A COMPARISON OF EDIT AND IMPUTATION SYSTEMS

Claude Poirier, Statistics Canada
Statistics Canada, 120 Parkdale Ave., 3-I R.H. Coats Bldg., Ottawa K1A 0T6, Canada

Key Words: Data Editing, Imputation, Evaluation

1. INTRODUCTION

For edit and imputation processes, survey managers have to decide whether they will develop custom-made systems or use existing software. Many statistical agencies develop and maintain generalized systems to offer managers basic tools for each survey step, but the choice of a system may sometimes be difficult. Internally developed generalized systems may offer only subsets of the required functionality, and so the managers perhaps have to look for potential systems outside of their agencies, be it for full implementation or simply to look for implementation ideas from other development teams around the world.

The goal of this paper is to evaluate the functionality of four editing and imputation systems. An empirical evaluation would also be interesting but the amount of common functionality across the systems is too limited to do so. The selected systems are the Generalized Edit and Imputation System from Statistics Canada, the New Imputation Methodology also from Statistics Canada, the Standard Economic Processing System from the U.S. Bureau of the Census, and Solas for missing data analysis from Statistical Solutions Inc. The four targeted systems are clearly not an exhaustive set of editing and imputation packages. Other products exist and may be part of future evaluations.

The four selected systems are described in Section 2. An evaluation and comparison exercise is documented in Section 3, elaborating on the systems' respective strengths and weaknesses, their expected future developments and the best use of each system.

2. THE FUNCTIONALITIES

2.1 The Generalized Edit and Imputation System

The Generalized Edit and Imputation System (GEIS) was developed at Statistics Canada to meet the requirements of the Canadian economic surveys. The current version, GEIS v6.5, is usually used after preliminary editing associated with the collection and capture phases and respondent follow-up have been completed. Linear programming techniques are used to conduct the localization of fields to be imputed and search algorithms are used to perform automatic imputations. More details are given in Statistics Canada (1998).

GEIS is usually applied in a step-wise fashion, and its structure facilitates this approach. The steps are edit specification, outlier detection, error localization, and automatic imputation. The first step, the edit specification and analysis, serves to identify the relationships which characterize acceptable records. The relationships are expressed as a set of linear edit rules:

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + \ldots + a_{1m}x_m &\le b_1 \\ \vdots \quad\quad \vdots \quad\quad \vdots \quad\quad &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \ldots + a_{nm}x_m &\le b_n \end{aligned}$$

where the a_{ij}'s and b_i's are user-defined constants, and the x_j's represent the m survey variables.

The second step aims at the detection of univariate outliers using the Hidiroglou and Berthelot method (1986). It identifies outlying observations based on the median and the quartiles of the population.

The third step is the error localization which uses a linear programming approach to minimize the number of fields requiring imputation. This is an application of the rule of minimum change as proposed by Fellegi and Holt (1976). The step identifies the fields that need to be imputed in order for the record to pass all the edit rules. The problem is expressed as a constrained linear program and solved using Chernikova's algorithm, as detailed by Schiopu-Kratina and Kovar (1989).

The final step is the imputation function which offers three imputation methods: Deterministic, Donor, and Estimators. Based on the edit rules, the deterministic imputation identifies cases in which there is only one possible solution that would allow the record to satisfy the rules. The donor imputation replaces the values to be imputed using data from the closest valid record, also referred to as the nearest neighbour. For a given record, a subset of the fields which do not need imputation are automatically used as matching fields, and the maximum standardized difference among these individual fields is used as the distance function. The donor pool includes the observations that satisfy all edit rules. The user can specify post-imputation edits. The imputation by estimators provides a wide set of techniques using historical or current information. Built-in estimators are

available in GEIS: Previous values, previous/current means, trends, and multiple regressions. User-defined estimators can also be specified.

The system works in MVS and UNIX environments. It was developed in C language and currently interacts with Oracle databases. It includes an interface that helps the user in specifying the parameters and edit rules, but the interface is not the easiest one to work with.

2.2 The New Imputation Methodology

The New Imputation Methodology (NIM) is another system developed at Statistics Canada. NIM targets the social surveys because it deals mostly with qualitative variables. The system uses donor imputation as a unique imputation method. As detailed in Bankier et al. (1996), its goal is to minimize the number of changes while making sure the imputation actions are plausible. It always performs record imputation based on a single donor.

