GENERALIZED EDIT AND IMPUTATION SYSTEM FOR ECONOMIC SURVEYS AT STATISTICS CANADA

J.Kovar, P.Whitridge and J.MacMillan, Statistics Canada
J.Kovar, 11C R.H.Coats Building, Tunney's Pasture, Ottawa, Ontario, K1A 0T6

ABSTRACT

With the advent of more powerful computers, Statistics Canada decided to develop a Generalized Edit and Imputation System (GEIS) for economic surveys. The system is embedded in the ORACLE Relational Database Management System, and as such, is portable across various computer architectures. The paper describes the methodology used in the development of the GEIS and provides the technical details of the available options. The system is presented as both a production and an evaluation tool. Its limitations and future enhancements are also discussed.

KEY WORDS: Nonresponse, Generalized Software, Edit Analysis, Donor Imputation

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, imputation. This predominantly manual 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 inconsistency of approaches among the various surveys, even in similar situations.

In 1985, Statistics Canada undertook a major project with the goal of redesigning all of the Bureau's economic surveys. As part of this Business Survey Redesign Project (BSRP), the development of generalized software is being emphasized, in an attempt to conserve resources and eliminate duplication. In developing the 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 is done at the preliminary editing stage. Only unresolved cases or cases of lesser impact would be passed to the Generalized Edit and Imputation System (GEIS) as a last resort, 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 development of the GEIS software 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 GEIS is quite flexible, supplying most of the options previously available in any given survey-specific system. To ensure that a particular application does not result in a self-contradictory and redundant set of edits, various analytical functions are provided within the GEIS. Moreover, by housing numerous imputation approaches in one system, the GEIS can be used effectively as a tool for evaluating these approaches. By automating the system, the edit and imputation process becomes more objective and reproducible. Using one system and one general strategy allows conformity between surveys, while the production of complete status reports facilitates evaluation of the process.

The paper is divided into seven parts. Section 2 provides an overview of the computing environment of the GEIS. Sections 3, 4 and 5 describe respectively the functions of editing, error localization and imputation. Definitions and concepts are provided followed by a discussion of the main features. The advantages and limitations of the system are considered. The paper concludes with a description of the outlier detection module in Section 6 and a short summary in Section 7.

2. COMPUTING ENVIRONMENT

The GEIS is embedded in the ORACLE Relational Database Management System (Oracle Corporation, 1985). Because ORACLE, and therefore the GEIS, is portable, the user can take advantage of the strengths of various computer 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. Furthermore, files can be moved easily from one environment to another as well as between machines. This is crucial to making the system available to the numerous existing surveys which process their data using very different computer systems and often many systems at a time.

As a result, the potential users of the GEIS must acquire a basic proficiency in ORACLE and the underlying Structured Query Language (SQL). However, relatively few commands are needed in

order to use the GEIS. Once the users become familiar with the computing environment, they will quickly appreciate the facility of data handling and the flexibility offered to them. In fact, because the database can be queried at any time, the users can monitor the edit and imputation process more effectively and thoroughly. In other words, the impact of any module can be assessed almost instantaneously by monitoring frequencies 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.

3. EDITING

The objective of editing is to determine whether a given data record contains invalid, missing, inconsistent or outlying responses. To accomplish this task, the edit component of the GEIS consists of three main parts: specification of edits, analysis of edits and application of edits. At this time the GEIS requires that all the edits be linear (i.e. of the form $aX + bY + \ldots + cZ < d$, where X, Y, Z are data items and a, b, c, d are constants) and all data values positive. Requirements outside these conditions can often be transformed to satisfy them (Kovar, MacMillan and Whitridge, 1988). As such, most of the users' edit requirements can be accommodated by the GEIS, although some edits may have to be specified in different ways. These constraints will be relaxed in future releases of the system.

Since the pattern of edit failures is often of greater importance than the individual edit failures themselves, the edits are always considered in sets. Edits are placed into such sets using the edit grouping facility in GEIS. The need for this facility becomes more evident in the case of larger, more complicated surveys. It has been noted that while the edits are often interrelated, they also tend to naturally cluster. 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.

In order to develop a successful application of the GEIS, the user essentially needs to specify only one thing: a set of conditions that describes a "clean" record. These conditions are specified by means of edits whose purpose is to identify acceptable and unacceptable records. This approach has been adopted successfully in the case of the Census of Population through the use of CANEDIT (Fellegi and Holt, 1976). In the case of economic surveys, however, defining an acceptable region of data points through the use of linear edits may be unfamiliar to users accustomed to the more traditional "detect and correct" approach. A detailed discussion of the uses of linear edits may be found in Kovar, MacMillan and Whitridge (1988) or Fitzpatrick (1988).

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 which 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 the GEIS in order to make the development phase easier.

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. For the most part, this task will be accomplished interactively, usually on a micro-computer. During the edit specification, the system performs some syntax verification including checking as to whether arithmetic operators have been correctly specified, and whether all variables referenced are, in fact, part of the questionnaire. A facility to update the edits, attach comments to the edits, and automatically date the changes is also provided.

