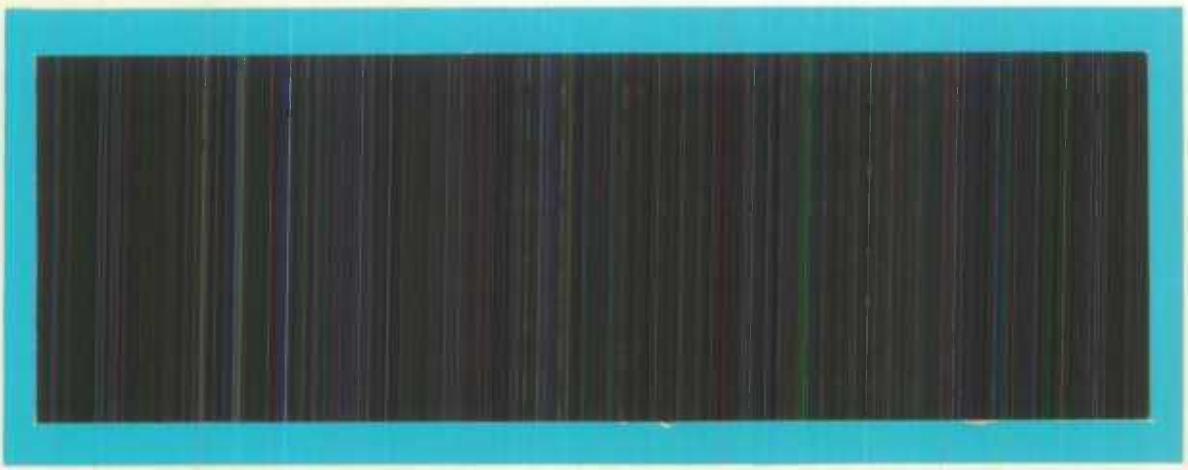
Statistics Canada | Statistique Canada



Methodology Branch

Social Survey Methods Division

Direction de la méthodologie

Division des méthodes d'enquêtes sociales

Canada

WORKING PAPER NO. SSMD-94-001-E
METHODOLOGY BRANCH

CAHIER DE TRAVAIL N0. DMES-94-001-E
DIRECTION DE LA MÉTHODOLOGIE

Use of the Generalized Edit and Imputation System for The Survey of Family Expenditures: A Feasibility Study

by

H. Bérard
Social Survey Methods Division
January 1994

ABSTRACT

The Generalized Edit and Imputation System (GEIS) was used to edit and impute the variables listed in the Family Expenditures Survey (FAMEX) balance sheet. The balance sheet is produced during the verification procedure of the questionnaire and contains data summarizing the household financial status. The GEIS application was evaluated to determine (1) if the imputation process could perform the balance edit i.e., ensure that the inflow and outflow of money within an imputed balance sheet are within 15% of each other, and (2) if the process yields acceptable estimates at the city level for the 31 variables listed in the FAMEX balance sheet.

The imputation of the FAMEX balance sheet can not be considered a typical GEIS application; there was only one predefined edit rule at the balance sheet level: the balance edit. All matching fields used in the distance calculation of the donor imputation module, as well as new edit rules, were determined based on existing relationships among variables (defined through cluster and correlation analysis). Imputation was successful for 97 of the 105 (92%) recipient balance sheets. Eight balance sheets remained unresolved and required manual intervention. Estimates of total for all but two imputed variables (for which matching fields were not defined) differed by less than 5% from their corresponding FAMEX total at the city level. It should be noted that in this application imputed values may not be consistent with non imputed values whenever edit rules involving the variables in question could not be determined.

RÉSUMÉ

Le Système Généralisé de Vérification et d'Imputation (SGVI) a été utilisé afin d'imputer des données correspondant au bilan des dépenses des familles qui est produit lors du processus de vérification du questionnaire de l'Enquête sur les dépenses des familles. Le but de cette procédure était (1) d'évaluer si le système d'imputation permettait de résoudre les cas problèmes où les entrées et sorties d'argent excédaient une différence de 15% dans un bilan financier, et (2), d'évaluer l'impact de cette procédure sur les totaux obtenus au niveau d'un centre urbain pour les 31 variables contenues dans le bilan financier.

L'imputation du bilan financier ne peut être considérée comme une application typique du SGVI; une seule règle de vérification était disponible et de plus, cette règle incluait toutes les 31 variables présentes. Les variables d'appariement utilisées dans la fonction de distance du module d'imputation par donneur, ainsi que des nouvelles règles de vérification, ont été déterminées suite à des analyses de groupements et de corrélations.

Quatre-vingt-quinze des 105 (92%) bilans financiers requérant un traitement ont pu être résolus par le système. Huit bilans financiers n'ont pu être résolus par le SGVI, ainsi ils nécessiteraient une imputation manuelle. À l'exception de deux variables (pour lesquelles aucune variable d'appariement n'a pu être déterminée), une différence inférieure à 5% a été observée entre les totaux obtenus après le traitement par SGCI et les totaux correspondant aux données "finales" de l'Enquête sur les dépenses des familles. Il est important de souligner que la concordance entre les valeurs imputées et non imputées dans les bilans financiers ne sera pas nécessairement respectée lorsque des règles de vérification, assurant la concordance entre ces variables, n'ont pu être déterminées.

