Statistics Canada Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

C.2

WORKING PAPER NO. BSMD-92-010E

CAHIER DE TRAVAIL NO. BSMD-92-010E

METHODOLOGY BRANCH

DIRECTION DE LA MÉTHODOLOGIE

THE USE OF THE GENERALIZED EDIT AND IMPUTATION SYSTEM (GEIS) FOR THE 1991 CENSUS OF AGRICULTURE

by

C.2

S. Legault and D. Roumelis
November 1992

L'utilisation du Système généralisé de vérification et d'imputation (SGVI) pour le Recensement de l'agriculture de 1991

S. Legault et D. Roumelis

RÉSUMÉ

Le Recensement quinquennal de l'agriculture a répertorié un peu moins de 300 000 exploitations agricoles en 1991. Le questionnaire comprenait environ 400 variables dont 300 étaient sujettes au processus de vérification et d'imputation. Cette étape du traitement a été accomplie en se servant du Système généralisé de vérification et d'imputation. Cette utilisation du SGVI constitue la plus imposante et la plus complexe application du système jusqu'à présent et va vraisemblablement servir de modèle pour de futures applications d'envergure. Dans un premier temps, ce document expose les grandes lignes du SGVI ainsi que le cheminement qui nous a amené à choisir ce système pour le Recensement agricole. Par la suite, les diverses étapes préparatoires, les stratégies utilisées pour les tests et la production ainsi que l'expérience acquise lors de la production sont décrites. Pour conclure, on présente un résumé de la performance du SGVI en terme de coûts encourus et d'autres facteurs pertinents ainsi qu'une série de recommandations visant à améliorer et faciliter l'utilisation du SGVI.

1) INTRODUCTION

The Canadian Census of Agriculture, conducted every five years, enumerates all agricultural holdings which produce any agricultural products intended for sale. There are approximately 280,000 of these holdings in Canada. The Census questionnaire itself contains over 300 variables and is divided into several logical sections (eg. crops, livestock, expenses, etc.). Edit and imputation is the process whereby incoherent, inconsistent or missing entries are detected and subsequently replaced with consistent, plausible values. Variables may be related within or between sections making edit and imputation considerations complex. The bulk of the automated edit and imputation for the 1991 Census was accomplished using the Generalized Edit and Imputation System (GEIS). The majority of the data was imputed using a donor imputation method which is the approach favoured by the Agriculture Division. Other types of imputation such as deterministic imputation and to a lesser extent imputation estimators were used. They are described below. The 1991 Census of Agriculture represented the largest and most complex application of GEIS to date.

Due to processing limitations in GEIS, 16 edit groups representing the different sections of the questionnaire were formed. The population was divided into 52 geographically contiguous data groups of approximately 6000 farms referred to as imputation regions. The imputation regions respect provincial boundaries and are made up of similar farms.

This paper discusses all aspects of edit and imputation (E&I) development, processing and results for the 1991 Census of Agriculture. Since pertinent documentation on the development and functionality of GEIS is already available (Cotton, 1991 and Kovar, McMillian & Whitridge, 1988), the emphasis in this paper will be placed on the experience and results of using the system. First, following a brief overview of GEIS in Section 2, the E&I approach for the 1991 Census of Agriculture and the preliminary steps taken in preparation for production are described in Sections 3 and 4. A comparison between the 1986 and 1991 E&I processes in terms of processing strategy, data quality and costs is presented in Section 5. The paper concludes with observations, comments and recommendations stemming from the production experience (Sections 6 and 7).

2) GEIS OVERVIEW

GEIS software is made up of several modules (error localization, donor imputation, etc..) which can be adapted in order to satisfy the edit and imputation requirements of a specific application or survey. GEIS is embedded in the ORACLE relational database system. It assumes that the data to be processed are numeric, continuous and non-negative with edits expressed in a linear form. GEIS also assumes that some preliminary editing has been done at the data capture stage and that respondent follow-up has been completed.

3) 1991 E&I PROCESSING STRATEGY

Essentially, E&I processing for the 1991 Census of Agriculture involved the following activities:

a) the pre-processor program
b) fine-tuning of the edits
c) GEIS processing
d) post-GEIS processing
e) imputation of total refusals

Data validation was carried out after the completion of imputation for a whole province. During this process subject-matter officers analyzed data for each variable at the macro level (and when necessary at the micro level) in order to make a recommendation to the management of the Agriculture Division regarding the suitability for publication (or not) of the data.

a) Pre-processor

The purpose of the pre-processor program was to prepare the data for GEIS processing. The pre-processor is, in fact, a complement of the error localization module of GEIS. It began as a series of decision logic tables developed for each section of the questionnaire, and was subsequently converted to a computer program. It can be viewed as the operation that customizes the data which would otherwise not be processed properly by GEIS. For the most part, this is done as follows:

