Statistics Canada Statistique Canada

Methodology Branch

Business Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes-entreprises

Canada

bcxcq

WORKING PAPER No. BSMD-88-007E

METHODOLOGY BRANCH

CAHIER DE TRAVAIL No. BSMD-88-007E

DIRECTION DE LA MÉTHODOLOGIE

C-1

OVERVIEW AND STRATEGY FOR THE GENERALIZED EDIT AND IMPUTATION SYSTEM

Z-190E?

by

J.G. Kovar, J.H. MacMillan and P. Whitridge

April 1988
(Updated February 1991)

OVERVIEW AND STRATEGY FOR THE GENERALIZED EDIT AND IMPUTATION SYSTEM

(Updated February 1991)

J.G. Kovar, J.H. MacMillan and P. Whitridge

ABSTRACT

The Generalized Edit and Imputation System (GEIS) currently being developed at Statistics Canada, is expected to meet the numerical edit and imputation requirements of most of the Bureau's economic surveys. The automation and to a larger extent the generalization aspect of the system have made it necessary to develop an edit and imputation strategy which is different from what has been used traditionally. This paper first provides an overview of GEIS followed by a presentation of the new strategy underlying its use.

RÉSUMÉ

Le système généralisé de vérification et d'imputation (SGVI), qui est présentement développé à Statistique Canada, devrait satisfaire les exigences de la plupart des enquêtes économiques du Bureau en ce qui a trait à la vérification et à l'imputation des données quantitatives. L'automatisation et particulièrement la généralisation de ce système ont rendu nécessaire le développement d'une stratégie de vérification et d'imputation qui soit différente de ce qui avait été utilisé jusqu'à maintenant. Ce document donne premièrement une idée générale du système suivie d'une présentation de la nouvelle stratégie sous-jacente à l'utilisation du système.

1. INTRODUCTION

Historically, the approach for edit and imputation for most surveys of economic production at Statistics Canada has consisted predominantly of detection and manual correction of errors as the records are received and reviewed. According to the type of error detected, any one of several courses of action may be taken, including follow-up with the respondent, manually supplying ad-hoc values to complete the erroneous fields, overriding the edit, excluding the record, or, often as a last resort, automated imputation. Being mostly manual in nature, this approach to edit and imputation is usually very subjective and generally not reproducible. As such, process statistics and status reports are rarely available, rendering impossible the assessment of the impact of the imputation.

The introduction of computers in survey processing resulted in little more than the automation of various stages of this manual, sequential, "detect and correct" approach. Moreover, developing software has been difficult because the specification of an edit followed by an action has required the programming of an unmanageable number of conditions. This often resulted in systems so large and complex that no survey record could pass all the edits. This tendency to overedit and the proliferation of multiple systems have lead to serious nonconformity of approaches between the various surveys, even in similar situations.

In developing generalized systems, the task of edit and imputation has been broken into two stages: preliminary editing, which is done at the data collection and capture stage, followed by edit and imputation. It is assumed that a substantial amount of correction and all follow-up and document control are done at the preliminary editing stage. Only unresolved cases or cases of minor impact would be passed to the Generalized Edit and Imputation System (GEIS), at which point an effort is made to resolve all problems by imputation. It is the latter edit and imputation system which is described here.

The GEIS software consists of methodologically sound modules to be assembled by a knowledgeable user, typically a methodologist in conjunction with a subject matter specialist. The system may not supply all the options available in a tailor made system, but in most respects it is quite flexible. By automating the system, the edit and imputation process becomes more objective and reproducible. Using one system and one general strategy yields conformity between surveys, while the production of complete status reports makes evaluation possible. To ensure that a particular application does not result in a self-contradictory and/or redundant set of edit rules, various analytical functions are provided within GEIS. The development of the system itself has been based on the Numerical Edit and Imputation System (Sande, 1979), and on the work of Fellegi and Holt (1976) for coded data.

The first section of the paper provides an overview of GEIS. The advantages and limitations of generalized software are considered. Each major part of the system, including editing, error localization, imputation and outlier detection is described. Definitions and concepts are provided, followed by a discussion of the main functions. The second section presents an edit and imputation strategy to be used by those developing a specific application using GEIS. The paper concludes with suggestions for further reading.