TABLE OF CONTENTS

PAGE

1. INTRODUCTION 1

2. METHODS 2
   - 2.1 Famex Edit and Imputation Process 2
   - 2.2 Study Objectives 4
   - 2.3 Study Database 5
   - 2.4 GEIS Overview 6
   - 2.5 Application of GEIS to FAMEX. 7
   - 2.6 Description of Imputation Process 8
     - 2.6.1 Edit Rules 9
     - 2.6.2 Relative Importance of Variables (Weights) 10
     - 2.6.3 Distance Calculation: Choice of Variables 11

3. DATA ANALYSIS 13
   - 3.1 Overall Performance of Imputation Process 13
   - 3.2 Impact of Imputation Strategy: Imputed Values 14
     - 3.2.1 Imputation of 'Mortgage' (C7) 15
     - 3.2.2 Imputation of 'Amount-of-Interest-and-Principal-Paid' (C9) 18
     - 3.2.3 Imputation of 'Other-Vehicle-Transportation-Cost' (BZ2) 19
     - 3.2.4 Imputation of 'Transportation-Cost' (AZ4 and AZ5). 20
     - 3.2.5 Imputation of 'Shelter-Expenses' (A5) 21
     - 3.2.6 Imputation of 'Change-in-Loan-and-Debts' (CZ2D13) 21
     - 3.2.7 Imputation of 'Change-of-Assets' (C1D4) 22
   - 3.3 Comparison of Estimates at the City Level 22
     - 3.3.1 Imputed by GEIS 23
     - 3.3.2 Imputed by FAMEX. 24

4. CONCLUSIONS 25

ACKNOWLEDGEMENTS 27

APPENDIX 1. DESCRIPTION OF VARIABLES 28

APPENDIX 2. LIST OF EDITS 35

APPENDIX 3. WEIGHTS 36

APPENDIX 4. CLUSTER ANALYSIS 37

LIST OF TABLES

PAGE

Table 1. Percent Change Between GEIS Imputed Values and FAMEX Standard Values 15
Table 2. Impact of the Imputation Process on 'Mortgage' (C7). 16
Table 3. Effect of Matching Fields on Imputation of 'Mortgage' (C7). 17
Table 4. Impact of Different Set of Matching Fields on Imputation Process of 'Amount- -of-Interest-and-Principal-Paid' (C9) 18
Table 5. Impact of Other Missing Variables on Imputation Process of 'Other-Vehicle-Transportation-Cost' (BZ2). 20
Table 6. Impact of Imputation Process for Variables Without Defined Matching fields . . 22
Table 7. Comparison of Estimates at the City Level 24
Table 8. Percent Change ((Standard - Test / Standard) *100) at City Level for Variables that Were Never Modified by GEIS but Were Modified by FAMEX 25

LIST OF FIGURES

PAGE

Figure 1. Current Famex Edit and Imputation Process. 3
Figure 2. Creation of database: Data Processed with GEIS (Test Data), and Standard Data. 6
Figure 3. Overview of Imputation Process used with GEIS 8
Figure 4. Description of the Imputation Process 14
Figure 5. Relationships Between Variable AZ5 (Transportation cost) and Variables AZ31 (Clothing and Personal Care Expenditures), AZ32 (Running the Home and Food Expenditures) and A11 (Household Furnishing and Equipment Expenditures) 21

1. INTRODUCTION

This study was initiated as part of the Methodology and Operational Reviews to Seek Efficiency (MORSE) process. The Generalized Edit and Imputation System (GEIS) was considered to have the potential to improve data quality and reduce the level of manual processing that is currently carried out in the Family Expenditures Survey (FAMEX) Edit and Imputation Process. GEIS has been developed as part of the Generalized Survey Function Development Project of Statistics Canada. It is a collection of modules which can be put together in order to satisfy the edit and imputation requirements of a survey [1]. It was successfully used the for the Census of Agriculture which constitutes the largest application of GEIS in both number of variables and records processed (280,000 agricultural holdings collecting information for some 300 variables) [2].

FAMEX provides estimates of the distribution of household expenditures and provides a measure of the economic well-being of the Canadian population [3]. More than 1600 variables are covered by the FAMEX 45 page questionnaire. Another important application of FAMEX is to monitor and periodically update the weights used in the construction of the Consumer Price Index, commonly known as CPI.

An initial proposal suggested that the study be restricted to one section of the FAMEX questionnaire. However after discussion with FAMEX subject matter staff, the focus of the project was changed to the balance edit, a step which is difficult to apply and requires considerable resources. Due to processing limitations it was also decided that the application of GEIS be restricted to the variable aggregates created at the balance sheet level. Note that these are different from the variables at the questionnaire level and are obtained by taking the appropriate sums of the variables in the questionnaire. Henceforth, the term variable means variable at the balance sheet level and not at the questionnaire level. The question addressed is: can GEIS produce acceptable estimates at the city level for the 31 variables listed in the balance sheet while ensuring that imputed balance sheets are not out of balance by more than 15% (i.e. ensure that the inflow and outflow of money within an imputed balance sheet are within 15% of each other).

2. METHODS

In this section, the current FAMEX Edit and Imputation process is discussed as well as the approach used in this study. Specifically, we present (1) a brief description of the current FAMEX Edit and Imputation Process, (2) the objectives of the study, (3) the database, (4) a GEIS overview, (5) how GEIS was applied to the FAMEX balance sheet, and (6) a detailed description of the imputation process.