Firstly, for cases of partial non-response, the pre-processor identified the fields to be imputed, by setting their values to -1, so that GEIS, which does not allow negative values, would automatically flag the field for imputation. Partial non-response occurred when respondents indicated that they produced a certain commodity, by virtue of their positive answer to the screening question, but omitted to provide any data. Secondly, for instance where respondents only provided the total (i.e. failed to provide a breakdown), the parts adding up to the total were "zero-filled" making them valid entries instead of missing entries. The pre-processor was used to flag such fields as fields to be imputed. This step was required since GEIS always looks for a solution that minimizes the number of (weighted) fields to be imputed. Thirdly, the pre-processor also dealt with the conditional edits which were required for the application but could not be expressed in a linear form, which is a GEIS requirement. Finally, it resolved disagreements between summation of the parts and total by re-adding the parts and substituting the sum for the total. The pre-processor was especially useful for the land portion of the questionnaire where many totals and sub-totals had to be reconciled.

It is important to note that the pre-processor might not have been necessary if an "intelligent" data capture system, such as the Data Collection and Capture (DC2) System

had been used.
The pre-processed data was then loaded to ORACLE in two data tables. Two tables were required since the number of variables to be processed in GEIS exceeded the maximum allowed in a single ORACLE data table.

For more information about the land strategy (pre-processor) please refer to the document entitled "New Land Strategy" by J. Mayda, 1990 updated by S. Legault, 1991.

b) Fine-tuning

GEIS relationship edits were developed using 1986 Census data (section 4 b). Hence, it was expected that some parameters might have to be adjusted in order to reflect changes in agricultural practices during the 5 year period. There were also a number of new variables in the 1991 Census for which quality problems were expected. That is, the behaviour of new variables, regarding the passing or failing edits, was unknown, and could have resulted in a reduction of the donor population that would in turn seriously affect the quality of the imputation for other variables in the same edit group. Therefore, incoming 1991 Census data were used to make last minute adjustments (fine-tuning) to the edit parameters, when deemed necessary, before imputation in GEIS was undertaken. This involved executing, for each province, the Apply Edit module of GEIS. This provided us with a complete report on how many records failed the different edits and the overall failure rate for each edit group. In addition, SAS routines were used to analyze in more detail the relationship edits and recommend whether to accept as is, modify or drop edits in consultation with subject matter officers (SMOs).

c) GEIS processing

GEIS was the core of the 1991 Census of Agriculture Edit and Imputation system. It was responsible for editing, determining which fields of a record should be imputed, and imputing data based for the most part on the nearest neighbour approach. That is, based on the distance calculation involving the matching fields. The closest "clean" record (not requiring imputation) is used as the "donor" for the values of the fields requiring imputation.

Firstly, the edits were applied to the data for each edit group and data group combination using the Apply Edit module. This GEIS facility produces complete reports on edit failures providing, among other things, the overall rejection rates and the frequency of failure of each edit and field. This gave an indication of what to expect in the subsequent steps, error localization and donor imputation, in terms of complexity, time and computer resources.

Secondly, the error localization module was executed. The purpose of this module was to identify the fields to be imputed for each record in error so that the record would satisfy all the edits. The fields to impute selection was made according to the rule of minimum change, that is, GEIS always finds the solution that minimizes the sum of the weights of fields that are identified for imputation. The weighting option in error localization was used in order to have some control over the selection of fields to impute since, for this particular application, not all fields had the same level of reliability or "importance".

The final step in GEIS processing was the actual imputation, primarily donor imputation, of the fields identified during error localization, that is, finding, for each edit failure, the closest donor record whose values allowed the recipient record to pass all the edits. In order to facilitate the search for a suitable donor, some edits (referred to as post-imputation edits), especially equality edits, were slightly relaxed in the first run of donor imputation. If a second run was required, the edits were relaxed even more. For most edit groups the search for a donor stopped after two runs. For some edit groups, estimator imputation using the current ratio estimator and/or deterministic imputation were used as a supplement to donor in order to reduce the amount of manual imputation required. The Deterministic Imputation module determines, for each field identified as requiring imputation, if there is only one possible value which would satisfy the original edits. If such a value is found, it is substituted for the erroneous entry. The current ratio estimator uses means from the current Census to impute for missing entries. For example, missing market values of specific types of machinery were imputed by multiplying the reported number of such machines with the average market value for the machine in the imputation region.

During the preparation period, a production manual was written describing all processing steps to be performed within ORACLE, including GEIS and non-GEIS jobs such as manual imputation and post-processing (Roumelis 1991). For details on production procedures please consult this manual.

d) Post-GEIS Processing

This phase of E&I processing actually refers to a series of steps conducted after GEIS was executed. The first step is the manual imputation of records which could not be imputed by GEIS. These records were identified using programs which referred to the ORACLE table that contains the imputation status at the record/field level (fieldstat table). Records were updated by manually entering values into prewritten update programs. The next step was to update the fieldstat table to indicate that imputation was completed. Backups of the data tables were then created in case the subsequent post-processor programs aborted and the data had to be recovered. Check programs were run to ensure that there were not any errors in the manually imputed data. Finally, post-GEIS processing programs were used to prorate, round and re-add the parts of each section to equal the total. There were two such programs.
One program, called the LANDSQL program, updated the land variables from the first data table. Another program, the post-processor, updated variables from the second data table and, for the most part, only prorated those variables which were imputed by referring to the fieldstat table. Data was then loaded to ADABAS for data validation.

