Survey Methodology, June 1986
Vol. 12, No. 1, pp. 49–60
Statistics Canada

Imputation Options in a Generalized Edit and Imputation System

P. GILES and C. PATRICK[1]

ABSTRACT

Statistics Canada has undertaken a project to develop a generalized edit and imputation system, the intent of which is to meet the processing requirements of most of its surveys. The various approaches to imputation for item non-response, which have been proposed, will be discussed. Important issues related to the implementation of these proposals into a generalized setting will also be addressed.

KEY WORDS: Modularity; Prototyping; Donor imputation; Regression models.

1. GENERALIZED SYSTEMS

Due to resource constraints imposed on surveys in recent years, especially in the area of development, the idea of generalized software has received considerable support. By generalized software, it is meant a set of computer programs, tied together into one system, which allows the user to select a suitable approach to the problem, from among several alternatives. For example, a user has a data file from which a sample of records is to be selected. A generalized sample selection system would offer the user the choice of various sampling schemes such as simple random or unequal probability sampling (with or without replacement), systematic, stratified, or cluster sampling.

A genuinely generalized system is, almost by definition, a complex object. The concept of modularity is an important device for the reduction of complexity, by allowing the overall task to be split into a number of simpler sub-tasks. Each of the sub-tasks, or functions, is performed sequentially. The user is offered several alternatives for each sub-task. Therefore, not only is the overall task able to be split into smaller, more manageable components, but also each sub-task can be performed in more than one way.

Figure 1 demonstrates how the edit and imputation task can be split into three sub-tasks. These three sub-tasks are editing, identification of fields to impute, and imputation. Each of the boxes, or modules, in a row employ different approaches to that particular sub-task. For example, C1 could employ some type of donor imputation, C2 could employ the imputation of a mean value, and so on. The user would select one of the modules from each of rows A, B, and C.

It should be noted that this representation of a generalized system for edit and imputation is not the only possibility. In fact, the actual proposal for a developmental project actually contains five sub-tasks, as opposed to the three exemplified here. This representation is given only for simplicity.

Each sub-task, or row in the example, would be a clearly defined function. The input files required, and the output files created, must have prespecified formats. This allows the user to concentrate on the choice of modules in each row, knowing that the system can handle the "housekeeping". (This refers to file handling and other mundane details about which the user would prefer not to worry.) Even though the system may accept all possible combinations of choices of modules, some combinations may not be desirable or even logically valid. It is usually the responsibility of the user to ensure that the pieces fit together.

[1] Philip Giles and Charles Patrick, Business Survey Methods Division, Statistics Canada, Tunney's Pasture, Ottawa, Ontario, Canada, K1A 0T6.