2.1 Famex Edit and Imputation Process: Current Approach

Initially, the FAMEX data are collected by a personal interview using a 45 page questionnaire. After data collection, the questionnaires are reviewed by the senior interviewer for completeness and consistency. A balance sheet (based on the collapsing of questionnaire cell entries) representing the household budget is manually created by the senior interviewer for each household. A calculation is included in the balance sheet to measure C, the absolute difference in percentage between inflow and outflow of money ($C \geq 0\%$). This calculation will be referred to as the balance edit. A document is said to be out of balance, if C is greater than a certain threshold value, say T. Note that different threshold values are used by FAMEX in different stages of the imputation process. A difference greater than 10% ($C > T$, $T = 10\%$), is the criterion used by the senior interviewer to indicate that the questionnaire needs to be reviewed further. Typically when a questionnaire is out of balance, different variables of the balance sheet that 'look suspicious' are checked first and the corresponding sections of the questionnaire reviewed, and when applicable, edited. Possible reasons of a failing balance edit may be several and diverse. For example, the imbalance may be due to one cell in the questionnaire that was transcribed as assets instead of debts, or a temporary resident of the household was considered as a permanent resident etc... Depending on the problem, the solution may be fairly simple, or may require a detailed analysis of the questionnaire. This manual operation can consume considerable resources depending on the household composition, demographic characteristics and the complexity of their budget.

The next step involves the capture of data and data editing at head office with an interactive processing system using validity and consistency edits. Another balance sheet is produced automatically by the processing system and documents that are still out of balance are identified. The editor is prompted by the system to question improbable data (according to edit

rules) and inconsistencies are manually resolved by studying the questionnaire contents.

After all the above steps, roughly 20% of the survey documents remain out of balance. The documents with an imbalance smaller than 15% are considered 'low' risk and are reviewed by the Operation and Integration Division (O&ID) quality officers, errors are corrected and the documents are accepted. Questionnaires that are out of balance by 15% or more are considered 'high' risk and are referred to FAMEX staff for final resolution. After examination and revision, the documents are accepted with or without any modifications or are rejected. This review consumes considerable resources, and the use of GEIS at this particular stage may result in considerable savings.

The final step in this process is the use of an automated edit and imputation system to perform imputation for any documents that require it. This includes consistency checks as well as imputation of any remaining missing data. Figure 1 is a diagrammatic representation of the above process.

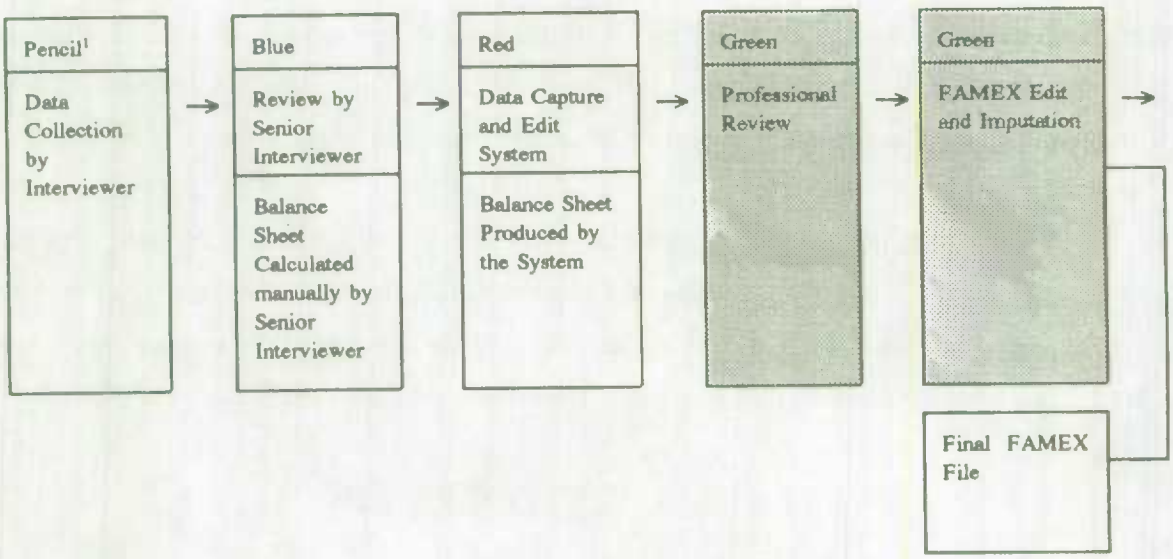


Figure 1. Current Famex Edit and Imputation Process.

[1] The colour associated with each stage corresponds to the colour used in the colour pencil study [see 4] and has no significance in this study. They are mentioned to facilitate the identification of each imputation stage.

2.2 Study Objectives

Under the current FAMEX procedure, resolving a problem balance sheet, i.e. a balance sheet that failed the balance edit, generally involves verifying individual cell entries of the questionnaire. In the context of an imputation system, this would require the simultaneous analysis of more variables that GEIS could possibly handle at once[2], as well as the determination of an extensive number of edit rules which are beyond the scope of a preliminary study. Hence, the application of GEIS was restricted to the variables created at the balance sheet level and after the red stage (see figure 2). For time and cost considerations, the analysis was restricted to the data for the city of Montréal. Specifically, the objectives of this study were:

(1) To determine if GEIS can resolve balance sheets that showed a difference between inflow and outflow of money being greater than or equal to 15% after the red stage.
(2) To evaluate how the estimates obtained for each imputed variable as a result of the above step compare with published FAMEX estimates at the city level.