Please see the production manual (Roumelis, 1991) for details.

e) Imputation of total refusals

The imputation for total non-response was performed outside of GEIS. Each total refusal was imputed with a single donor. Any useful information on non-respondents that was gathered, either from the Farm Register or through field collection reports, was to be accounted for during imputation. Each total refusal record was paired with a non-refusal in the same imputation region using total farm area (TFAREA) and farm type as matching variables, when this information was available. Once the records were matched, the data from the donor was transferred to the recipient. The rationale in using TFAREA and farm type as matching fields was that these quantities would provide a good indication as to what type of donor we should be looking for. When such information could not be determined, a donor in the geographic proximity of the recipient was randomly selected.

This approach was preferred over using GEIS for two important reasons. Firstly, it was estimated that the amount of jobs to be run would be extensive and would add significantly to an already busy production schedule if GEIS were used. Secondly, since one of the matching variables was discrete (farm type), it could have led to inappropriate donor selection. For example, suppose that there are five recipient records all with farm type equal to 5 and TFAREA unknown. In the donor population there are 100 records with farm type equal to 5. The same donor record will be selected for all five recipients. However, GEIS has since been improved and it would now select, at random, one donor out of the 100 possible donors for each of the recipients.

4) PRELIMINARY STEPS

a) Feasibility study

The use of GEIS was considered for several reasons. Firstly, the 1986 E&I system was being pushed to its physical limits and the only new requirements it could deal with were changes to the questionnaire. The system, which was not fully documented, required a dedicated mainframe system with operator assistance in order to function adequately. Secondly, at Statistics Canada, there is strong encouragement by management to incorporate generalized systems in as many areas as possible. The goal is to conserve resources by eliminating the duplication of effort through the use of common approaches and methods for different surveys. As well, current thought in the Bureau is to move away from editing survey questionnaires sequentially, one at a time, to macro- and selective editing top down approaches where efforts are concentrated on verifying high impact records. The rationale is that it is more efficient to spend less time examining records of low contribution to an estimate, and focus on records which have significant impact. As for the quality and coherence of the smaller records, it would be ensured by GEIS.

A working group was formed in late 1989 to investigate the possibility of using GEIS for the 1991 Census of Agriculture. The prototype focused on the livestock section of the questionnaire and involved approximately 40,000 records from the previous Census. Edits were also developed for other sections of the questionnaires in order to have a better estimate of the overall processing costs. The main issues addressed were:

1) the determination of edits suitable for GEIS;
2) the quality of the imputed data;
3) assessment of wether the processing schedule could be met in 1991 and;
4) estimation of the cost of using GEIS.

The conclusion from the prototype experiment was that the 1991 Census of Agriculture data could be processed using GEIS while meeting the methodological and systems requirements. The projected cost of implementing and using GEIS was higher than that of the estimated budget for 1991. However, it was noted that GEIS would yield some long term savings since it could be reused in future Censuses.

For more information, please refer to the document entitled "Business case for using GEIS for the 1991 Census of Agriculture" (Working group on the use of GEIS, 1989).

b) Edit rules determination

The edit strategy in GEIS is to define the profile of an acceptable or 'clean' record using a set of linear edits. These edits are then combined to define a region called an acceptance region. Records which satisfy all the edits fall inside the acceptance region and those which do not satisfy at least one of the edits fall outside the region. Such records will subsequently be imputed so that they can enter the acceptance region. Edits in GEIS have two important purposes:

1) at the editing stage, to identify fields that require imputation.
2) at the imputation stage, to identify the criteria under which a donation is considered successful since, with GEIS, constraints are put on the donor to ensure that the recipient will satisfy all the specified edits after imputation.

The edits are central to GEIS processing. Failure to establish a suitable set of linear edits could have negative repercussions both on processing and on the quality of the data produced. Firstly, relationship edits which were used to identify inconsistencies between fields, were determined. Two steps were taken in order to establish such edits. The first step was to consult with experts in different areas in order to identify the variables that they felt were correlated. 1986 Census data was then used to measure and confirm the strength of the correlation or linear association between two variables. The second step was to determine the actual edits between the correlated pairs of variables identified in the first step. Linear regression models were used to establish the parameters of the edit rules which were of the type $Y < mX + b$. The rejection rates associated with each edit were calculated and the parameters (slopes and intercepts) were adjusted in order to obtain acceptance regions that were satisfactory. The parameters of some of the relationship edits differed from one province to an other. In addition to relationship edits, edit groups were made up of logical edits ($A+B+C=D$) and conditional edits ($A>k\times B$ with $k=$a sufficiently small constant to ensure that if $B>0$ then $A>0$). Conditional and logical edits were derived following analysis of the 1991 Census of Agriculture questionnaire.