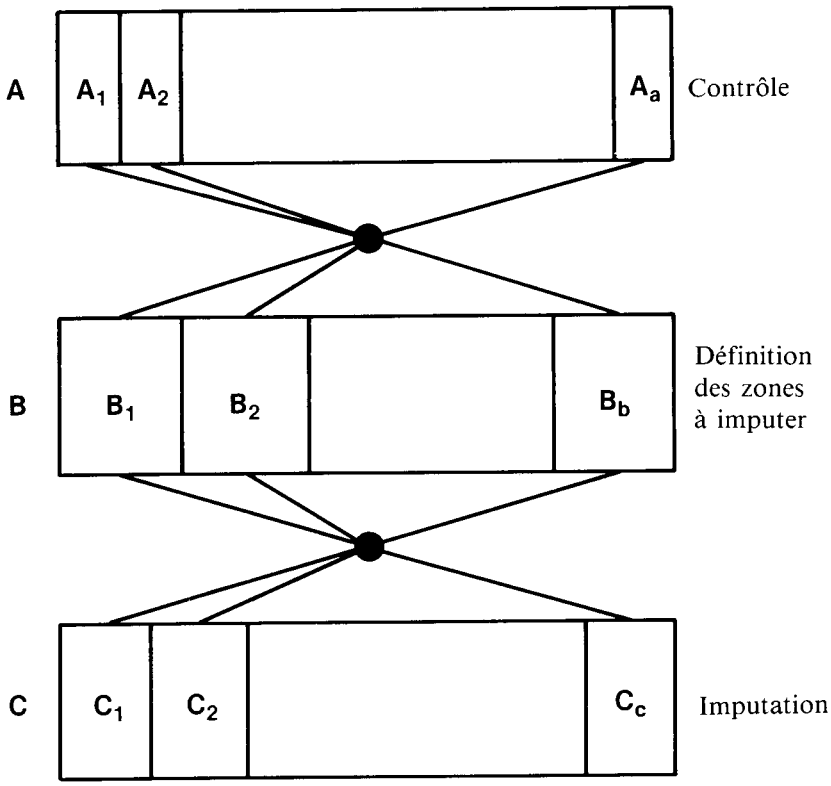

Figure 1. Generalized System Example - Edit and Imputation

A modular approach to the development of a processing system has an important consequence. From a certain point of view, the system is always "under development", since additional modules embodying new approaches and enhancements to "old" modules, can always, and in principle should, be added. This open-endedness also means that the very important concept of prototyping can be easily accommodated. Prototyping is an approach wherein a subset of modules are developed initially. The system would then be available to some of the users. Subsequently, additional modules are developed to meet the requirements of additional users. Thus, the key advantage of prototyping and modularity is that piecemeal improvements to the system are deliberately anticipated and more easily accomplished. A minimal, but imperative, requirement of such an approach is that a framework (as shown in Figure 1) and a host environment (format of data files and programming language) must be carefully defined and specified very early in the overall developmental process.

In addition to the foregoing developmental advantages, others may be gained after the system is in place. The user has considerable flexibility in choosing the path to proceed. If several alternatives seem equally viable, one can use historical data to choose among them, by testing the various alternatives prior to data collection. This can be accomplished without an undue expenditure of effort. Once the generalized system is developed there is a reduction in resource requirements for each of its users, with a corresponding reduction in elapsed time to implementation.

There are some disadvantages to following a generalized route. The utilization of generalized software in a production environment may be less efficient than the corresponding custom-designed system. The initial resource requirement will be higher for a generalized system as compared to a customized system. However, this higher cost must be assessed against the

substantially higher costs of repeated custom-designed implementations. Nor is it reasonable to expect a generalized system to satisfy every specific requirement. In this situation, the user has two options. The first option is to develop a user-written module. This would not require the same degree of effort as a complete customization. However, if this occurs frequently, the purpose of the generalized system is defeated. The second option is for the user to modify the specifications in order to fit the generalized system mold. If the system has been well-designed, any required compromise should not result in a serious deterioration of data quality. It should also be recognized that compromises to the original specifications are usually and frequently required during the development of a customized system.

2. BACKGROUND TO IMPUTATION

The term "imputation", in this document, refers to a certain class of procedures for handling non-response. The input is a data captured file. The imputation procedure creates a file with individually "clean" records; a "clean" record being one which has no missing values and which satisfies all the specified edits. In order to create a clean record, a value must be estimated for each missing value.

The edits, specified by the user, are logical constraints on the values that each variable can assume. The set of edits, as a whole, define the acceptance region for the data. For categorical data, an edit is specified as a set of combinations of acceptable data values. The acceptance region can be represented as a set of lattice points in N-space. For numerical data, an edit is a linear equality or inequality. The requirement of linearity is not unduly restrictive, since a non-linear edit can be made linear by either algebraic manipulation or by adding supplementary variables, which are suitably defined non-linear functions of survey variables. The acceptance region for numerical data is a set of convex regions in N-space. The reason that there may be more than one convex region is that conditional edits are possible. Conditional edits are edits which pertain to only a subset of records. For example, the edits which are relevant to a particular record may be very different, depending on whether the variable Sex is recorded as Male or Female.

If one or more edits fail for a particular record, it may not be obvious which variable(s) is/are in error, and, by implication, to be imputed. For example, a failed edit is $A + B \leq C$. The data record under consideration has data values $A = 10$, $B = 5$, $C = 12$. There are seven combinations of variables to change which would result in a clean record. These are A, B, C, A & B, A & C, B & C, and, A & B & C. Without any other information or decision rule, each of these choices is equally valid. The problem of how to decide which variable(s) to impute will not be discussed in this document. It will be assumed that, for each record, the variable(s) to impute have been identified. No distinction is made between variables to impute due to missing values and variables to impute due to edit failures.

3. PROPOSED IMPUTATION TECHNIQUES

This section is comprised of four sub-sections, which define all the proposed imputation techniques. These are Deterministic Imputation, Donor Imputation, Regression Models, and Other Imputation Estimators. The use of regression models and the section on other estimators is restricted to numerical data. The other two sub-sections apply both to numerical and categorical data.

