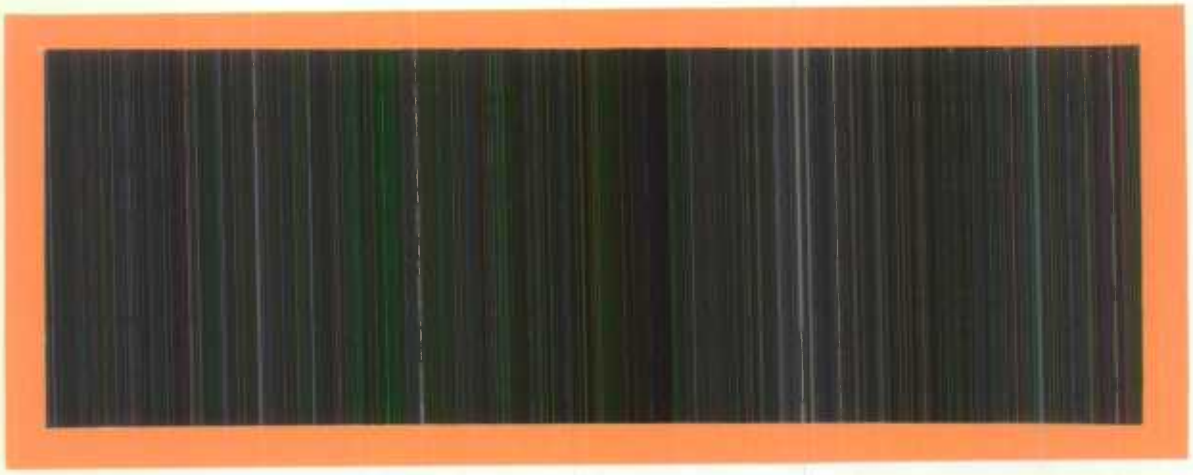

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes entreprises

Canada

WORKING PAPER NO. BSMD-85-063E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL NO. BSMD-85-063E

MÉTHODOLOGIE

REVIEW OF EDIT AND IMPUTATION IN STATISTICS CANADA

by

Philip Giles, Simon Cheung, George Kriger
Georges Lemaître, Sylvie Michaud, Charles Patrick

October 1985

* This is a preliminary version.
Do not quote without author's permission. Comments are welcome.

REVIEW OF EDIT AND IMPUTATION IN STATISTICS CANADA

A. BACKGROUND

The Edit and Imputation (E & I) Research Team was formed as one of the initiatives of the Methodology Research Committee. One of its main objectives is the development of a generalized E & I system (or systems) which will meet most, if not all, of the E & I requirements in Statistics Canada. One of the first steps taken was to conduct a review of the current E & I picture in Statistics Canada. This review was designed mainly to answer three questions. These questions are:

(i) Which are the various E & I methodologies currently being used in Statistics Canada?

(ii) What are the computer-related requirements of the current E & I systems?

(iii) Which E & I processing systems, if any, are suitable to be modified into a generalized system?

The decision was made to solicit the required information from the methodologists to the extent possible. A list of projects, and methodology contact persons, was compiled from the methodology division section chiefs. It was recognized that some projects could be missed in this way. However it was felt that the identified projects would provide a good knowledge of E & I in Statistics Canada. Projects currently in the developmental stage were purposely excluded. Two of these that are known are the Integrated Agriculture Surveys (IAS), and the Wholesale/Retail Trade Annual Survey Redesign.

A questionnaire was designed by the E & I project team. One questionnaire was completed for each of the identified projects, jointly between the contact person and a member of the E & I team.

A list of the projects for which a questionnaire was completed is given in Appendix A. A copy of the questionnaire, with the summary statistics in italics, is given in Appendix B.

B. RESULTS

The questionnaire was intended to give a general picture of E & I in Statistics Canada. When completing the questionnaires, it was found that it was not always possible to slot the information into the categories presented on the questionnaire. In these cases, explanations were written on the questionnaire. Along with the fact that multiple responses to certain questions are possible, the result is that the number of responses to various questions may be different. This was not regarded as a major problem to the analysis of the results.

The results obtained to the questions in Section A indicate the expected. The surveys conducted in Statistics Canada vary greatly in terms of their characteristics, aside from the data content. The number of numeric-only data files is higher than expected. Also further investigations should be conducted to determine the type of numeric/categoric mixture in the surveys which collect mixed data.

The types of editing which are performed indicate that a generalized system must be able to do more than to process the data file sequentially, one record at a time. Comparisons between records and with outside sources are required. This point is quite important to the development of a generalized system.

The question on manual editing was not very informative. It is suspected that most of the "Other" responses should be in the category "Insufficient resource to automate".

Reweighting is the most prevalent form of correcting for unit non-response. For item non-response, a number of different approaches are used.

When deciding which of the variables cause an edit, or set of edits to fail, most surveys treat the edits on an individual case by case basis. That is, a subject matter decision is made on the priority of variables.