2. OVERVIEW

2.1 Generalized Systems

A generalized system is a collection of computer programs which can be used in a variety of situations. That is, the system is not tailored to a specific application. The development of such software has been emphasized as part of the Business Survey Redesign Project (BSRP). The BSRP is attempting to conserve resources by eliminating the duplication of effort through the use of common approaches and methods for business surveys. In other words, the Bureau can no longer afford to build specialized, survey specific systems, which require annual modifications and maintenance.

Although the development of generalized software is initially expensive, it pays for itself by eliminating the need for the development of customized systems and the maintenance of multiple systems. Of course, a generalized system cannot offer the specificity and speed of a tailor made program, but the compromises required of the user are expected to have no detrimental effect on the quality of the data while offering greater flexibility. Since a generalized system is intended to replace many customized systems, resources for developing and in particular, for maintaining the software, become concentrated. Updating the software and its documentation therefore becomes faster and easier since the expertise is centralized. As such, the support that can be offered to the user is improved in quality and availability.

The number of functions that GEIS must perform results in the requirement for the development of complex software. The complexity is reduced by developing the system in modules, each of which performs a subtask for one of the major system tasks of editing, error localization, imputation, or outlier detection. Edit analysis for example, is a subtask of editing. There are numerous advantages to modularity (Giles and Patrick, 1986). First, it is conducive to ongoing development of the system. With the scope of surveys GEIS is expected to serve, not all specifications can be anticipated in the initial stages of development. It may be necessary to add modules over time and to enhance existing ones. The addition of a new module is much simpler than attempting to incorporate the function into existing code. Changes may be made independently to an existing module thereby reducing the chances of introducing new errors elsewhere in the system. Secondly, modularity lends itself to prototyping: the development of an initial set of modules which is used for only a few surveys. User reaction to the prototype will influence the direction to be taken for subsequent system releases as modifications and enhancements are suggested based on experience. Thirdly, modularity facilitates comparison of different edit and imputation strategies for a given application. In fact, it allows different applications to use different strategies by rearranging the same basic components. As well, because the system is constructed in modules, there is the opportunity for smaller surveys to afford the cost of the software by selecting appropriate, ready-made modules. A survey does not have to reach some specific size to benefit from processing through GEIS.

GEIS is embedded in the ORACLE relational database management system. A database management system, in general, facilitates the organization and the manipulation of data. This is particularly important to the system, not only in actually performing some of the main tasks, but in monitoring the edit and imputation processes. Summary statistics such as the number of times an edit was failed, or the number of times a particular record was used as a donor during imputation are produced to help the user with the monitoring. In addition to the above, ORACLE was chosen mainly because it is portable across different computer architectures.

2.2 Editing

The objective of editing is to determine whether a given data record contains incorrect, missing, inconsistent or questionable responses. To accomplish this task, the edit component of GEIS consists of four main parts: specification of edits, analysis of edits, production of edit summary statistics tables, and outlier detection. At this time GEIS requires that all edits be linear and all data values non-negative. Even with these restrictions most of a user's edit requirements can be accommodated by GEIS, although some edits may have to be specified in different ways.

The specification of the edits is done interactively, possibly on a micro-computer, utilizing two or three different screens. Through any one of these screens the user specifies the edit identifier, whether the edit is a pass or fail condition, and the edit itself, by providing the variables with their coefficients and the constant. There is also a facility to update the edits, attach comments to the edits, and automatically date the changes. Other screens are used to group the edits. These groups are used to create edit sets that will be applied either to sections of questionnaires (e.g. income versus expenses) or to subsets of the population (e.g. different SIC groups) or, in some circumstances, to different edit and imputation functions such as error localization and post-imputation (to verify that the imputed record is satisfactory). The system performs some syntax verification including checking as to whether arithmetic operators have been correctly specified, whether the edits are linear, and whether all variables referenced are, in fact, part of the questionnaire.