Further edit analysis is only possible as a result of the assumption of linearity of the edits and positivity of the data. Linear programming techniques are used to analyze the edit set beyond mere syntax (Sande, 1979). When the Check Edits function is invoked, the 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 redundant edits, if any; that is, edits which do not further restrict the feasible region of data values in the presence of the other edits. By identifying these redundant edits, the system implicitly defines the minimal set of edits.

The system then generates the acceptable ranges for all variables, the extreme points of the feasible region, and the set of implied edits (Sande, 1979). In particular, the Extreme Points module generates a set of records which would pass all edits but which represents the vertices 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.

All 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. They are intended to help the user create a consistent, minimal edit set that describes accurately the variable relationships and constraints. The edits must be derived bearing in mind the intent of the questionnaire, the accounting rules, reasonableness of entries, and the general requirements of the survey. 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.

The GEIS then applies the edits to the data and classifies the records as pass or fail (Giles, 1986b). This information is retained internally to be used by other modules of the GEIS. 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 has been made very easy in the GEIS.

4. ERROR LOCALIZATION

Error Localization is the link between the edit and the imputation phases. It is the process of determining which fields of a record should be changed and/or imputed. Clearly, missing items have to be imputed. However, when a record fails one or more edits, there might be several combinations of fields that could be changed so that the record would pass the set of edits. The GEIS finds all those combinations which will minimize the number of fields to be changed. The user also has the option of assigning weights according to the reliability of the fields and minimizing a weighted number of fields to be imputed. The module thus performs three distinct functions. It finds those sets of fields which if changed would make it possible for the record to pass all the edits. Secondly, it selects the optimal set(s), where optimality is in the Fellegi-Holt sense, i.e. minimal disruption of existing fields. Thirdly, the module chooses one set at random, if several optimal sets were identified.

On the technical side, the error localization problem is recast as a cardinality constrained linear program (Sande, 1979) and is solved using Chernikova's algorithm (Rubin, 1973). The module is relatively self-contained and thus the user does not need to interface with it a great deal. The fields which need imputation are flagged internally, for use by subsequent modules.

5. IMPUTATION

Imputation is the procedure of supplying valid values for those fields of a record that are missing or have been identified for change as a result of the error localization. The new values must 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. 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.

Initially, as the fields to be imputed are identified, the system also checks if there are any fields on the record that are determined uniquely by the edits and the valid data, and performs this imputation. This facility is especially useful in a situation where one achieves a partial imputation and is forced to recycle the record for another try.

In general, the 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 determined based on the reported values. Technically, to ensure that the edits are satisfied, a number of nearest neighbours is found, and the closest one which produces a record that satisfies the edits is used to impute for the record, provided that such a donor exists. If no donor can be found, the record remains unimputed, and an alternate method must be used.

The second category consists of various imputation estimators that replace the missing or invalid values using a model. 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 and regression estimators. Details of the methods and exact formulae may be found in Giles and Patrick (1986) or Giles (1986a).

Note that the donor method 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 is their interrelationship. On the other hand the model based methods are unlikely to preserve the structure of the data as well as donor imputation would (Bureau, Michaud and Sistla, 1986), since none ensure that the edits will be satisfied. For most general applications, donor imputation should perform satisfactorily. 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, may be quite appropriate in the case of monthly surveys. The GEIS allows the user to make these decisions by packaging all the methods together in order to facilitate comparison.

As with editing, imputation results can be monitored based on tabulations generated by the 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. Some of this information will be used by the system to generate quality indicators of the imputed data. Other information can be used by the user to improve the particular application.

6. STATISTICAL EDIT

The system also provides a facility for outlier detection. This module, referred to as the statistical edit, considers all the data records at once and therefore cannot 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 statistical edit is considerably different from the edits which have been discussed so far in that it is 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, the statistical edit 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. As well, the module can be used to flag fields for imputation by the GEIS, or for differential treatment by the estimation phase outside the GEIS. Most importantly, in its univariate form, a derived statistical edit itself can be used as a linear edit in future applications, most appropriately, in the case of monthly surveys.

7. CONCLUDING REMARKS

The Generalized Edit and Imputation System evolved from the work of Fellegi and Holt (1976) and Sande (1979). It is embedded in the ORACLE Relational Database Management System, thus offering a great deal of flexibility and portability. It aims to provide a complete and consistent data set, in preparation for the final stages of survey processing: estimation, tabulation and dissemination. To this end, missing and inconsistent entries are identified using the edits, fields are flagged for change by the error localization module, and the record is cleaned up by imputation. Moreover, the GEIS provides the user with a choice of imputation methods and can thus also be used effectively as an evaluation tool.

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.

Fellegi, I.P. and Holt, T. (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.

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. Presented at the U.S. Bureau of the Census Third Annual Research Conference, March 1987.

Giles, P. (1988). Generalized edit and imputation of survey data. **Statistics Canada Special Issue of the Canadian Journal of Statistics,** to appear.

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., MacMillan, J.H. and Whitridge, P. (1988). Overview and strategy for the Generalized Edit and Imputation System. Statistics Canada, Methodology Branch Working Paper No. BSMD 88-007.

Oracle Corporation (1985). **ORACLE Relational Database Management System.** Oracle Corporation, Menlo Park, California.

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, Phillipines.

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.