A balance sheet is considered to be resolved, if after the edit and imputation process all missing variables in the balance sheet are imputed and the inflow and outflow of money differs by less than 15%. A threshold value of 15% was used since it corresponds to the criterion used by FAMEX, after the red stage, to identify household documents that are reviewed by FAMEX staff for final resolution.

The relative percent change between values imputed by GEIS and standard FAMEX values, was used as a criterion to measure the efficiency of GEIS. More specifically, the relative percent change, denoted by ϕ_{ik} for the i-th variable estimate (i = 1 to 31), was calculated at two levels, i.e. when considering only the values requiring imputation (k = 1), and at the city level (k = 2):

$$\phi_{ik} = (V_{ik}^{GEIS} - V_{ik}^{FAMEX}) / V_{ik}^{FAMEX} * 100$$

---

[2] GEIS can handle approximately 40 variables at a time, depending of the complexity on the edits. In larger surveys it is usually necessary to divide the variables into independent sets called edit groups. The edits are applied simultaneously within an edit group, however, each edit group is treated independently. When it is impossible to divide the variables in independent sets, i.e. one variable must be treated in at least two edit groups, the imputation for the second edit group is conditional on the imputed value in the first edit group [5].

Where, V_{ik}^{GEIS} and V_{ik}^{FAMEX} are respectively the estimates of total for GEIS imputed values and FAMEX standard values for the i-th variable estimate at the k level.

Note that consistency (i.e. relationships are retained) among variables at the balance sheet level is not required in this study. For example, a large 'Mortgage' value may be given to a renter, that initially did not have any mortgage payments, in order to make the inflow and outflow of money within 15% of each other.

2.3 Study Database

The data used for this study were obtained from a 1990 FAMEX sample for the city of Montréal coded under the colour pencil schemes [4]. Four-hundred and ninety-six (496) household documents were used to create 496 balance sheets. It should be pointed out that the balance sheet in itself does not include any information about the household composition and demographic characteristics[3]. It contains only values of many budget components representing expenditures or assets etc.

One balance sheet contains 31 values (or fields) corresponding to the 31 variables under study. For example, one of the variables represents rented living expenditures, and it is equal to the sum of five different cells (components) of section C08 of the questionnaire. A detailed description of the variables can be found in appendix 1. A variable is considered missing if at least one of its major components (components that were considered to have a large contribution to the variable total) at the questionnaire level is missing.

The balance sheets used were produced from edited questionnaires that were processed by the FAMEX Edit and Data Capture System i.e. after the red stage and before professional review. These balance sheets will be referred to as Test data. (see Figure 2).

Using the FAMEX published data, a similar balance sheet was produced for each of the 496 household documents using the same 31 variables. These balance sheets will be referred to as Standard data. A diagrammatic description of the above steps is given in Figure 2.

---

[3] Due to the complexity of the data and data files, extensive data manipulation would have been necessary to incorporate information about the household composition and demographic characteristics in the data base originally created for this project in 1991.

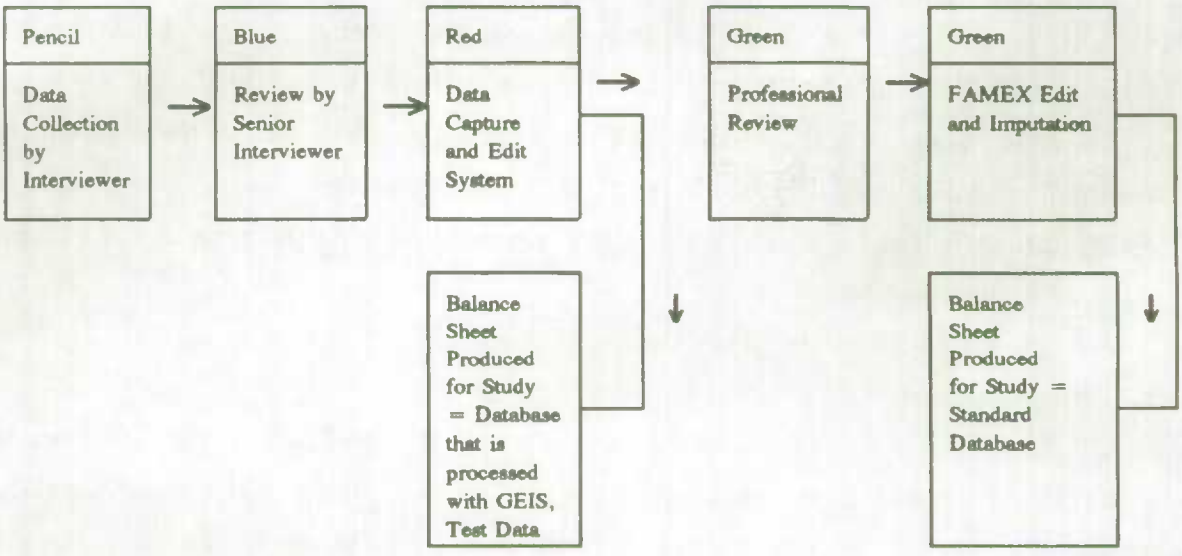


Figure 2. Creation of database: Data Processed with GEIS (Test Data), and Standard Data.

2.4 GEIS Overview

GEIS is a collection of modules which can be put together in order to satisfy the edit and imputation requirements of a survey. Data to be processed by GEIS are assumed to be numeric, continuous and non-negative. The edits used in GEIS must be expressed in linear form. GEIS consists of three major components: specification and analysis of edits, error localization and imputation [5]. These are discussed briefly below.