Further edit analysis is possible as a result of the assumption of linearity of the edits and non-negativity of the data. Linear programming techniques are used to analyze the edit set beyond mere syntax (Sande,1979). When this function is invoked, GEIS verifies the consistency of the edits, that is, it ensures that the set of edits is not self-contradictory. For consistent edit sets, the system also identifies any redundant edits, that is, edits which do not further restrict the feasible region of data values in the presence of the other edits. By identifying the redundant edits, the system identifies the minimal set of edits. The system then generates the acceptable ranges for all variables, the extremal points of the feasible region, and the set of implied edits (Sande, 1979). All three of these diagnostics can aid the analyst in verifying that the edits specified are meaningful (Giles, 1987 and 1988; Sande, 1988), and act as a check on the correct entry of the edits.

GEIS applies the edits to the data and internally classifies the records as pass, miss, or fail (Giles, 1986b). Tables which provide information on counts of edit failures cross-classified by various dimensions are produced.

Finally, the system provides a facility for outlier detection. This module considers all the data records at once and therefore should not be applied at the preliminary edit stage, unlike the linear edits. The method is based on the work of Hidiroglou and Berthelot (1986). Given the data, the module determines upper and lower acceptance bounds for each requested variable or for the ratio of the variable's current to previous values. The module can serve two distinct purposes: to determine the edit bounds (using previous data) or to identify outlying values which can be flagged for imputation or for other considerations in subsequent modules (e.g. exclusion from calculation of trends).

2.3 Error Localization

Error localization is the process of determining 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 number of fields to be changed. The user also has the option of minimizing a weighted number of fields to be imputed. In this way it is possible to indicate to the system which fields are considered more or less reliable. The fields which need imputation are flagged internally, for use by subsequent modules. The error localization problem is recast as a cardinality constrained linear program (Sande, 1979) and is solved using Chernikova's algorithm (Rubin, 1973). Details on the use of Chernikova's algorithm in Error Localization and other parts of GEIS may be found in Schiopu-Kratina and Kovar (1989).

After the fields to be imputed are identified, the user may choose to have the system determine if there are any fields on the record that can be imputed in only one way. This imputation is performed if any such fields are found. In other words, if the set of edits together with the valid entries determine a unique solution, such a solution is imputed. This is called deterministic imputation.

2.4 Imputation

Imputation is the procedure that supplies valid values for those fields of a record that have been identified for change as a result of the error localization. The new values should be supplied in such a way as to preserve the underlying structure of the data and to ensure that the resulting data record will pass all the required edits. In other words, the objective is not to reproduce the true micro-data values, but rather to establish internally consistent data records that will yield good aggregate estimates. At the present stage of development, GEIS provides two broad categories of imputation.

The first is a donor imputation method based on the nearest neighbour approach. In this case, the invalid and missing values are replaced by values from a similar, clean record. The similarity of records is judged based on some or all of the valid, non-missing values. This procedure operates on a set of variables defined by an edit group and tends to preserve the structure of the data, since all needed variables in one edit group are imputed at the same time. That is, not only are the values themselves imputed but so are their interrelationships. To ensure that the edits are satisfied, several nearest neighbours are found, and the closest one which produces a record that satisfies the post-imputation edits is used to impute for the record, provided that such a donor exists. The post-imputation edits may be different from the original edits.

The second category consists of various model-based imputation estimators which are deterministic in nature and consist of replacing the missing or invalid values using a predetermined method. The available methods include most of the traditional procedures such as the imputation of a previous observation for the same respondent, a mean of current or previous observations, a previous observation adjusted by a trend, as well as methods based on ratio estimators. Details of the methods and exact formulae may be found in Giles (1986a) or Cotton (1991). It must be noted that these methods are unlikely to preserve the structure of the data as well as donor imputation (Bureau, Michaud and Sistla, 1986). None ensure that the edits will be satisfied. As such, they are primarily intended to serve as backup methods for the donor imputation, but they can also be useful when imputing repeated subannual survey data.

In the case of imputation estimators, the system will permit a choice of method of imputation by field, as well as the sequencing of imputation methods. In other words, each field can be potentially imputed using a different method, or more than one method if prior methods prove unsuccessful. The sequencing should increase the chances of imputing successfully by providing alternatives if a preferred method fails.