NIM is used after the collection and capture editing has been completed. It uses edit rules to identify records that need imputation. A failed-edit record is identified if at least one of the rules is true. The rules are defined through decision logic tables (DLT), using SPIDER, a Statistics Canada package. Table 1 provides a simple example of edit rules for a two-person household. An observation that satisfies either of the two rules will be flagged as a failed-edit record.

Table 1: An example of edit rules

	Edit Rules	
	1	2
Person1 is married	N	
Person2 is married		N
Person2 is the spouse of person1	Y	Y

When failed edit and passed edit records are identified, the system tries to find, for each record to be imputed, a record that can be used as a donor. The search targets a donor coming from the set of passed records and being close to the failed edit record. In this process, the distance between a failed record f and a passed record p is defined as follows:

$$D(f,p) = \sum w_j D_j(f,p)$$

where w_j is a user-defined weight associated with the variable j, and $D_j(f,p)$ is a distance function associated with variable j (this distance function may be different for each variable). In making a choice amongst the records in the donor pool, the system takes into account all feasible actions for each potential donor. A feasible action is the transfer of donor data into a set of recipient's fields such that the newly imputed record, say a, passes the edit rules. NIM will randomly select a donor p and a final action a from the feasible actions which minimize the following composite distance, D_{fpa}, for the failed record f:

$$D_{fpa} = \alpha\, D(f,a) + (1-\alpha) D(a,p) \qquad 0 \le \alpha \le 1$$

where α is a user-defined constant. In this equation, an α close to one would give more importance to the minimum number of changes than to the similarity of the imputed action and the passed record. Variations can be made by accepting not only the minimum D_{fpa} but also some near minimum changes as possible imputation actions.

In practice, the function described above becomes costly to minimize as the number of passed-edit records and potential actions grows. Highly efficient algorithms were introduced to alleviate the potential shortcoming.

The system was developed in the C language and runs in a mainframe environment. It works jointly with SPIDER, a PL-1 program that handles DLTs. The system was used successfully for the 1996 Canadian Census of Population.

2.3 The Standard Economic Processing System

Sigman (1997) describes the basis behind the development of the Standard Economic Processing System (StEPS). The system is to replace 15 existing systems used for U.S. economic surveys. Its development was initiated in 1996 by the U.S. Bureau of the Census to provide integrated tools for the processing of survey steps. As detailed in the system concepts and overview document (U.S. Bureau of the Census, 1996), StEPS is more than just an editing and imputation system. It includes a module to control the collection of information, a data review and on-line correction module, an estimation and variance calculation module, and a tabulation and disclosure module. It can provide general diagnostic tables, including response rates, imputation rates, etc. For the purpose of the present evaluation, the focus is constrained to the editing and imputation modules.

The data editing module of StEPS v1.0 allows simple verifications such as ascertaining the presence of data values for required items, range verifications, and verifications of valid categories. It also provides more complex tests such as balance tests which verify the additivity of items against selected totals, and survey rules to verify field relationships within observations. The edit options offer the basic functionality and should

a complex rule be required, the user can provide his own program statements. Such program coding is made easy by special windows integrated in the menus. In case of edit failures, concurrent users can individually modify reported data in an interactive manner.

StEPS has two modules for imputation, referred to as "simple imputation" and "general imputation". The simple-imputation module performs deterministic imputations and flags the resulting imputed values as if they had been reported. The imputation formulas used by the simple-imputation module are defined by the user through the use of SAS windows. Any group of SAS statements, regardless of their complexity, can be used to define the imputation formula.

The general-imputation module aims to replace with valid values, any invalid values identified in the above editing process. The imputation techniques available in StEPS are mostly estimator type techniques. This includes the imputation by auxiliary data items, sum of data items, historical values, means, trends, ratios, and multiple regressions. All estimator functions can be evaluated from weighted or unweighted data. Similar to GEIS, the system can exclude several types of records from the calculation of estimators. Furthermore, for the ratio and mean estimators, StEPS allows the exclusion of records based on upper and lower bounds U and L.

The prorating transformation represents another imputation action offered in StEPS. The function consists of adjusting every component of a sum in order to obtain a known total. Currently, StEPS can prorate multiple one-dimensional sums that have a common total. Future versions of StEPS will be able to prorate nested one-dimensional sums ($A+B=C$ and $C+D=E$) and two-dimensional sums.

