ALL	PEI	67	27	8	7	3	9
ALL	ALBERTA	1098	534	146	2	75	311
20 + 120	ALL	1858	967	245	17	126	579

* THESE UNITS WERE AUTOMATIC FOLLOW-UPS WHICH WERE NOT FOUND TO BE OUT OF SCOPE

APPENDIX 4

CELL SIZE DISTRIBUTION FOR 1987 FILE

CELL SIZE CLASS	FREQUENCY
0	0
1-9	22
10-19	4
20-39	8
40-59	4
60-99	3
100-199	4
200+	2

APPENDIX 5

FREQUENCY OF NON-ZERO VALUES REPORTED IN 1987
(BY VARIABLE)
(OUT OF 1858 IN-SCOPE RECORDS)

VARIABLE	1987 ORIGINAL FILE	1987 FINAL PRODUCTION FILE
Net Sales	1541	1851
Gross Commissions	101	94
Receipts from Repairs	739	871
Receipts from Rentals	162	191
Receipts from Food Services	49	46
Other Operating Revenue	258	276
Total Net Sales	1592	1858
Non-Operating Revenue	439	418
Total Opening Inventory	1553	1841
Total Closing Inventory	1562	1841
Purchases	1511	1842
Salaries	1566	1842

APPENDIX 6

SEQUENCE OF GEIS IMPUTATION
(34% OF DC2 SELECTIVE FOLLOW-UPS)

TRADE GROUP	RECORDS WITH FIELDS TO IMPUTE	DONOR IMPUTATION AT SIC4 LEVEL	DONOR IMPUTATION AT TRADE GROUP LEVEL	ESTIMATOR IMPUTATION
20 (N=179)	33/179	31/33	2/2	--
120(N=596)	113/596	102/113	6/11	5/5 BY DIFTREND ESTIMATOR
110(N=145)	24/145	24/24	--	--
100(N=365)	84/365 (*)	84/84	--	--
OTHER (N=573)	72/573	63/72	5/9	**

* 3 FIELDS WERE MANUALLY FLAGGED FOR IMPUTATION DUE TO PROCESSING DIFFICULTY (FIELDS TO IMPUTE OF FIRST SIMULATION USED)

** OF 4 REMAINING RECORDS, 3/5 FIELDS WERE IMPUTED BY DIFTREND ESTIMATOR, ONE EACH BY CURRENT RATIO ESTIMATOR AND CURRENT MEAN ESTIMATOR