Almost all imputation techniques can be formulated in a prediction framework, described by Rubin (1976), as follows. A joint distribution, $f(X_1, \ldots, X_N)$, summarizing the

statistical behavior of the population of complete records is specified. This can be done whether the individual variables are quantitative or qualitative. Without loss of generality, for a record i which requires imputation, the N variables can be partitioned into $X_1, \ldots, X_{m_i}$, which require imputation, and $X_{m_i+1}, \ldots, X_N$, which do not require imputation. A conditional distribution $f(X_1, \ldots, X_{m_i} \mid x_{m_i+1}, \ldots, x_N)$ can be derived. Imputed values, $y_1, \ldots, y_{m_i}$, are chosen for $X_1, \ldots, X_{m_i}$ from the set.

$$\{y_1, \ldots, y_{m_i} : f(y_1, \ldots, y_{m_i} \mid x_{m_i+1}, \ldots, x_N) > 0\}$$

Various selection mechanisms can be employed. However, as stated above, some of these are relevant only to certain types of data variables.

It should be noted that there is nothing new or radically different in these proposals. They are based on work done previously, both in Statistics Canada and outside. The discussion on donor imputation is based on Fellegi and Holt (1976). The model-based approach to determining a value to impute is discussed by Little (1982). Other related papers of interest are Sande (1976), Kalton and Kasprzyk (1982), and Kalton and Kish (1981).

3.1 Deterministic Imputation

The first type of imputation is called deterministic imputation. This occurs when only one value can satisfy the edits. If more than one variable is to be imputed for a particular record, a deterministic solution may be possible for some, or all, variables. The check for determinacy should be done before proceeding to other imputation procedures.

Deterministic imputation may arise in very simple, and easily detectable situations. For example, suppose that there is an edit $A + B = 10$. The record under consideration requires A to be imputed and B has value 6. Obviously, $A = 4$ is the only value which will satisfy the edit. Another example demonstrates this for categorical variables. Suppose an edit is stated as "If the relationship to the household reference person is wife, then sex must be female." If the reference record has "wife" as the value of "relationship to the household reference person", and the variable "Sex" requires imputation, then the only valid imputed value is Sex = Female.

However, a typical survey situation will have several edits, rather than just one. This may mean that an existing deterministic solution may not be apparent. The procedure for checking for deterministic imputation is to find the reduced acceptance region defined by the active edits and the "good" data values. The active edits are defined as the subset of edits in which the variable(s) to be imputed are participant. This can also be expressed in the notation of the prediction framework given at the beginning of Section 3. The conditional distribution $f(X_1, \ldots, X_{m_i} \mid x_{m_i+1}, \ldots, x_N)$ will specify a unique value for some or all of the variables $X_1, \ldots, X_{m_i}$.

An example serves to illustrate the procedure for identifying deterministic imputation. Note that while the example is written with numerical variables, an analogous situation exists for categorical variables.

There are three edits:

$$X + Y \leq 16,$$

$$Y + Z \leq 4,$$

$$X - 3Z \leq 8.$$

The reference record has values

$$X = 11 \text{ and } Y = 3.$$

The variable Z is to be imputed.

It is not apparent whether or not a determinancy exists. This first step is to consider all active edits. In the example, there are two edits which contain the variable Z.

$$Y + Z \leq 4,$$

$$X - 3Z \leq 8.$$

Next, the known values of X and Y are inserted into these edits, and the reduced acceptance region is determined.

$$3 + Z \leq 4,$$

$$11 - 3Z \leq 8.$$

Solving these inequalities gives the following solution.

$$Z \leq 1,$$

$$Z \geq 1.$$

It is now obvious that $Z = 1$ is the only possible valid imputed value.

In most "real-life" situations, the incidence of deterministic imputation should be low. The contrary would indicate that the edits are more restrictive than necessary or desirable, and should lead to a re-examination of the edit specifications. However, in the sense that it reduces the imputation problem, deterministic imputation is a useful first step.

3.2 Donor Imputation

Donor imputation is a method which pairs each record requiring imputation, the candidate record, with one record from a defined donor population. In order to determine the value to impute, one approach is to directly copy the value from the donor record onto the candidate record. For numerical variables, if suitable auxiliary information is available, more complex methods may be used to determine the value to be imputed. Further discussion on imputation estimators for donor imputation is given in Section 3.3.