The system is developed entirely in SAS and works in a UNIX environment. A complete graphical user interface is available. The file and variable naming convention eases the processing of historical edit and imputation. The database architecture is based on a data point model. A record includes three basic components: the unit identifier, the field name and the value itself. These "skinny" records are different from the usual "fat" records which include all survey variables. In this architecture, the empty cells are simply dropped, as opposed to the usual architecture. For the user, this translates into a more efficient database where no record layout has to be maintained to port information across processes.

2.4 Solas for missing data analysis

Solas for missing data analysis is produced by Statistical Solutions Ltd., a statistical software company based in Ireland with offices in the United States. Websites for the company can be found at *www.statsol.ie* and *www.statsolUSA.com*. Solas v1.1 was designed for the imputation of missing data, primarily in biostatistical research. Although the documentation claims support for both numeric and ordered categorical variables, the imputation functions are more widely applicable to numeric variables. Solas also includes data analysis tools but these are used in the imputation process rather than being the main feature of the system.

The system does not include any edit function. It simply imputes fields having missing data. Its main feature is the multiple imputation option, a technique developed by Rubin (1978). It also includes the standard hot deck method, and two estimator type imputations, namely the current mean and the historical imputation.

The hot deck imputation attempts to find matching records which are similar to records to be imputed with respect to auxiliary matching variables and precedence rules defined by the user. Exact matches are targeted, and if no records are found, Solas will automatically drop the matching variables one by one given the precedence rules until it has found an exact match. The process is completed when an exact match is identified. Whenever several exact matches are found, one can be selected randomly to impute all the missing fields of the record to be imputed. If absolutely no matches are found, a random selection from the entire pool is possible. Due to the possible rareness of exact matches for continuous variables, these variables can be categorized within Solas prior to imputation.

The multiple imputation is a repetitive execution of an imputation strategy. It can be applied to both cross-sectional and longitudinal data and many imputation parameters can be controlled by the user. Solas will impute several, say M, values for each missing field. The results can be combined to produce overall estimates with variances for the variables of interest. The User reference manual (Statistical Solutions, 1997) describes the theory well.

The current mean is one of the two estimator functions. It consists of imputing the missing values with the basic mean of the other records in the imputation class. For ordered categorical variables, the mode is used. The historical imputation is the second estimator type imputation. It simply imputes the value from the previous data period, with no transformation. The value

is copied as is.

Solas provides the capability to define a weighting variable, referred to as a "case frequency variable". As its name suggests, the weights are defined for each record, as opposed to the weighting concept introduced for GEIS and NIM to put more emphasis on some variables. Thus, a weighted observation will be processed by Solas as repetitions of an original observation.

The Solas system works on IBM compatible personal computers with 80-486 or higher processors. It can read and write data in different formats including ASCII, SAS, DBase, FoxPro Paradox, Excel, Lotus, BMDP, and others. The system can be installed by almost any user, with no support required. Its interface is user-friendly and the help function is quite complete.

3. COMPARISON OF THE SYSTEMS

3.1 Their strengths

The four systems we analyzed can process data by imputation classes or groups. Comparing GEIS with NIM, we note that the first one targets uniquely numeric variables while the second targets mainly qualitative variables. The strengths of GEIS are its capacity to find minimum changes for any set of rules, and its automated donor imputation function driven by the edit rules. This imputation function runs with almost no intervention from the user since it derives the matching fields by itself. The flexible estimator module of GEIS, the several diagnostic reports and the on-line tutorial, coupled with a continuous user support constitute the good aspects of the system.

NIM, on the other hand, finds the donor before it identifies the minimum number of changes needed. Because the minimum changes do not necessarily guarantee a plausible imputation, NIM was developed to meet two objectives at once: to minimize changes and assure plausible imputations. Of the four systems we evaluated, NIM is the only one that includes a generic distance function for the donor imputation. This means the user can define the distance function for each matching field. Its first use for the 1996 Canadian Census was a success with the processing of 11 million households within a month (Bankier et al, 1997).