For more information about the edits refer to the document "1991 Census of Agriculture: GEIS edits, weights and matching fields" by S. Legault, 1991.

c) Modular testing

Preparation to incorporate GEIS into the Census of Agriculture included the execution of modular tests, that is, the testing of the various GEIS modules using Census of Agriculture data from 1986. These tests were conducted by methodologists between January and June of 1991. The reasons for conducting these tests are many and will be described in turn.

1) Modular testing was conducted to estimate CPU and MSU consumption and elapsed real time in order to plan a computing resources budget and schedule for the production period.

   For several sections of the Census questionnaire, that is, several edit groups, all GEIS modules were executed and costs recorded. The procedure was run twice, once during mainframe prime time and once at night, since the charges for these times differ. It was decided to use imputation estimators later in the development of the GEIS application and therefore this module was not tested during the modular testing phase. In quantifying costs there were some confounding effects. Firstly, a new charging system was introduced in April of 1991, halfway through the modular testing period. Secondly, toward the end of the modular testing period, version 6 of the ORACLE database system was introduced, confounding comparison between resource consumption in the modular and integrated tests.

   Based on the results from modular testing, GEIS processing costs were expected to be over six million MSUs, equal to the entire processing budget for the Census.

   An attempt was made to reduce costs by varying job parameters and scenarios, then comparing these results to the original costs. Only one parameter or scenario was altered for each test in order to isolate factors affecting processing. Scenarios tested included the following:

   i) the difference between the use of -1s or nulls for error localization. It was thought that -1s enabled error localization to run more quickly.

ii) the use of ORACLE indexes on the stamp number (the unique Census of Agriculture identifier) and/or imputation region variables. Generally, indexes allow data in an ORACLE table to be accessed more quickly.

iii) dropped or altered edits. Fewer or less complex edits enable GEIS to run more quickly.

iv) different values for the GEIS parameter for the size of the first and second pass (ie. n_1 and n_2) in donor imputation were tested to assess the impact on CPU consumption.

The results of these experiments showed that only the addition of an index on the stamp number significantly and consistently reduced costs of GEIS jobs. However, an additional index on imputation region was shown to reduce the cost of non-GEIS programs such as the post-processors. If a variable appears frequently in the condition of a program statement, as imputation region does, processing becomes faster if an index on that variable is used. Altering other parameters did not significantly reduce costs. Since not all possible scenarios were tested, these findings are not exhaustive.

2) Modular testing was also conducted to establish the sequence of GEIS job execution to ensure that the highest quality imputation was eventually performed. When using GEIS, after error localization is performed, the methods of imputation and the sequence of execution to be used are the decision of the user. It was previously decided that donor imputation would be the primary method of imputation to be used. However, it was found, during testing, that deterministic imputation could be useful for certain edit groups. If there is only one possible value a variable can take in order to satisfy the edits, the deterministic imputation module assigns this value. Later it was determined that the imputation estimator module, using the ratio estimator based on current data, was useful when an auxiliary variable was present, such as in the tree fruits and machinery edit groups. Please see the production manual for details on where these modules were used.

3) In the error localization module, the user may limit the execution time to be spent finding a solution for any single record. If a solution can not be found in the specified time, each field in the edit group is assigned a Time Limit Exceeded status in the fieldstat table. Modular testing enabled methodologists to measure the frequency of the occurrence of TLEs and determine appropriate time limits for each edit group. TLEs posed a problem in complex edit groups such as machinery. Since TLE records were to be imputed manually, it was desired to keep their frequency of occurrence low. In edit groups where TLEs were frequent, it was decided to rerun error localization with a higher time limit in a second execution.

4) Different field weights for the error localization module were also tested. The weights in GEIS are used to reflect the level of reliability of each variable. Variables which are

usually well responded to are assigned higher weight making them less likely to be selected for imputation. Field weight determination has a direct impact on the selection of fields to impute and subsequently on the quality of imputation. This was especially relevant in the land portion of the questionnaire which was made up of numerous totals and sub-totals spanning seven edit groups. An appropriate weighting strategy was required for error localization in order to maximize the use of the information provided by the respondents. Therefore, different field weight structures were tested until a satisfactory approach was established.

For the most part, the weighting strategy involved assigning higher weights to the totals of sections indicating that reported totals were considered more reliable than parts and hence making them less likely to be imputed. Some edit groups were treated differently. The machinery edit group, for example, which asks the number and current market value of several types of farm machinery, had a lower weight placed on the total. This did not indicate a lack of confidence in the reported total but was done for logical reasons. Each part value of the total value was involved in a relationship edit with its quantity. If the ratio between the value and quantity of a given part was not acceptable, the value and then the total were imputed. A higher weight on the total would prevent it from being imputed, thereby forcing another value to be adjusted in order that parts continue to add to the total.