Usually, the donor population is defined as all records in the current survey which have no variables to be imputed. Referring to the prediction framework described at the beginning of Section 3, then this situation implies that $f(X_1, \ldots, X_N)$ is the empirical probability function. However, other approaches to defining the donor population are possible. For the remainder of the discussion on donor imputation, it will simply be assumed that a donor population has been defined.

Donor-candidate pairs are formed using matching variables. Matching variables are defined as variables which do not require imputation on the candidate record and are "highly correlated" with the variable(s) requiring imputation. Preferably, the matching variables should also have "low correlation" with each other. Two matching variables with "high correlation" would have the same discriminatory power as one alone, but would have the effect of doubling the weight given to one alone.

For categorical variables, a donor record is chosen, using some random process, from amongst potential donor records having the same values for the matching variables to those for the candidate record. Since numerical variables can assume many more values than categorical variables, it is very unlikely that an exact match on matching variables would be possible. Therefore, for numerical data, a distance function is used to define similarity. This distance function is a function of the matching variables on the candidate and potential donor records. The chosen donor is the record with minimum distance from the candidate record. Usually, the matching variables are transformed for the purpose of distance calculations in order to remove the effect of scale in which the variable is recorded. For example, it would be quite worrisome to the user if the formation of the donor-candidate pairs was dependent on whether a length variable was recorded in metres or feet. The proposed transformations and distance functions are discussed below.

The matching variables to be used can be a user input, or determined by an automated procedure. Usually, due to time considerations, all decisions must be made prior to data collection. Therefore, if the determination of matching variables is a user input, the user must specify the matching variables for each pattern of variables to be imputed. If there are N variables on the file, the user must make $(2^{**}N) - 2$ input specifications. Obviously, the value of N does not have to be very large in order for this approach to become unmanageable. In order to reduce this number, the matching variables may be specified by stratum. All candidate records in a particular stratum would use the same matching variables. In this situation, it is possible (depending on how careful the user is in specifying the matching variables) that a particular candidate record may have a matching variable which requires imputation. All in all, the user who inputs the matching variable specifications, is warned that this decision may result in a large increase in the work required.

One possible approach for automatically determining the matching variables is proposed. This procedure can be used, analogously, for both categorical and numerical data. Basically, the procedure is as follows. At a minimum, the set of matching variables must contain the variables sharing in the edit rules with the variables to be imputed. As defined earlier, these are the active edits. This approach seems intuitively reasonable, since it is desirable that the matching variables be correlated with the variable(s) to be imputed. The variables in the active edits constrain the range of possible values to be imputed. This implies a type of dependence, or correlation structure.

The use of this matching procedure, together with direct transcription, has one important consequence for categorical variables. All imputed values are guaranteed to pass the edits. This is very important as it is required in order to create a clean record. Without this guarantee, the user must re-edit the records, and possibly adopt a secondary imputation procedure. For numerical data, similarity as defined by a distance function does not guarantee this outcome. However, the closer the distance between the donor and candidate record is to zero, the greater the probability that the imputed values will satisfy the edits.

The determination of matching variables using this automated procedure can be illustrated by an example.

There are five edits:

$$\text{I.} \quad A + B \leq \alpha_1,$$

$$\text{II.} \quad B - E \leq \alpha_2,$$

$$\text{III.} \quad C + 2D + 3E \leq \alpha_3,$$

$$\text{IV.} \quad A + C + D \leq \alpha_4,$$

$$\text{V.} \quad A - 2B + C \leq \alpha_5.$$

There are five survey variables A, B, C, D, E and α_1, α_2, α_3, α_4, α_5 are known scalars. The candidate record under consideration has variable B only to be imputed.

The first step is to identify the active edits. In this example, there are three active edits. These are edits I, II, and V.

The second step is to determine the active variables. The active variables are defined as all variables which are contained in at least one of the active edits. In the example, there are four active variables: A, B, C, E. Note that, by definition, the active variables contain all variables to be imputed.

The third step is to determine the matching variables, as those active variables which do not require imputation. For this example, the matching variable are A, C, E.

In addition to the determination of matching variables, donor imputation for numerical data requires the choice of a data transformation and the choice of a distance function.

Two types of data transforms are proposed. For both of these, each variable is to be transformed independently. The two proposed transformations are a rank value transform and a location-scale transform.

For the rank value transform, the values for each variable are sorted. Then, the rank values are divided by a suitable constant such that all values are in the range from zero to one. The transformed values are distributed uniformly in that range.