Various reports are generated to facilitate the monitoring of the imputation process. The tabulations include items such as the number of times a record was used as a donor, the number of records which had 1, 2, 3, or more fields imputed, and the number of imputations attempted. If desired, the user may query the database in order to obtain additional information, such as which fields of any specific record were imputed and what methods were used.

3. STRATEGY

This section of the paper provides a general outline of how a user of GEIS should approach assembling the various modules to suit an application. To this end, we first describe the computing environment, followed by the specific edit and imputation steps. A more detailed description of the functions performed in each module may be found in Cotton (1991), along with numerous examples. Recent experiences of several applications are described in Whitridge and Kovar (1990).

The objective of GEIS is to provide a complete and consistent data set, in preparation for the final stages of survey processing: estimation, tabulation and dissemination. It is assumed that before the data are passed to GEIS all attempts to follow-up the respondents have been carried out, and that all reporting data and document control variables, which are items that GEIS will not process, have been cleaned up. The input data file may thus contain missing or inconsistent entries, either because all follow-up avenues have been exhausted or because the size of the unit has not warranted extensive follow-up. These unresolved cases will be cleaned up by imputation.

Note that in the present release of the system, all negative values are treated as incorrect. If the user has variables which can be both positive and negative, this problem can be overcome by expressing them as a difference of two positive variables, provided that there are not too many such cases. For example, "profits" can be expressed as "gains - losses". Alternately, a large constant can be added to the variable in question and all related edits modified accordingly.

In general, it is assumed that edit and imputation is the last processing stage before estimation and, for the most part, time constraints rule out manual intervention and repeated attempts at imputation through interactive parameter control.

3.1 Computing Environment

GEIS is embedded in the ORACLE relational database management system. All GEIS facilities may be accessed through a menu system. While it is not absolutely essential to become familiar with ORACLE and the underlying Structured Query Language (SQL), a basic proficiency will give the user added flexibility in assembling the GEIS modules. In fact, because the database can be queried at any time, the users may monitor the edit and imputation process more effectively and thoroughly. In other words, the impact of any module can be assessed almost instantaneously. Furthermore, because ORACLE, and therefore GEIS, is portable, the user can take advantage of the strengths of the various architectures. That way, for example, the edit specification and analysis may be done interactively at a micro-computer, while the time consuming tasks such as edit application, error localization and imputation may be done more effectively using the mainframe computer.

The GEIS modules themselves may be invoked through the use of a menu system which performs a verification of associated parameters. The appropriate syntax rules are described completely in the GEIS Operations User's Guide (GEIS Development Team, 1991). The sequencing of the commands and their parameters is at the user's discretion. As such, it is assumed that the user not only understands the subject matter at hand, but also is familiar with the basic concepts of editing and imputation. Even though some steps have been taken to warn the user of potential pitfalls, it is impossible to ensure that the system will not be used incorrectly. Therefore we would strongly suggest that GEIS applications be assembled by methodologists in conjunction with subject matter officers, while computer specialists may be required to establish a complete production environment.

3.2 Editing

In order to develop a successful application of GEIS, the user essentially needs to specify only one thing: the description of an acceptable or "clean" record. This is the set of conditions which a record must satisfy so it will be acceptable for further processing. These conditions are specified by means of linear edits whose purpose is to identify acceptable and unacceptable records. Note that no information as to how to react to the individual edit failures is provided to the system by the user, thus simplifying the development substantially. It is the system itself that identifies the fields to impute. While this seems overly simple at first sight, one must appreciate the importance of specifying the edits well, since lack of any other information that would drive the system implies that the quality of the imputed data can be only as good as the quality of the edits. Because of the importance of the edits, many analytic functions are supplied in GEIS in order to make the development phase easier. Further discussion as to how a user should approach this phase can be found in Whitridge and Kovar (1990).