5) GEIS has a built-in facility that selects appropriate matching fields for a recipient based on failed edits. These are used to calculate the distance between the recipient and the potential donor. GEIS also allows the user to specify additional match fields (called user-specified match fields). These fields are used in the distance calculation regardless of whether or not the system would have chosen them, because the user has deemed that this information is important in the search for a good donor. Different sets of user-specified match fields were tested for each edit group.

d) Integrated testing

The integrated test was conducted by production staff after they were trained to operate GEIS. GEIS training took place in early July 1991. The emphasis was on GEIS job submission outside of the GEIS menu system because using the menu system was considered too time consuming for the volume of job submissions expected. Two people were originally trained.

Integrated testing ran from July through September 1991. The purpose of this procedure was to orchestrate the job execution sequence, estimate costs of submitting a full job stream both during the day and at night, test computer capacity by running four job streams concurrently, and prepare personnel for the rigorous production schedule. A job stream is the series of GEIS jobs, submitted in the predetermined sequence established in the modular testing phase, for one data group or imputation region.

As previously mentioned, by the time of this test a new mainframe computer charging system and version 6 of ORACLE had been installed. An updated version of GEIS (version 6.3.2) was tailored for use with the new ORACLE version, and was used for Census production.

A problem which caused massive increases in CPU consumption was encountered repeatedly throughout this testing period before it could be solved. Toward the end of the testing period it was discovered, through consultation with SDD, MCC and ORACLE Corporation, that the fieldstat table was becoming fragmented, that is, stored inefficiently, becoming extremely expensive to read from and write to. The solution to the problem was to periodically export the table, drop the table from ORACLE and reimport. This enables the data to be stored more efficiently. Unfortunately, although invaluable experience in running the system was gained, this problem prevented adequate estimates of CPU consumption from being produced during integrated testing. Other tables, such as the actual data tables and the donor_map (which contains list of recipient records and donors) table, are subject to this problem as well but not to the same degree.

5) PRODUCTION EXPERIENCE

As previously mentioned, the Census of Agriculture was the largest GEIS application thus far. The Census enumerated 280,000 agricultural holdings collecting information for some 300 variables. In terms of edit and imputation processing, this translated into over 3500 GEIS jobs and 1500 non-GEIS jobs to be submitted. The rate of processing necessary to complete production by the deadline was one imputation region every two days. Generally, several regions were processed concurrently, averaging two days per region.

The production schedule lasted five months from October 10, 1991 to March 13, 1992. From October to January, one technical officer ran all procedures, often working overtime. When it became evident in January that the production schedule might not be met, another technical officer and another, faster, disk pack were added to the project. The rationale here was that because ORACLE is input/output bound, using another pack would allow more reading and writing to the disks in the same amount of time. Generally, production proceeded as planned. In retrospect, fewer problems were encountered than expected and the resulting delays added up to about two weeks. Because of the tight schedule, the delays sometimes appeared worse than they were. These problems, their solutions, and their effect on production are presently discussed in the following lines in approximately chronological order.

1) One of the first problems encountered was an ORACLE error, namely the ORACLE deadlock error (ORACLE error 0060). This error occurs when two commands from different programs attempt to update the same record. The deadlock occurred a few times causing job execution to abend (abnormally end). The problem was traced to an SQL program called TREEMACH (see production manual) which was used as an

addition to the pre-processor in order to enable estimator imputation to be used for the tree fruits and machinery sections of the Census questionnaire. A condition used in several statements was erroneously allowing the program to update records in imputation regions other than the region desired, while GEIS was already executing in these regions. The conditions were corrected and the job streams affected were re-executed after a one day delay. The error was never again encountered.

2) Fragmentation of the fieldstat table was encountered only twice during the production period, around early December and toward the end of production in late February. The first occurrence was indicated by an explosion in CPU time, a symptom of this problem which was first observed during integrated testing. The second occurrence was indicated by ORACLE error 4187 which caused jobs to abend. In consultation with ORACLE Corporation, it was suggested that fragmentation might be the problem. As mentioned previously, the solution was to export the fieldstat table, drop the table from the ORACLE account, and re-import the table, which solved the problem. In each instance, the delay in processing was a couple of days.

3) ORACLE error 4288, which was encountered several times in early January 1992, and error 4289, encountered in late January, caused GEIS job streams to abend. No positive explanation for either error was offered from database support personnel. However, it was thought that the errors may have occurred due to disk space problems. Partly in response to the first error, and partly due to concerns that the production schedule might not be met due to similar delays in the future, another technical officer and another faster disk pack were employed to assist. Savings in turnover time were realized. In response to the error 4289, the buffer size (a concept similar to RAM on a microcomputer) was increased. Neither problem occurred again but it is not known for sure if the changes mentioned are responsible.