The location-scale transform is of the form,

$$y^T = \frac{1}{b}(y - a),$$

where y^T is the transformed value,

y is the original data value,

a, b are user-specified parameters.

Two popular choices for these constants are, one, that a be the sample mean and b be the sample standard deviation, and, two, that a be the sample minimum and b be the range of values in the sample. Other options may be possible.

In choosing a data transform, there are robustness and outlier considerations. The rank value transform is very robust against changes in data values, and pulls outliers closer to the other data values. This may or may not be desirable. There are no bounds on the transformed values, using the location-scale transform with the mean and standard deviation. These parameters are also sensitive to outliers. The choice of the minimum value and range would restrict the transformed values between zero and one. However, these are very sensitive to extreme values. One very large value could cause all of the transformed values, except one, to be virtually zero.

In considering the choice of distance function, a family of distance functions are proposed. These are the weighted $\mathcal{L}^p$ norms, where p is a user-specified constant. The general form of these functions is

$$D(X, Y) = \left[\sum_{k=1}^{r} w_k |x_k - y_k|^p \right]^{1/p},$$

where x_k, y_k are the r matching variables on the two records,

w_k are user-specified weights,

p is a user-specified constant.

The weights are used if one wishes some of the matching variables to contribute more to the distance calculation than others. The default values are for all weights to be set to one.

Three particular choices of a value for p are of special interest, $p = 1$, $p = 2$, and $p = \infty$. For $p = 1$, this function calculates the city block distance. For $p = 2$, the Euclidean distance is calculated. The limiting case of this function, when $p = \infty$, yields the minimax distance. For this choice of p, the function is written as

$$D(X, Y) = \underset{1 \le k \le r}{\text{Max}} [w_k |x_k - y_k|].$$

One final point to be discussed about donor imputation is the concept of a "penalty" for donor usage. This penalty would reduce the number of times that a particular donor record is used. For donor imputation of categorical data, a donor record is selected from the donor population without replacement. This strategy has to be modified slightly if the size of the candidate population is greater than the size of the donor population.

For numerical data, the distance function is modified by increasing the distance calculation according to the number of times a particular donor is used. One possible approach is to use $D'(X, Y)$ to calculate distances, where

$$D'(X, Y) = D(X, Y) \times (1 + ud),$$

where u is the "penalty" imposed by the user,

d is the number of times that donor record has been chosen.

An implication of the imposition of a penalty on the distance function, is that the choice of a donor record for each candidate record is now dependent on the order of the candidate records.

3.3 Regression Models

This section discusses imputation estimators which result from the use of regression models. For this discussion, only two models are used. These are:

MODEL I : $y_i = \alpha + \epsilon_i$, $\quad \text{Var}(\epsilon_i) = \sigma^2$,

MODEL II: $y_i = \beta x_i + \epsilon_i$, $\quad \text{Var}(\epsilon_i) = \sigma^2 x_i$.

Note that these models are special cases of the more general formulation of regression models, which has the form

$$\underset{\sim}{y} = \underset{\sim}{X}\underset{\sim}{\beta} + \underset{\sim}{\epsilon},$$

$$\text{where } E(\underset{\sim}{\epsilon}) = \underset{\sim}{0}, \; V(\underset{\sim}{\epsilon}) = \underset{\sim}{V}$$

Model II is used when auxiliary data is available. Otherwise Model I is used. Both models have one parameter to be estimated. Using least-squares, the parameter estimates are:

$$\hat{\alpha} = \bar{y},$$

$$\hat{\beta} = \frac{\bar{y}}{\bar{x}}.$$

Before stating the various proposed estimators, some notation will be introduced.

Let t be the subscript for time t, the present survey,

y_{it} be the variable under study for unit i and time t; this is the value to be imputed for candidate records,

x_{it} be the auxiliary variable (correlated with Y) for unit i and time t,

R be the subscript for all non-respondents at time t (i.e., y_{it} is known),

NR be the subscript for all non-respondents at time t (i.e., y_{it} is to be imputed),

C, D be superscripts which denote either a candidate or donor record, whenever the distinction is required.

Several explanatory notes are required along with the notation. First, R and NR are as defined in the current survey, regardless of the reporting history of each record. Second, the values for the variables $y_{i(t-1)}$, x_{it}, $x_{i(t-1)}$ may themselves have been imputed. The only restriction is that they are not missing. Third, the notation does not include the concept of imputation classes. Imputation classes are essentially post-strata, in that they define sets of records which are judged homogeneous within, and heterogeneous between groups. However, both the notation and the imputation estimators are readily extendible to include imputation classes.