Many of the linear edits specified to GEIS may resemble those specified at the preliminary edit or data collection and capture stage. However, there are three notable differences. First, for a preliminary edit, there may be multiple actions specified depending on the severity of the edit failure. In GEIS there are no "degrees" of failure: all failures will be automatically imputed. Secondly, preliminary edits need not be linear or even numeric, unlike GEIS edits. Thirdly, most data collection and capture edits are to be applied to units at the reporting level, while it is assumed that for business surveys, GEIS will typically process data at the statistical level. This may necessitate a modification or an exclusion of the preliminary edits and possibly the addition of other new edits when deriving the edit set to be used by GEIS.

3.2.1 Linear Edits

Defining an acceptable region of data points through the use of linear edits will likely be unfamiliar to users accustomed to the more traditional "detect and correct" approach. In fact, quite often, translating previous edit and imputation specifications will be a more difficult task then implementing new edit and imputation requirements. Thus deriving the requirements from the specifications, either implicitly or explicitly, will have to be attempted in the case of existing surveys. It is these requirements that are then used to derive the edits for the specific application.

Some requirements may be expressed in terms of linear edits quite readily while others may be difficult, and at times impossible (Fitzpatrick, 1988). In particular, range edits and bounds are very naturally linearized. For example, the edits "sales < 1,000,000,000" or "0 < wages < 1,000,000", are bound and range edits respectively, which are in a linear format. Also simple to express are magnitude relationships between variables such as "wages < sales" or "salaries < .9*profits", and balancing (accounting) edits such as "closing inventory = opening inventory + purchases - sales". Furthermore, some conditional edits can be recast as linear edits, while others must be dealt with by splitting the file associated with the questionnaire. For example, the condition: "if sales are greater than zero, then purchases must also be greater than zero" can be linearized by the edit "purchases > constant*sales" with a sufficiently small constant, though in this case the system may impute sales=0 to resolve a conflict. On the other hand, the edit "if 1000 < SIC < 1999 then X + Y < Z" may only be processed by treating the 1000 to 1999 SIC range as a separate file. Due to space, time and effort limitations not many of these ranges should be specified, and the extent to which they are required might have to be re-examined.

Furthermore, certain edits which appear to be non-linear, such as the bound on a ratio of a pair of variables: $X/Y < 10$, may be rewritten in a linear form as $X < 10*Y$. Others like $X/Y < Z$, may only be linearized through a suitable transformation: $\log X - \log Y < \log Z$. Note, however, that a given edit set can only contain a transformed variable or the original variable, but not both, in order not to disrupt the error localization function. Thus, one transformation may necessitate other transformations. Still, there will remain edits which cannot be linearized, and whose purpose must therefore be reconsidered, so that other approaches may be evaluated. This, once again, stresses the need for a clear set of edit and imputation requirements.

In summary, the edits must be derived bearing in mind the intent of the questionnaire, its implied accounting rules, validity ranges, and the general requirements of the survey. Care should be taken to decide where given edits are applied. For example, edits which deal with expected response patterns (e.g. if sales > 0 then purchases > 0) are best placed at the preliminary (data capture) stage, while bounds and range edits are appropriate at the GEiS stage. For more details see Kovar and Whitridge (1991).

3.2.2 Edit Specification and Application

The actual specification, input and analysis of the edits should be done well before the data is available, as soon as the survey questions have been finalized. This task can be accomplished on line, so that syntax errors detected by the machine can be corrected immediately. Furthermore, the edit analysis functions, including the consistency and redundancy check, and the generation of extremal records and implied edits, are intended to help the user create a consistent, minimal edit set that describes accurately the variable relationships on the questionnaire, or part thereof.

In particular, the Check Edits module will first perform a consistency check to determine whether or not the edits are self-contradictory. Following the confirmation of consistency, the system will check for redundant edits which are those that do not restrict the region where acceptable records must lie. By removing all redundant edits, the user can create a "minimal set of edits": a set of edits which defines the same region as the original set, but whose further processing is more efficient due to the reduced size.

Secondly, the Extremal Points module generates fictitious records which would pass all edits but which are in the corners of the acceptance region. Such records may suggest to the user that some edits should be changed, or other, more restrictive edits, should be added to the existing set. On the other hand, the Implied Edits module generates linear combinations of the input edits, thus uncovering conditions which are being imposed on the variables but which have not been stated explicitly in the original set of edits. Implied edits which indicate that some variables are being overly constrained may suggest a review of the original edits.