While most surveys report that the imputed values satisfy the edits, it is perceived that a number of these satisfy the edits only because the imputed file is re-edited.

There was an even split on Question B6: Is the imputation dependent on the order of records, order of variables, or the generation of random numbers? Most of the surveys reporting "Yes" are thought to use a hot - deck approach, with the file processed sequentially.

Results different from those anticipated were received on the question on whether the choice of donor was dependent on the number of times each donor had been used. Due to the system implementation, a few surveys dropped a record from the donor list once it had been used. To use a donor more than once, the system must be re-run.

The question on the age of the computer system revealed that there were some old, some new, and some currently being revised. Half of the systems have been developed in the past five years.

It was interesting to see the number of positive responses, when asked if the computer system could be useful for other applications. It is thought that, for the most part, these "other applications" would be very limited.

There was generally a even split of opinion on the quality of the computer system documentation. The correlation between the response to this question and the amount of input from methodologists (particularly the methodologists completing the questionnaire) into the system is unknown.

Most of the computer systems were written in either PLI or COBOL, hardly a surprising result.

It is thought that those systems reported as an adaptation of another system were simply a revision of the previous system for that survey.

It is encouraging to note the large number of surveys with adequate knowledge of the E & I stage in both methodology and the systems area. It is felt, however, that this reflects an optimistic attitude of the respondents, rather than accurate reality.

Also encouraging is the relatively large number of surveys which are considering an alternate approach to E & I.

The reasons given for considering an alternate approach to E & I can generally be summarized into one. The computer system is obsolete, due to changing technology or to a survey redesign.

C. CONCLUSIONS

There are a number of important conclusions resulting from the review of Statistics Canada E & I.

1. There are no suitable candidates for generalization in Statistics Canada, other than those previously known by the project team. These are the PSTAT Numerical E & I system (NEIS), SPIDER, and to a lesser extent, CANEDIT and FIBCOC. As noted earlier, new systems currently under development were excluded from this review.

2. Traditionally, a far greater proportion of resources has gone towards editing (detecting errors) than towards imputation (resolving the edit failures). The structure of the edits is often very complex, the result of much preparation and study. On the other hand, the imputation procedure is generally mathematically straightforward, with a large amount of subject matter intervention. Manual imputation is frequently performed.

3. There appears to be a general willingness of subject matter divisions to adopt more sophisticated approaches to imputation, if the software is available.

APPENDIX A

List of Projects Reviewed

1. Labour Force Survey
2. Census of Agriculture
3. Census of Population
4. Vacancy Check Study - 1981 Census
5. Family Expenditure Survey
6. Food Expenditure Survey
7. Household Facilities and Equipment Survey
8. Survey of Consumer Finances
9. Absence from Work Survey
10. Travel to Work Survey
11. Canadian Travel Survey
12. National Farm Survey
13. Farm Credit Corporation Survey
14. Egg Producer Survey
15. Farm Price Survey
16. Farm Wages Survey
17. Farm Tax Data Project
18. Remote Sensing Project
19. Potato Objective Yield Survey
20. National Livestock Survey
21. Other Agriculture Mail Surveys (1)
22. Consumer Price Index - Rent Component
23. Industry Selling Price Index
24. Capital Expenditures Survey
25. Census of Construction
26. SEPH (Survey of Employment, Payroll and Hours)
27. Census of Manufactures (QUIPS) (2)
28. Census of Manufactures (SFES) (2)
29. Current Shipments, Inventories and Orders (CSIO)
30. Annual Traveller Accommodation Survey
31. Fare Basis Survey
32. Full Civil Aviation
33. Charter Survey
34. TRACC II
35. Private Trucking Origin and Destination Survey
36. For Hire Trucking Survey
37. Annual Retail/Wholesale Survey
38. Monthly Retail Survey
39. Monthly Wholesale Survey
40. Small Area Business Data Development
41. International Trade - Imports and Exports
42. Annual Survey of Corporation Taxation Returns
43. Tax Record Access
44. Business Register Master File
45. Periodicals Survey (3)
46. Disability Survey
47. Hospital Morbidity Survey
48. Transportation Survey for Special Care Facilities
49. Uniform Crime Reporting Program

50. Caseload Level 1 - Adult Criminal Courts
51. UCR - Homicide Program
52. Youth Court Survey

Footnotes

1. The list of surveys encompassed in one E&I questionnaire by "Other Agriculture Mail Surveys" are:

   a) Greenhouse Survey
   b) Nursery Survey
   c) March Intentions Survey
   d) March Stock Survey
   e) July Stock Survey
   f) December Stock Survey
   g) June Crop Survey
   h) August 1st Yield Survey
   i) August 15th Yield Survey
   j) September Yield Survey
   k) November Yield Survey
   l) Summerfallow and Stubble Survey
   m) Maple Survey
   n) July Sheep and Wool Survey
   o) Honey and Bee Survey
   p) Vegetable Processing Survey - Intentions
   q) Vegetable Processing Survey - Harvest
   r) Vegetable Processing Survey - Contracting

2. The Census of Manufactures has two E & I processing systems. Since there is a great difference between the two, QUIPS (Questionnaire Image Processing System) and SFES (Short Form Estimation System), two questionnaires were completed.