The edit strategy in GEIS is to locate acceptable or clean records by means of the conditions (edits) which a particular record must satisfy so that it will be acceptable for further processing. The quality of the edits has a direct impact on the final data quality of the imputed values.

The next step, i.e. the Error Localization module, identifies the fields to be imputed, but it does not actually perform any imputation.

In GEIS, donor imputation technique based on the nearest neighbour approach is the primary method of imputation, although other methods are available. Fields that require imputation are imputed by transferring the corresponding values from the closest acceptable donor. The 'closeness' is determined based on a set of matching fields (other variables) which may be chosen by the user, the system or both. In GEIS, the system automatically determines the matching fields for each record based on the original edits, the recipient's pattern of fields

to be imputed, and the values for fields that are retained. However, the user also has the option of specifying additional match fields. A donor is deemed to be acceptable if transferring its values to the recipient will allow the recipient to pass a user-specified set of post-imputation edits (see [5] for more details).

2.5 Application of GEIS to FAMEX.

This section presents the motivation behind the imputation strategy. A more detailed description of the imputation strategy itself is presented in section 2.6

GEIS usually determines which fields are to be used in the distance function for imputation purposes. This is done by analyzing the variables involved in all the edit rules related to the variable requiring imputation. Consequently, if one edit rule includes all variables under study, as it is the case for the balance edit, all non missing variables that do not require imputation are systematically picked by the system to be used as match fields. The use of uncorrelated variables as matching fields can considerably reduce the efficiency of the imputation procedure. Hence, a strategy was developed to identified the variables that should be used as match fields for each variable that required imputation.

For the imputation process to be meaningful, it is necessary that each balance sheet pass the balance edit (household level) while still producing acceptable estimates for each variable at the city level. These two goals can sometimes be conflicting in the sense that for a particular balance sheet, any imputed value that is within the acceptable range for that variable could make the balance sheet pass the balance edit. However, the imputed value itself may be inadmissible given the context (e.g. a high mortgage value is given to a renter that does not have any mortgage payments). This type of inconsistency could yield estimates of total for particular variables that are unacceptable at the city level. Hence, consistency among variables, even if not required in this study, is desirable at the balance sheet level in order to get acceptable estimates at both the balance sheet and city level. Specific match fields and when applicable, edit rules were determined in order to maintain the existing relationships among variables at the balance sheet level.

The imputation process was done in three stages, the requirements (post-imputation edit rules) being less restrictive at each successive stage. Edit rules that were considered desirable but that were not absolutely necessary were allowed to be less restrictive in subsequent stages in order to allow the system to pick acceptable donors (see appendix 2). Figure 3 is a diagrammatic

representation of the above process (Figure 3).

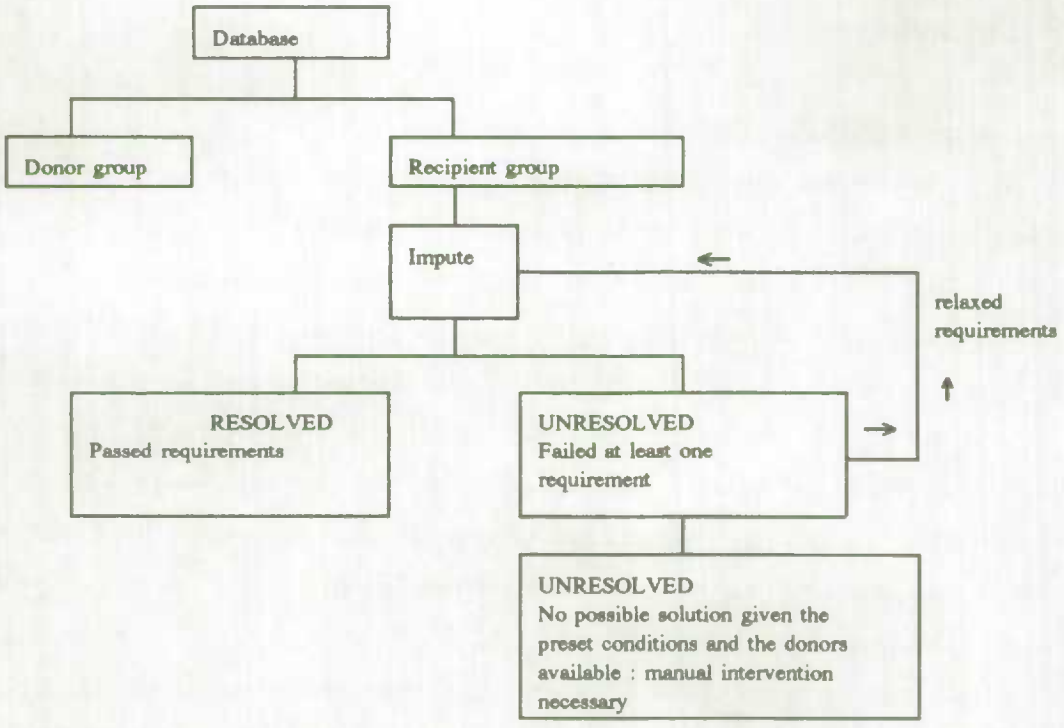


Figure 3. Overview of Imputation Process used with GEIS

2.6 Description of Imputation Process

The donor imputation module was used to perform the imputation (the closest donor is found based on a distance function). In this GEIS application, the same donor is used to impute all the values requiring imputation within a balance sheet (household record). The use of the donor imputation module requires (1) the development of pre-imputation and post-imputation edit rules, (2) the determination of the variables that should be picked for imputation to resolve a failing balance sheet (weights) and (3) the determination of match fields, i.e. variable(s) that should be considered in the donor imputation distance function. These requirements are described in the following sections.