The foregoing stages of analysis are likely to be performed repeatedly, in order to arrive iteratively at a satisfactory edit set or group of edits, for each logical part of a questionnaire and possibly industry grouping. As the actual data are passed through the system, careful monitoring of the edit results is essential. The generated reports include counts such as the total number of edit failures for a given record, the number of times a given edit was failed, and the number of records that had a given number of edit failures. This information can be used to improve the questionnaire design, survey procedures and most notably the edits themselves. Because the specification of edits is an evolutionary process, the addition, deletion, modification and documentation of edits have been made easy in GEIS.

3.2.3 Edit Groups

As mentioned above, edits are always considered in sets since often the pattern of edit failures is of greater importance than the individual edit failures themselves. Edits are placed into such sets using the edit grouping facility in GEIS. Each edit must belong to at least one group but it can belong to many groups. In any case, it is specified only once. The need for this facility becomes more evident in the case of larger, more complicated surveys. For example, it may be necessary to process logical parts of the questionnaire separately, such as crops, livestock, and expenses sections of a farm questionnaire, using very different edits. Secondly, some industries may have to be processed independently while sharing many common edits. Thirdly, sophisticated users will appreciate the flexibility offered through this facility in fine tuning the edit application - error localization - imputation interface.

3.2.4 Outlier Detection

The Outlier Detection module is considerably different from the modules which have been discussed so far in that it provides an inter-record edit rather than an intra-record edit. In other words, it compares values for given fields between records, rather than a set of fields within a given record. Most notably, it may be used as a stand alone module, without any reference to imputation, in order to identify outlying fields, either for manual inspection or other considerations. On the other hand, outlier detection may be used in conjunction with the edit and imputation process. In particular, the identification of outlying values is useful for imputation evaluation, or for exclusion of records from the donor population or from contributing to the averages used by imputation estimators. As well, the module can be used to flag fields for imputation by GEIS, or for differential treatment at the estimation phase outside of GEIS. In its univariate form, a derived edit can be used as a linear edit in future applications.

3.3 Error Localization and Imputation

To guide the system in performing this last task, the user will have a choice of donor imputation as well as other imputation estimators which have been described briefly in Section 2.4. The fields to be imputed will have been selected automatically using the given set of edits and taking into account the particular pattern of responses for the given record. A limit may be placed on the processing time to be spent on any individual record so that a small number of intractable records do not consume a great deal of execution time. These records must be resolved manually.

For most general applications, donor imputation should perform satisfactorily. It has the advantage of imputing all relevant fields at the same time, thus preserving as much of the underlying data structure as possible. There are, however, situations when other methods may do better for some specific variables. This is to be established using prior subject matter knowledge or data analysis. For example, historical imputation, possibly trend adjusted, is likely quite appropriate in the case of monthly surveys. Other methods are made available as back-up methods in the event that preferred methods fail for one reason or another. For example, donor imputation can fail when no donor can be found so that the resulting imputed record would satisfy all post-imputation edits. Historical imputation can fail when previous values are unavailable, or the trend cannot be calculated with sufficient reliability.

The record can be broken into segments and imputed by donor imputation in different phases by making use of the edit grouping facility. This allows for more donors at any given phase, different stratification on each pass, and the use of imputed data for subsequent matching or adjustment.

The choice of imputation estimators is made by specifying selected methods for each field. Sequencing of methods is accomplished simply by specifying the same field several times, each time associated with a different method, in descending order of preference. The user can also fine tune the calculations associated with some of the imputation estimators. For example, in calculating trends, imputed values can be excluded as can records satisfying user specified criteria and records identified by the outlier detection module.

As with editing, imputation results can be monitored based on tabulations generated by GEIS. This information will include frequencies such as the number of records which were imputed, the number of times a certain field was imputed, and for donor imputation, the number of times a record was used as a donor. This information can be used to improve the particular application.