4) In general, problems due to ORACLE errors were not easily solved. This is partially due to the fact that the messages received when ORACLE errors are encountered are not very descriptive. As well, some errors were only identifiable by database support personnel or ORACLE Corporation. However, even when simple ORACLE errors were encountered, where the accompanying message was understood, production personnel often did not know how to cope and were forced to consult methodology. Often, if an error affected subsequent jobs, which it usually did, a job stream would have to be re-executed, causing a one day delay. This type of problem was encountered perhaps a half dozen times.

6) COMPARISON BETWEEN 1986 AND 1991

In comparing the 1986 and 1991 E&I systems, three important issues will be discussed: the E&I methodology used in each of the Censuses, the quality of the data imputed and the cost.

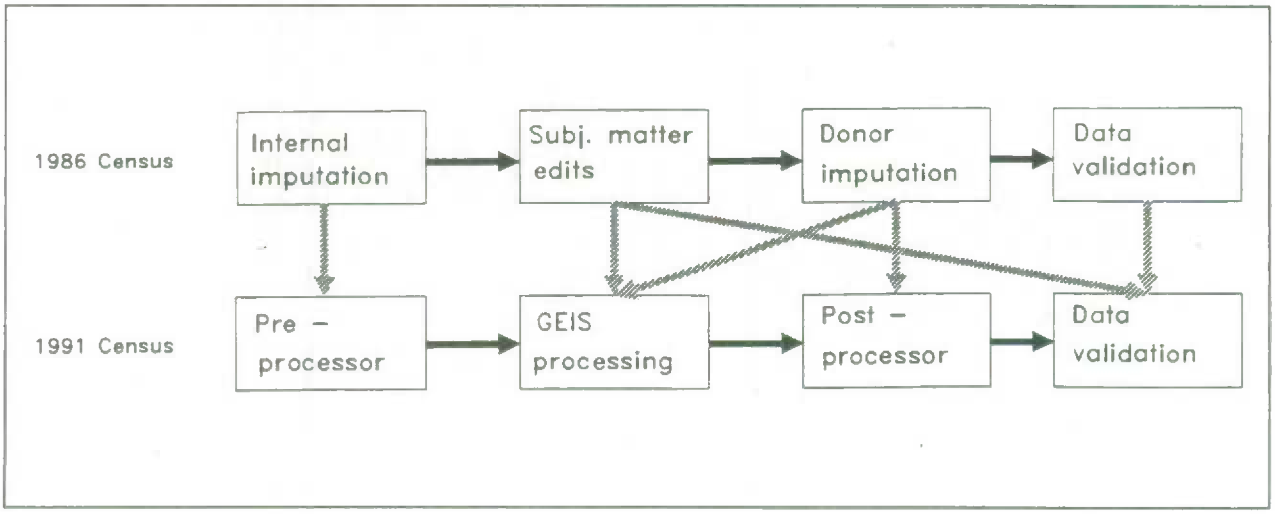
a) Processing strategy

Figure 1.(next page) shows the flow of E&I processing for both the 1986 and 1991 Censuses. The major difference between the two processes is the integration of the subject-matter edits (SME) and data validation (DV) processes in 1991.

In 1986, the SME process consisted of the identification and correction (if necessary) of edit failures (outliers, unusual relationships or large discrepancies between parts and totals) before donor imputation. Questionnaires containing edit failures were verified manually by subject-matter officers. Values were then changed, if necessary, based on SMO knowledge, follow-up with respondents or outside data sources. SMO also had the option of flagging fields for donor imputation if they could not be imputed manually.

As for the data validation process, it consisted of the production of several sets of tables aimed at providing information necessary for the SMO's to defend and certify their estimates before the Certification committee. Changes were made to records if an SMO felt that some records were faulty.

FIG. 1.
Edit and Imputation process: 1986 compared to 1991



It was noted during the feasibility study that some records (approx. 3%), verified and accepted (or corrected) at the SME stage of the 1986 Census of Agriculture, could later be flagged as being in error and be imputed by GEIS. This was identified as a problem since it meant that SMO's, in the event of imputation of a top contributor field by GEIS, could end up correcting the same record twice (once in SME and once in DV). Therefore, the E&I team recommended that the two processes be merged into an enhanced data validation process. This new approach would remove duplication between the two processes as well as putting more emphasis on macro analysis and follow-up of only high impact records, following Statistics Canada strategy.

This meant that subject-matter edits would no longer be applied before donor imputation. The functionality of the process was however maintained by the improved DV process. In addition, some of the subject-matter edits were linearized and integrated into GEIS edits. This was a major change of methodology for the Census of Agriculture since, for the first time, SMO's would not get the opportunity to look at any questionnaires before imputation was completed.

Another important modification was the processing environment. In the last Census, all stages of processing where carried out using the ADABAS database management system. In 1991, some activities such as the pre-processor and data validation were carried out in ADABAS while GEIS and post-GEIS processing utilized the ORACLE relational database system. Therefore, the data had to be loaded and unloaded into ORACLE in the middle of production, adding complexity and time to an already demanding job submission schedule.