2.6.1 Edit Rules

The Pre-imputation and post imputation edit rules were defined to determine (A) Recipients and Donors, and (B) specific requirements (edit rules) that must be met by the imputed records. All the rules used as pre-imputation edit rules were also used as post-imputation edit rules.

(A) Recipients and Donors

For a record to be labelled as a recipient it must satisfy at least one of the following criteria:

1. At least one of the variables in the balance sheet is missing.
2. The balance sheet shows a difference between 'inflow' and 'outflow' of money equal or greater than 15%.

A record is considered a potential donor if it satisfies the following criteria:

1. The data in the balance sheet is 'clean' e.g. it does not require further modification.
2. The balance sheet shows a difference between 'inflow' and 'outflow' of money smaller than 10%.

The chosen donor among all potential donors will be the one that is considered the 'closest' to the recipient based on the distance function. Note that some records may not belong to either the recipient or donor category.

(B) Specific Requirements

Edit rules were used not only to check the initial entries (pre-imputation edit rules), but also to ensure that the imputed data (post-imputation edit rules) met different requirements at the balance sheet level (household level). Some edit rules were defined to be consistent with the FAMEX imputation procedure, and others, to maintain the existing relationships among variables at the balance sheet level. A complete list of these edit rules can be found in appendix 2. Edit rules that conform with FAMEX imputation procedures were as follows:

1. A linear form of the balance edit was developed to measure the difference in the inflow and outflow of money. Balance sheets that failed this edit were identified by the system as being unresolved.

2. Edits defining the minimum and maximum value that a variable can take. These were defined based on the maximum and minimum value for each variable observed in the final FAMEX data (standard database) at the city level.
3. To be consistent with FAMEX procedures, edit rules were required to force the imputed values for five of the variables to be greater than zero.

In addition to the above edit rules, relationships among variables were used to define additional variable dependent rules using cluster analysis and correlation techniques. Among all the variables under study that required imputation, only one showed a strong correlation with another variable. A strong correlation between 'Mortgage' (C7) and 'Amount-of-Money-Borrowed-or-Renewed' (D1) was observed in the standard data. It justified the use of a special edit rule requiring that the imputed values show a similar relationship. Typically, D1 had a zero value when 'Mortgage' (C7) varied between 0 and 20,000 and increased linearly with 'Mortgage' when 'Mortgage' was larger than 20,000. Hence, for the imputation of C7, the recipient's value of D1 was used to guide the system towards a 'better' donor i.e. a donor that has a similar value for D1.

Operationally, this was accomplished by specifying that the value for D1 for the recipient and a potential donor should be within 60 % of each other. For example, a recipient with a value for D1 of 100,000 could only find a donor that had a value for D1 between 60,000 and 166,667. This also ensured that a recipient with a zero value for D1 could only have a donor that also had a zero value for D1. A criterion of 60% was chosen due to the limited number of potential donors with 'Mortgage' larger than 20,000.

It should be pointed out that this edit rule would not be necessary if D1 was the only matching field used when C7 is missing. The donor imputation module would choose automatically the closest donor in the donor group. The use of the edit rule ensured that the relationship between D1 and C7 was maintained independently of the matching fields used.

2.6.2 Relative Importance of Variables (Weights)

In any edit and imputation system, the selection of fields to impute must be based on some criterion. The strategy used by GEIS is to minimize the number of fields requiring imputation. The user is able to exert some influence on the fields that are selected for imputation by assigning weights to variables. GEIS then minimizes the sum of the weights of fields that are

identified for imputation (see [5] for more details).

Although this is not the actual procedure that GEIS uses to determine the fields to impute, the general strategy for the FAMEX application is the following. In the current imputation strategy, missing data must be imputed. Hence, a minimum change in a failing balance sheet could be achieved by selecting only missing values for imputation. However the values required to resolve the balance sheet may be outside the range of possible values for the missing variables. In this case, another variable, not actually missing is picked to be imputed as well in an effort to resolve the balance sheet. Weights are assigned to variables in order to determine which variables(s) will also be flagged for imputation. Note that these weights have nothing to do with sampling weights but are used to determine which variable(s) should be preferentially flagged for imputation to resolve a failing balance sheet. Weights were defined in the following manner.

Variables were ranked in increasing order according to (1) the percent change of a particular variable from the red stage to the final stage, and (2) the relative importance of this change to the total budget. In the ranking procedure, twice as much importance was given to the second criterion since variables that represent larger amount of money have a higher potential to resolve a failing balance sheet in a minimum number of steps. Variables with the highest rank were given a small weight, i.e. they would be preferentially picked for imputation. Essentially, 5 variables were identified to be picked for imputation, C1D4, AZ4, C7, C9 and D1 (see appendix 3).

2.6.3 Distance Calculation: Choice of Variables

The donor imputation module finds the closest donor for a particular recipient based on a distance function between donor and recipient matching fields. Matching fields correspond to variables in the balance sheet that are compared when calculating the distance between a donor and a recipient.

In a typical GEIS application, matching fields are determined directly by the system based on the defined edit rules or specified by the user. In this study, due to the nature of the balance edit, all non missing variables were systematically picked by the system as match fields. Depending on the balance sheet, between 23 to 30 variables were picked as matching fields. The list of matching fields provided by the system were not used as is. Matching fields were defined by the user to prevent the use of uncorrelated variables as matching fields. It should be noted that this step required extensive data manipulation.