The StEPS project was initiated following a decision from upper management to build an integrated and standardized product, implemented in SAS, that is to be used by up to a hundred economic surveys. Therefore, the major strength of StEPS is its integration of several survey processing modules: Information management, data review and on-line correction, editing, imputation, estimation, variance calculation, tabulation and disclosure. For the edit and imputation modules, both the survey specifications and the implementations are integrated into the graphical interface. That means, a survey manager can provide his specifications directly through the system. The product standardization resulted in a good file and variable naming convention which simplifies all the processes. The completeness and effectiveness of its set of estimator imputations is comparable to the one offered in GEIS.

Solas presents a good multiple imputation function with many control options for that method. The nice graphical interface of the system represents another good aspect. Solas is easy to install and user-friendly. Its on-line help function is adequate for the functionality Solas provides. Once imputation is completed, a copy of the resulting data sheet appears on the screen. The imputed values are shown in blue, in contrast to the reported values which appear in black. Finally, the small size and the portability of the system makes it very practical. Some empirical evaluations have shown that the system is relatively quick.

3.2 Their weaknesses

The foundation software of GEIS makes the system sometimes "too heavy" to run. Also, a user that built his or her own edit system will in most cases want a direct access to the imputation function. Unfortunately, the current imputation function cannot be run independently from its edit function. GEIS only deals with numeric variables. In the editing process, it assumes each variable takes non-negative values, which is not always true in practice, especially for financial surveys. Pre-processors have to be developed to overcome the problem.

NIM was developed essentially for the Canadian Census which surveys persons within households. In its current form, it may be difficult to reuse NIM for a wide variety of surveys. Although its generalization is being considered, its feasibility has not been demonstrated yet. NIM can process quantitative variables along with qualitative variables, but the performance of the system with more than a few quantitative variables has yet to be demonstrated. Some recent theoritical results, however, suggest this may be feasible.

StEPS does not provide a minimum change functionality nor any other automated error localization module. Thus, for every combination of errors, the user has to specify which fields need to be imputed. Also, StEPS does not offer a donor imputation function. This means the imputation strategy for a brand new survey,

with no historical data nor administrative information, is limited to estimator imputations based on current values. Although the SAS windows in which users specify special rules are practical, a certain level of SAS knowledge is a prerequisite. In practice, this is not always available and thus some SAS training has to be provided in addition to the system-specific training.

As for Solas, the functionality aside from the multiple imputation is very basic. There is no control on the number of required records from which the information is extracted to perform the group mean or the donor imputation. The historical method includes no control on the imputation status of the historical information before its use in the process. Finally, no imputation summary report is produced after the imputation has been completed. Solas was developed for biostatistics, more than for complex surveys where we observe high numbers of variables linked together with complex relationships.

3.3 A subjective comparison

The four processing systems being evaluated can be compared in terms of their functionality. The goal here is to qualify rather than quantify the quality, flexibility, efficiency and reliability of each implementation. For this subjective comparison, a zero to three-stars rating, where three stars represent the best, is used in table 2 below to differentiate the implementations. A three-star (***) rating is given to a function when its implementation offers the sub-functions or options being required by a wide range of survey applications. This does not mean, however, that no improvement can be made to the function. A two-star (**) rating is given to an implementation having a less complete set of options. A one-star (*) rating means the implementation offers a partial functionality. That is, either its assumptions are too restrictive or its options are not generalized enough to make good use of the function. No stars are assigned when the functionality is not offered at all.

In that table, the minimum change refers to the automated identification of the minimum set of variables that need to be imputed. On the other hand, user-defined change consists of the pre-identification of variables to be imputed in case of an edit failure. In the general category, the integration refers to the possibility of using the system within a suite of products that provide other survey functions, like sampling, data collection and capture, estimation, etc. A reusable code is a program that can easily be adapted to any survey, regardless of its collection structure, its database structure and variable names. The portability depends on the platforms and foundation softwares being required to install, compile and run the system. Note that both the size and the cost figures are approximated and dated August 1999.