3. Culture Division conduct a large number of small surveys. The approach to E & I is very similar from survey to survey. Therefore, one survey was selected to represent all surveys in Culture Division.

Appendix B

Questionnaire and Survey Tabulations

The questionnaire is given in this appendix. The summary statistics are given for each question in italics. The counts reflect questionnaire returns as indicated in Appendix A. One questionnaire does not always mean one survey.

Review of E & I Systems in Statistics Canada

A. BACKGROUND

1. Survey/Project Name

2. Frequency of data collection:

Sub-annual	☐	20
Annual	☐	18
Less frequent than annual	☐	6
One off	☐	2

Note: The next two questions are intended to gain some information on the size of the dataset that is processed (i.e. E & I Processing).

3. Approximate number of records

Median = 20,000 *Low = 200* *High = 25,000,000*

4. Approximate number of variables subject to editing

Median = 30 *Low = 1* *High > 2,000*

5. Type of data processed:

Numeric	☐	20
Categoric	☐	1
Mixed	☐	29

B. METHODOLOGY USED FOR E & I

Note: Multiple responses to Questions B1-B4 are possible.

1. Approach to editing:

Automatic editing within record	☐	47
Manual editing within record	☐	25
Editing in comparison with previous survey	☐	20
Editing in comparison to other data sources	☐	16
Editing in comparison to other records	☐	15
Other - specify	☐	4

2. If Manual Editing is used, give reasons

Insufficient resources to automate	☐	4
Edits too complex to automate	☐	9
Simplicity of edits make it not cost efficient	☐	4
Other - specify	☐	15

3. Approach to correcting non-response (report separately for unit and item non-response).

	Unit		Item	
Report missing value as separate category	☐	2	☐	11
Reweighting	☐	21	☐	1
Determine value from other fields on record	☐	0	☐	21
Transfer value of missing field from another record	☐	9	☐	20
Determine value as a function of fields from another record	☐	1	☐	13
Determine value from another source - specify	☐	7	☐	16
Determine value by another means - specify	☐	4	☐	9
No Action		3		3
Manual		2		3

4. In the situation where there is a failed edit, how is the field(s) to impute determined?

Subject matter decision (pre-determined for each edit)	☐	30
Minimum change	☐	3
Random choice	☐	1
Other - specify	☐	18

5. Are the imputed values guaranteed to satisfy the edits?

Yes	☐	20
No	☐	13
Don't Know	☐	2

6. Is the imputation dependent on the order of records on the file, the order of variables on a record, or on the generation of random numbers?

Yes	☐	18
No	☐	14
Don't know	☐	3
Not Applicable	☐	16

7. If donor-candidate pairs are required, does the choice of a donor take into consideration the number of times each record has been used as a donor?

Yes	☐	11
No	☐	10
Don't know	☐	1
Not Applicable	☐	26

C. E & I COMPUTER SYSTEM

Note: If no automated E & I processing, go to Section D.

1. When was the computer system initially developed? Year ____

Median = 1980 *Low = 1967* *High = 1985*

2. Could this system be used for other applications?

Yes, with minor modifications	☐	*7*
Yes, but with major modifications	☐	*6*
No	☐	*32*
Not Known	☐	*1*

3. Is the documentation of this system:

Good	☐	*15*
Adequate	☐	*12*
Poor	☐	*10*
Non-existent	☐	*3*
Don't Know		*3*

4. What programming language or package is used by the system:

SPIDER	☐	*4*	FORTRAN	☐	*4*
CAN-EDIT	☐	*1*	PLI	☐	*21*
Sande Numerical System	☐	*1*	COBOL	☐	*17*
Other - specify	☐	*8*	SAS	☐	*3*

5. Was the system:

An adaptation of another system	☐	*7*
Programmed from scratch	☐	*38*

D. OTHER INFORMATION

1. Is a person available to provide detailed information on:

	Yes	No	Don't Know	Not Applicable
(a) Methodology	☐ 43	☐ 0	☐ 3	☐ 2
(b) Computer System	☐ 42	☐ 2	☐ 2	☐ 1

2. At this time, is the project considering an alternate approach to E&I?

Yes	☐ 18
No	☐ 27
Not Known	☐ 3

3. If "Yes" to Question D2, Why?

Old System was temporary only	☐ 1
Data quality was not always acceptable	☐ 3
Improved efficiency is necessary	☐ 9
Survey (or part of) is being redesigned	☐ 6
Other - specify	☐ 2

4. Person(s) providing information

5. Person(s) completing report.

E. A brief description of the system covering points not discussed in the previous questions.