Thus, estimators can be classified according to:

(i) the choice of model, I or II,

(ii) the imputation group, and,

(iii) the variables in the regression used to estimate the parameter.

The data on the records in the specified imputation group are precisely the data used to estimate the parameter(s) in the model. This concept allows considerable flexibility. For example, it could allow the preclusion of outliers from the calculation of the parameter estimate. After the parameter is estimated, it is used for prediction purposes to determine the imputed value. According to the notation, Y_t is always the variable predicted.

Based on the two models, eight imputation estimators are proposed. Even though there are eight proposed estimators, this list can be augmented in the future. These additional estimators could be derived, for example, by choosing other models, possibly incorporating more variables.

Scanning the list of eight, one can see that these are the familiar imputation estimators that have been used traditionally.

Estimator 1: The value from the previous survey for the same unit is imputed. $y_{i(t-1)}$

Estimator 2: The mean value from the previous survey is imputed. $\bar{y}_{(t-1)}$

Estimator 3: The mean value of all respondents to the current survey is imputed. $\bar{y}_{tR}$

Estimator 4: The value is copied directly from the donor record to the candidate record, y_{it}^{D}

Estimator 5: A ratio estimate, using values from the current survey is imputed.
$$\frac{\bar{y}_{tR}}{\bar{x}_{tR}} x_{it}$$

Estimator 6: A ratio estimate, based on values on the donor and candidate records is imputed.
$$\frac{y_{it}^{D}}{x_{it}^{D}} x_{it}$$

Estimator 7: The value from the previous survey for the same unit, with a trend adjustment calculated from an auxiliary variable, is imputed.

$$\frac{y_{i(t-1)}}{x_{i(t-1)}} x_{it}$$

Estimator 8: The value from the previous survey for the same unit, with a trend adjustment calculated from the change in reported values to variable Y, is imputed.

$$\frac{\bar{y}_{tR}}{\bar{y}_{(t-1)R}} y_{i(t-1)}$$

It is interesting to contrast the difference in estimators when one fixes all classification items but one. For example, the difference between estimators one and two is due only to the difference in choice of imputation group, as is also the case for estimators three and four, and, estimators five and six. The difference between estimators one and seven is due only to the choice of model. The same is true for estimators three and five, and, estimators four and six. It should also be noted that estimators four and six are those used in donor imputation, which were discussed in Section 3.2.

3.4 Other Imputation Estimators

The choice of imputation techniques is dependent upon the assumptions made by the user about the non-responding population. When using donor imputation, one assumes that there are some respondents which are similar to each non- respondent. If one imputes the mean from the current survey, the assumption is that the mean value of the respondents is the same as the mean value of the non- respondents. Similarly, one can go through all the estimators and list the implied assumptions. The first estimator proposed in this section tries to ease the somewhat restrictive (and usually untrue) assumptions required in the previous section. It pays for this by being more complex. It is called the chain-link estimator, given by Madow and Madow (1978).

The derivation of this estimator is described. First, by assuming that the rate of change (trend) of the non-responding and responding populations are the same as observed in the previous survey, the population mean of the variable Y for the non-responding population in the current survey is estimated.

$$\bar{y}_{NRt} = \frac{\bar{y}_{NR(t-1)}}{\bar{y}_{R(t-1)}} \bar{y}_{Rt}.$$

One then determines the imputed value according to the auxiliary variable.

$$y_{it} = \frac{\bar{y}_{NRt}}{\bar{x}_{NRt}} x_{it}$$

$$= \frac{\bar{y}_{NR(t-1)}}{\bar{x}_{NRt}} \quad \frac{\bar{y}_{RT}}{\bar{y}_{R(t-1)}} x_{it}$$

Note that this amounts to a more complex application of the Regression Model approach discussed in Section 3.3. First, temporarily impute $y_{it} = \bar{y}_{NRt}$, as given above. Then, use Model II, and define the imputation group as being all non- responding records to the present survey for variable Y. The response variable is Y_t. The regressor variable is X_t. The resulting estimator is as given above.