Table 2: A subjective comparison of systems

Characteristics	GEIS	NIM	StEPS	Solas
Type of variables:				
- Numeric	***	*	***	***
- Qualitative	*	***	**	*
Editing:				
- Data verification	*	*	***	
- On-line correction			***	
- Minimum changes	***	***		
- User-defined changes			***	
- Outlier detection	**		**	
Imputation:				
- Deterministic	***		*	
- Donor (random)	**	**		***
- Donor (closest)	***	***		
- Estimators	***		***	*
- Prorating			***	
- Multiple imputation		*		***
General:				
- User-friendliness	*	**	**	***
- On-line help			***	***
- On-line tutorial	***			**
- Diagnostic reports	***	***	***	
- Integration			***	
- Reusable code	***	**	**	***
- Portability	**	**	**	***
- User support	***			*
Other information:				
- Size of code (Mb)	20	1	200	7
- Cost ('000 US$)	20	--	--	1

3.4 Future developments

GEIS recently entered a major redesign phase. First, independent modules for the edit and the imputation functions are being created in order to ease the direct access to either of the two. Changes to the input and output statements will be made to make possible the interactions with SAS datasets, in addition to the Oracle databases. Finally, the development of new functionality including a prorating function and mass imputation function was recently initiated.

The NIM development team is currently defining the theory to allow a better and more complete processing of numeric variables mixed with qualitative variables. Its generalization is being investigated in order to make the code easily reusable for other surveys. Work was

initiated to move the entire program into C language, to use generic DLTs and to make the system portable by using flat files.

Future versions of StEPS will be able to prorate nested one-dimensional sums and two-dimensional sums. Improvements and some functionality will be added to the edit module. A long-range plan of the StEPS development team is to investigate the possibility of a minimum change function. This may be implemented using the Chernikova's algorithm that GEIS uses.

The future version 2.0 of Solas should be available in 2000. Plans are to include more diagnostics and control functions, at least for the donor and mean imputations.

3.5 The best uses of the systems

It is possible to identify the context in which these systems can better serve survey statisticians. If a small and simple survey is being developed on micro computers, with a tight schedule and budget, Solas may present a good cost/benefit ratio. On the other hand, in the case of large scale business surveys for which long questionnaires and complex field relationships are developed, GEIS or StEPS would be more appropriate. The required functionality is probably the main factor that would differentiate the two. Another factor to consider is the foundation software. Indeed, the existing in-house expertise with C/Oracle or SAS, the cost of these commercial products and the potential benefits of their acquisitions for the working units should be considered. Finally, a social survey that targets persons within households would clearly derive more benefits from NIM then from the other three systems. The possible generalization of NIM may eventually make the system more suitable for all kinds of social surveys and maybe some business surveys.

In summary, the performance of each system depends on the survey requirements: Numeric, or qualitative variables? Automated minimum changes, or user-defined changes? Donor, or estimator techniques? Good support/ high costs, or low support/low costs? . . .

4. CONCLUDING REMARK

In the evaluation of software, we can often say that the more complete the functionality, the less user-friendly the system is likely to be. In practice, we are tempted to forget this rule and to expect a full set of options and controls with a simple and user-friendly interface. When a system grows in complexity, the development of training tools is suggested in order to improve its uses. Also, for systems like StEPS, GEIS or NIM, there is an increasing need for auxiliary skills in SAS, ORACLE/SQL, or C. These auxiliary skills may encompass a better understanding of imputation so that users can better choose imputation options and keep induced errors to a minimum. This may also provide the users with some tricks to adjust input data so it better fits into the available methods or even generate variations of the existing functionality.

REFERENCES

Bankier, M., Luc, M., Nadeau, C., and Newcombe, P. (1996). "Imputing Numeric and Qualitative Variables Simultaneously". Statistics Canada Technical Report, 120 pages.

Bankier, M., Houle, A.-M., Luc, M., C., and Newcombe, P. (1997). "1996 Canadian Census Demographic Variables Imputation". *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

Fellegi, I.P. and Holt, D. (1976). "A Systematic Approach to Automatic Edit and Imputation". *Journal of the American Statistical Association*, 71, 17-35.

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

Rubin, D.B. (1978). "Multiple imputations in Sample Surveys - A Phenomenological Bayesian Approach to Nonresponse". *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

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

Sigman, R. (1997). "How Should we Proceed to Develop Generalized Software for Survey Processing Operations such as Editing, Imputation, Estimation, etc.?". Technical report from the U.S. Bureau of the Census.

Statistical Solutions, (1997). "Solas For Missing Data Analysis 1.0: User Reference". Cork, Ireland, Statistical Solutions Inc.

Statistics Canada (1998). "Functional Description of the Generalized Edit and Imputation System". Statistics Canada Technical Report.

U.S. Bureau of the Census (1996). "StEPS: Concepts and Overview". Technical report from the U.S. Bureau of the Census.