To identify the match fields, i.e. variables that are strongly correlated, variables were grouped using an oblique component analysis (VARCLUS procedure, SAS). It should be pointed out that this method was not used for a strict statistical analysis but was only used as a tool to help determine which variables should be grouped. The multiple factor method consists of representing the factors (clusters) by reference axes passing through the centroids of the respective groups of variables that are not forced to be orthogonal [6]. This is of primary interest when it is suspected that a variable can be associated with more than one cluster[4]. Variables are assigned to clusters to maximize the variance accounted for by the cluster components. The cluster component is represented by a linear combination of all the variables in the cluster. The number of clusters formed is dependent on a user or system specified criterion. Based on the default criterion used in SAS, 14 clusters were originally identified: 9 of these clusters were formed by only one variable.

The clusters defined by this technique were further analyzed by examining a two-dimensional spatial representation of cluster components. The purpose of the spatial representation was to determine if a particular variable should be associated with more than one cluster. Based on this analysis two of the originally defined clusters were regrouped into one cluster. In summary, twenty-two of the variables were regrouped into five clusters. Nine variables could not be associated with any other variable(s), hence no matching fields were identified for these variables (appendix 4).

When a variable within a cluster requires imputation, the other variables identified as being part of the same cluster are used as match fields and in the distance calculation. Note that all missing variables within a balance sheet are imputed simultaneously when a donor is found i.e. the same donor is used to impute all missing values. The donor must satisfy the requirements (edit rules, match fields) related to each missing variable. Consequently, when many variables belonging to different clusters required imputation in one balance sheet, all the match fields corresponding to the different clusters are used in the distance calculation. This method differs from a block imputation method where each variable can be imputed independently of the others within a balance sheet. In block imputation different donors may be used to impute variables within the same balance sheet.

---

[4] Note that the term cluster is not used in the sense of sampling design. Here the cluster is a collection of variables.

In some cases, matching fields could not be identified for any of the variables requiring imputation within a recipient balance sheet. When no matching fields are defined the distance function cannot be used. In such cases, the system determines which donors can be used to resolve the balance sheet. If there is more than one potential donor the system will pick one at random. If there is none, the balance sheet remains unresolved.

3. DATA ANALYSIS

The data analysis covers two aspects: (1) a study of the overall performance of the imputation process by considering how many balance sheets were resolved, (2) a study of the impact of the imputation strategy, i.e. specifically how the edit rules, match fields and weights had an impact on estimates of total obtained for each variable when considering **only the imputed values** (ϕ_{i1}), and a comparison of the estimates obtained from GEIS (Test data) and FAMEX (Standard data) for each variable **at the city level** (ϕ_{i2}).

3.1 Overall Performance of Imputation Process

Preliminary analysis of the data showed that some of the variables in the 376 balance sheets that passed the balance edit were further modified by FAMEX in later stages (green stage). Hence, this group of balance sheets could not qualify as donors since the data were not 'clean' or 'final'. If these balance sheet were used as donors, any difference found when comparing the GEIS imputed data and the final FAMEX data could be due to donor data that was not 'final'. To resolve this problem the 'clean' data from the final published data was obtained for these 376 balance sheets. The 'clean' 376 balance sheets constitute the donor group.

Of the 496 balance sheets produced after the red stage, 105 were classified as recipients. Eighty-five of these balance sheets had at least one missing value for a variable. The other 20 balance sheets did not have any missing value but they failed the balance edit set at 15%.

Eighty-eight (88) cases were resolved in the first imputation stage, 6 at the second stage and 3 at the third stage for a total of 97 resolved cases. Hence, the imputation process was successful for 97 of the 105 (92%) recipient records. No donors could be found, from the 376 available donors, that could make the 8 remaining balance sheets pass the balance edit. A diagrammatic representation of the imputation process is presented in Figure 4.