The second estimator proposed in this section can be used when one has data on variable Y for several previous surveys. It does not use auxiliary variables, or data from other records. The behavior of each non-respondent is considered independently of others. This method is called exponential smoothing. It is a standard econometric forecasting technique. There is one user-specified parameter. It allows the flexibility of changing the relative contribution of the various data values. Algebraically, the estimator is given by

$$y_{it} = \frac{1-A}{1-A^t} \sum_{r=0}^{t-1} A^r y_{i(t-r-1)},$$

where $0 < A < 1$, is prespecified.

The closer A is to zero, the more weight is given to recent data. If $t = 1$, this reduces to imputing the value for the previous survey.

4. PAST WORK IN STATISTICS CANADA

Statistics Canada has made efforts in the past to develop a generalized edit and imputation system. Two of these will be highlighted, as they form the basis for the current proposal. These are the CAN-EDIT system and the Numerical Edit and Imputation System (NEIS).

4.1 CAN-EDIT

CAN-EDIT is itself, not a completely generalized system. However, the methodology that it employed is. The system is based on the work by Fellegi and Holt (1976) on imputation for categorical data. It was developed for processing the 1976 and 1981 Canadian Censuses of Population and Housing.

CAN-EDIT adopted a donor imputation approach. The matching variables were determined automatically, using the procedure described in Section 3.2. The CAN-EDIT system employed what it called primary and secondary imputation. If a candidate record could not be imputed in primary imputation, it was sent to secondary imputation.

In primary imputation, all imputed values are taken from the same donor. The matching variables were determined based on all variables to be imputed. A record would fail primary imputation if no donor record had identical values on the matching variables.

In secondary imputation, each of the variables to be imputed are treated independently and sequentially. The procedure for determining the matching variables is the same. However, by considering only one variable at a time, the number of matching variables will, in general, be less than under primary imputation. (There cannot be more, but the number may be the same). This implies that the potential donor population is larger. There are a few disadvantages to secondary imputation, as compared to primary imputation. First, it is possible to choose, as a matching variable, a variable which is to be imputed. There is no value to match on. Second, this approach does not make use of the joint distributions of the variables. The imputed values for two variables may satisfy the edits, each may be a very valid value, but which may occur in the population in combination only rarely.

4.2 Numerical Edit and Imputation System (NEIS)

The NEIS is a first prototype of a generalized E&I system for numerical data. It was written as a set of modules in the PSTAT statistical package. Subsequent prototypes have never

been developed. This system was developed by Gordon Sande (1979). It is felt that the methodology is very sound, and should be incorporated in a new system. However, PSTAT may no longer be a suitable software environment. The NEIS was used, in a production environment, by the 1981 Farm Energy Use Survey. The methodology was employed in the development of the 1981 Census of Agriculture processing system.

The NEIS, similar to CAN-EDIT, used a donor imputation approach with matching variables determined automatically using the procedure described in Section 3.2. However, as explained in that section, the determination of matching variables in this fashion for numerical data will not always result in the imputation procedure producing a clean record. The strategy adopted to reduce this problem is to select the closest r donors. If the closest donor does not impute values which satisfy the edits, then the next closest donor is considered, and so on.

The NEIS gave the user no choice of transformation or distance function. It used the rank value transformation and the weighted $\mathcal{L}^\infty$ norm for distance calculations.

5. CONCLUSION

The proposals presented would allow considerable choice to a user of a generalized edit and imputation system. As mentioned, it does not close the door on additional approaches. However, it is felt that a system which is developed with these components would be suitable for a large number of users. It has been the experience of the authours that the ultimate power and usefulness of such a system is not apparent until one starts to use it. As testing proceeds, it becomes clear that there are more capabilities and extensions than first appear.

REFERENCES

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

KALTON, G., and, KASPRZYK, D. (1982). Imputing for missing survey responses. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 22-31.

KALTON, G., and KISH, L. (1981). Two efficient random imputation procedures. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 146-151.

LITTLE, R.J.A. (1982). Models for non-response in sample surveys. *Journal of the American Statistical Association*, 77, 237-250.

MADOW, L.H., and MADOW, W.G. (1978). On link relative estimators. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 534-539.

RUBIN, D.B. (1976). Inference and missing data. *Biometrika*, 63, 581-592.

SANDE, G. (1976). Searching for numerically matched records. Technical Report, Business Survey Methods Division, Statistics Canada.

SANDE, G. (1979). *The Numerical Edit and Imputation Subsystem for PSTAT — A User's Guide.* Research and General Systems Subdivision, Statistics Canada.