There was also a considerable increase in the number of jobs to be run. In the past, one job submission per province was required for each of the stages. Even the donor imputation stage was run at the province level for all edit failures but required a dedicated mainframe system with operator assistance. In 1991, donor imputation with GEIS necessitated several job submissions, one for each data group - edit group combination, because of the limit on the number of variables and records GEIS can handle at one time.

b) Quality of the data imputed

A priori, the quality of the imputed data in 1991 was expected to be improved for several reasons:

- improved linear edits were developed following consultation with experts and thorough analysis of the 1986 Census data.
- these edits were fine-tuned using incoming 1991 Census data.
- GEIS ensures that any imputed record is internally consistent, that is, satisfies all the (post-imputation) edits simultaneously.
- GEIS imputes records in a consistent and reproducible manner.

It is difficult to measure to what extent these factors influenced the quality of the imputed data. There is, however, no doubt that the concepts used in GEIS were a great improvement over the methodology used in the past. The latter method primarily involved manual imputation and was prone to large biases depending on who performed the imputation and/or when the imputation was actually performed. Furthermore, comments from data validators (whose mandate is to certify each estimate) indicated an overall satisfaction with the imputed data. Therefore, we believe that the quality of the imputation was quite acceptable and comparable to the level of quality obtained in the past but further studies should determine how good it really was. We do not want, at the present time, to imply that the improved E&I process automatically translated into improved data quality.

c) Cost

It is estimated that the edit and imputation process (pre and post-processors + GEIS) required 500 hours of CPU time which amounts to approximately $200 000. This represents an increase of $100 000 (in 1991 dollars) over the cost for the equivalent operation in the 1986 Census. However, the use of GEIS and the improved E&I and data validation methodologies lead to a small decrease in human resources (person-years) as compared to 1986. In addition, we assume that some of the supplementary cost will be absorbed by long term gains. As mentioned earlier, GEIS is a system that can be easily reused in the future. It also does not necessitate any maintenance by the Agriculture Division and is constantly refined and improved based on comments and suggestions made by different GEIS users.

70% of the total CPU time required during GEIS processing was needed to perform the donor imputation phase and 20% for error localization. The remaining 10% was used to perform estimator and deterministic imputation. Prior to production, many improvements were made to error localization in order to make it less expensive. These enhancements resulted in a very important decrease in computer resources. Efforts should now be put on reducing time comsumptions in the donor imputation module since even modest savings could significantly reduce the overall cost. Non-GEIS procedures such as the post-processor and the LANDSQL also proved to be very expensive. Ways at making these programs more efficient is also suggested (see conclusions and recommendations section).

7) CONCLUSIONS AND RECOMMENDATIONS

The Census of Agriculture is the largest application of GEIS both in number of variables and records processed. The Census experience has shown that the software can be succesfully used for large applications and can be a viable alternative to customized systems. Furthermore, the use of GEIS has had positive repercussions on other stages of processing such as data validation where the macro and selective editing aproaches made this stage more efficient than in the past. Based on the experience gained throughout the Census development and production period, a number of recommendations for future consideration have been put forth by both methodology and agriculture personnel. These are summarized below.

1) It is suggested that a production manual be produced before training production staff. It is felt that the manual was not consulted enough because staff became accustomed to working from their own notes. Often, when a problem was encountered and their notes did not help, methodology was consulted where the manual could have been used. This was especially true where the SQL language was concerned.

2) Non-GEIS production procedures and troubleshooting required extensive use of the SQL language. Thus a facility with SQL was necessary. However, due to time constraints, the language was not learned adequately by the production staff. Consequently, much consultation with methodology was required throughout the production period. More training regarding SQL procedures is recommended.

3) Non-GEIS procedures were often set up as SQL scripts, that is, as programs consisting of a series of SQL statements. The post-processor programs were of this form. Composed of nearly 2000 lines each, they were highly inefficient. The programs were written before it was discovered that a structured programming language, called PL/SQL, was available for use. It is recommended that this language be used for SQL programming in the future. It is also recommended that SDD review and optimize any such programs, a procedure that was planned but not carried out.

4) It is recommended that more staff be trained to run production in the future. The staff employed in 1991 were overworked.

5) GEIS was generally run outside of the menu system for this application, because switching screens constantly was considered too cumbersome. In order to reduce the job submission workload, possibly by several times, it is recommended that a method be developed, inside or outside of the GEIS menu system, where GEIS parameters could be specified and a series of jobs be submitted, with a minimum of screen switching and/or scrolling. This recommendation received support from the director of Agriculture Division. One idea discussed was a "super screen" where GEIS parameters for any GEIS jobs could be specified and submitted from the same screen, and submitted together in a job stream if desired.

6) The donor_files created during each execution of donor imputation should have been dropped by the system after the imputation was completed, but were not since they were used for data validation purposes. This resulted in space problems during testing, after which ORACLE account sizes was increased to accommodate the problem. Still, the files had to be dropped periodically because of the space they occupied, interrupting processing. It is suggested that these files be moved to a different environment (OS file, SAS dataset, etc..) in order to free valuable ORACLE space.