4. CONCLUDING REMARKS

Many of the GEIS modules have parameters that can be changed by the user to fine tune the application. A description of these options is not within the scope of this document, but the interested reader may refer to several other sources for more information. Cotton (1991) discusses the steps performed by each module and the methodological reasons behind the GEIS procedures so that potential users can better understand and evaluate these procedures. Many simple examples are provided. As well, the GEIS Operations User's Guide, (GEIS Development Team, 1991) describes how to install GEIS and how to run the menu system. This guide describes the fields which appear in each menu and the actions that each menu performs.

Three other documents are currently in preparation. The GEIS Application Strategy Guide explains how to conduct a feasibility study, how to develop edit and imputation strategies, and how to monitor the progress of an application. Secondly, the GEIS Application User's Guide gives the technical details of how to get started in GEIS and how to customize GEIS to suit the needs of a specific application. Thirdly, the GEIS Tutorial provides the user with a data set in machine readable form and steps the user through the entire system from creating ORACLE tables, loading data, performing edit and imputation to unloading the data after imputation is complete.

At its present stage of development, GEIS is able to accommodate the edit and imputation requirements of many surveys and has already been used successfully by a number of applications. New modules and refinement of existing modules are always under consideration. The GEIS development team will continue to accept suggestions for additional features, and will respond by incorporating these improvements as time, resources and need permit.

ACKNOWLEDGEMENT

The authors wish to thank Cathy Cotton for her help in updating this document.

REFERENCES

Bureau, M., Michaud, S. and Sistla, M. (1986). A comparison of different imputation techniques for quantitative data. Statistics Canada, Methodology Branch Working Paper No. BSMD-87-002.

Cotton, C. (1991). Generalized Edit and Imputation System Functional Description. Statistics Canada Technical Report.

Fellegi, I.P., and Holt D. (1976). A systematic approach to automatic edit and imputation. **Journal of the American Statistical Association 71**, 17-35.

Fitzpatrick, T. (1988). Report on the feasibility of using the Generalized Edit and Imputation System for the Annual Wholesale, Retail Surveys. Statistics Canada Technical Report.

GEIS Development Team (1991). Generalized Edit and Imputation System Operations User's Guide. Statistics Canada, Research and Generalized Systems Subdivision Technical Report.

Giles, P. (1986a). Generalized edit and imputation - part II. Statistics Canada Technical Report.

Giles, P. (1986b). Methodological specifications for the generalized edit and imputation system. Statistics Canada Technical Report.

Giles, P. (1987). Towards the development of a generalized edit and imputation system. **Bureau of the Census Third Annual Research Conference Proceedings**, 185-193.

Giles, P. (1988). Generalized edit and imputation of survey data. **Canadian Journal of Statistics 16, Supplement**, 57-74.

Giles, P. and Patrick C. (1986). Imputation options in a generalized system. **Survey Methodology 12**, 61-72.

Hidiroglou, M.A. and Berthelot, J.-M. (1986). Statistical editing and imputation for periodic business surveys. **Survey Methodology 12**, 73-83.

Kovar, J.G. and Whitridge, P. (1991). Generalized Edit and Imputation System: Overview and Applications. **Revista Brasileira de Estatística**, to appear.

Rubin, D.S. (1973). Vertex generation in cardinality constrained linear programs. **Operations Research** 23, 555-565.

Sande, G. (1979). Numerical Edit and Imputation. Presented at the 42nd International Statistical Institute Meeting, Manila, Philippines.

Sande, I.G. (1988). A Statistics Canada perspective on numerical edit and imputation in business surveys. Presented at the Conference of European Statisticians, Geneva, Switzerland, February 2-5.

Schiopu-Kratina, I. and Kovar, J.G. (1989). Use of Chernikova's algorithm in the Generalized Edit and Imputation System. Statistics Canada, Methodology Branch Working Paper No. BSMD-89-001E.

Whitridge, P. and Kovar, J.G. (1990). Applications of the Generalized Edit and Imputation System at Statistics Canada. **American Statistical Association 1990 Proceedings of the Section on Survey Research Methods**, Anaheim, California, August 6-9, to appear.