7) Although the (relationship) edits were approved several months prior to production, subject-matter officers had several comments regarding the edit strategy during the course of production. It appears as if the GEIS concepts were, perhaps, not fully

assimilated by the SMO's. More time should be dedicated in the future to ensure that they understand the approach and that their participation in the process is increased.

8) The TREEMACH program, which was a late addition to the E&I process, could not be incorporated in the pre-processor. It is suggested that such a process be part of the pre-processor in the future in order to reduce costs and the number of jobs to be submitted.

9) The fine-tuning process turned out to be very helpful especially for the tree fruits and machinery sections. As a result of this operation, several ratio edits involved in these sections were updated following consultation with SMO's. The process did not require a lot of time and resources and enabled us to monitor effectively the quality of the incoming data. It is imperative that such a process be repeated in 1996.

10) There were some important keying errors made at the data capture stage that resulted in records becoming top contributors at the province level. GEIS could not detect such mishaps unless the erroneous fields were involved in a relationship edit. GEIS works under the assumption that data capture errors have already been resolved. However, GEIS could detect and correct errors of this sort if an upper bound limit was to be assigned to each field. This would, however, increase significantly the number of edits and subsequently the cost associated with each GEIS job. It is recommended that basic preliminary data capture edits be performed (perhaps on DC2) before processing data through GEIS.

11) The process which was used to impute the total non-response was not optimal. It appears that the information regarding the type of farm (on the recipient record) was not taken into consideration when looking for a suitable donor. A system problem was the source of this omission and could not be fixed during production because of resources and time constraints. GEIS, which now has the random donor feature, should be considered for the imputation of total refusals in 1996.

12) SMOs were reluctant to introduce relationship edits because they thought this would eliminate some "valid" unusual farms and that during data validation they would be required to change data, altered by GEIS, back to their original values. This was a very time-consuming and difficult process during data validation. It might be a good idea to have a system built for data validation where original values could be re-inserted easily. Please see the memorandum "Use of GEIS" (Shields, 1990) for more information.

13) In order to aid the manual imputation process, a GEIS or SQL screen should be developed which will automatically generate a screen containing blanks where values for any of the variables in a specified edit group can be entered or updated directly.

14) Manual imputations during GEIS processing were tedious and not user-friendly. Furthermore, some of the manual imputations resulted in records that were outside the

acceptance region. A menu driven facility that would enable updates to records and then perform edit checks would be desirable

15) There were some difficulties in using two ORACLE tables to store the Census data. For one, some variables had to be included on both tables. Therefore, updates on the second tables had to be made if variables in the first table were modified. This lead to additional job submissions and increased the possibility of mishaps. It also made job submission slightly more difficult since different edit groups referred to different data tables.

16) The approach used for imputation in GEIS appears to be satisfactory. The great majority of the donors were found during the first run of the donor imputation module. The second run, which had less restrictions on the donor record usually resolved the remaining faulty records. The proportion of records requiring manual imputation was below 1% of all records requiring imputation (i.e. approx. 3 000 records). The assignment of weights in error localization and the strategy used in the land portion of the questionnaire were appropriate and produced the expected results. As a result of the apparent success of the process, no major changes in the approach is recommended. Any changes to the approach used in 1991 should be performed with care, especially regarding the land strategy. One important recommendation would be to investigate ways of reducing the cost of donor imputation jobs.

17) There were too many jobs to be submitted (≈5 000). The technical officer had to keep track of several jobs simultaneously making her task that much more complicated. It also lead to contention problems that resulted in significant slowdowns.

At the present time, it is still too early to measure the performance of GEIS in an exhaustive manner. Further evaluations should provide us with more in-depth results pertaining to the behaviour of GEIS in 1991. However, we can safely conclude that the use of GEIS was very successful in achieving its objectives, that is, improving the efficiency of the whole E&I process while maintaining (and possibly increasing) the quality of the final product. Moreover, Census of Agriculture management has expressed great satisfaction with the software and recommended its use for other surveys.

REFERENCES

Kovar, J.G., MacMillian, J. and Whitridge, P. (1988), "Overview and strategy for the Generalized Edit and Imputation System". (Updated February 1991). Statistics Canada, Methodology Branch Working Paper No. BSMD-88007E/F.

Cotton, C., (1991), "Functional description of the Generalized Edit and Imputation System". Statistics Canada technical manual.

Legault, S., (1991), "1991 Census of Agriculture: GEIS edits, weights, and matching fields". Statistics Canada technical document.

Mayda, J., (1990), "New Land Strategy". (Updated by S. Legault, 1991).
Statistics Canada technical document.

Roumelis, D., (1991), "1991 Census of Agriculture - Edit and Imputation Production Manual". Statistics Canada technical manual.

Working group on the use of GEIS for the 1991 Census of Agriculture, 1989), "Business case for using GEIS for the 1991 Census of Agriculture". Statistics Canada technical document.

Shields, M., (1990), "Use of GEIS". Memorandum.