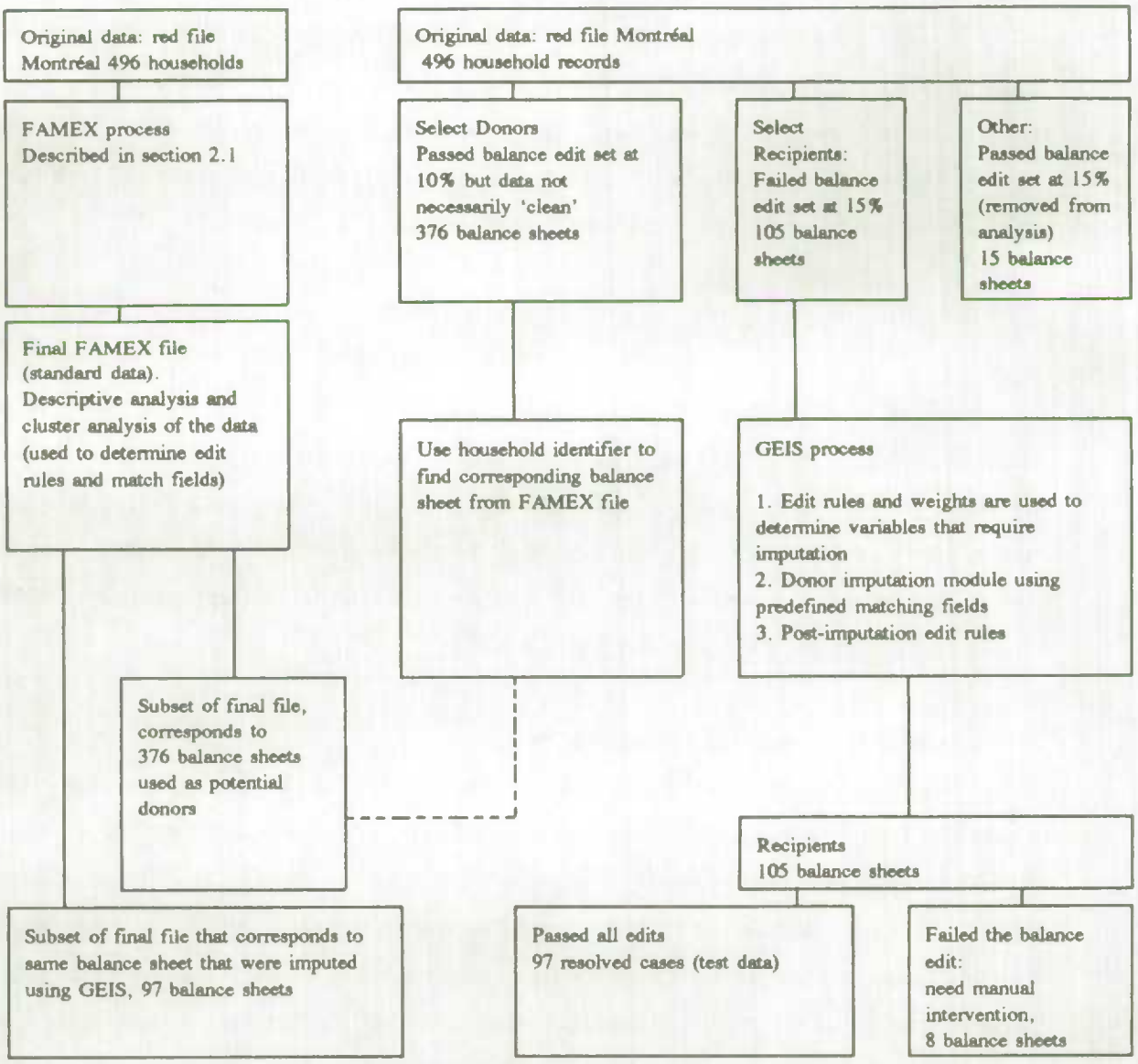


Figure 4. Description of the Imputation Process

3.2 Impact of Imputation Strategy: Imputed Values

Depending on the variable, less than three or between 9 and 55 fields required imputation. Evidently when only a few fields required imputation (three and less), conclusions can not be drawn on the efficiency of the imputation technique since edit rules were not defined

to ensure consistency at the record level. The impact of the imputation strategy is discussed separately for the eight variables that required at least 9 fields to impute: C7, C9, AZ5, AZ4, A5, BZ2, CZ2D13, C1D4[5]. Table 1 presents the results for the 9 variables where this analysis was applicable. The correlation of the variable requiring imputation with the closest cluster gives an indication of the quality of the matching fields used.

Table 1. Percent Change Between GEIS Imputed Values and FAMEX Standard Values*

VARIABLES	C7	C9	AZ5	AZ4	A5	BZ2
Correlation with closest cluster	0.79	0.39	0.84	0.83	0.65	0.26
Number of fields requiring imputation	55	24	14	14	10	9
Number of fields resolved	51	23	13	11	9	8
GEIS estimates for the resolved cases	1288879	157262	43551	93479	11887	35474
Famex estimates for the resolved cases	1168373	164262	41827	62023	8174	28060
GEIS estimate - FAMEX estimate for the resolved cases	120506	-7000	1724	31456	3713	7414
% change with FAMEX estimates for the resolved cases (ϕ_{ii})	10.31**	-4.26	4.12	50.72	45.43	26.42

* Matching fields were identified for C7, AZ4, AZ5 and A5. For variable C9 and BZ2 matching fields defined in the 'closest cluster' were used in the distance calculation.
** The percent change for C7 is calculated as follow: (1288879-1168373)/1168373 *100 = 10.31

3.2.1 Imputation of 'Mortgage' (C7)

The variable that required imputation most frequently was 'Mortgage' or C7. It constitutes as well the only variable for which a specific post-imputation edit rule could be defined. In all, 'Mortgage' required imputation in 55 balance sheets. After the three imputation stages, 51 of the 55 'Mortgage' values requiring imputation were imputed. Donors for 'Mortgage' were found for 42 balance sheets in the first stage, for 6 balance sheets in the second stage, and finally, for 3 balance sheets in the third stage. No donors could be found for 4

[5] See appendix 1 for a description of these variables.

balance sheets in which 'Mortgage' required imputation. These 4 balance sheets would require manual intervention. The effect of the special post-imputation edit rule and the match fields are discussed in the following sections.

The impact of the post-imputation edit rules is assessed by comparing the results obtained during the successive imputation stages when the edit rules related to the imputation of 'Mortgage' were made less restrictive. The total of the 42 'Mortgage' values imputed at the first stage differs from its corresponding standard total by only 2.6 % (Table 2). After removal of the edit rule relating 'Amount-of-Money-Borrowed-or-Renewed' (D1) and 'Mortgage' (C7) in the second stage, 6 additional records were resolved. At this point the total for the 47 resolved cases (41 + 6) differs from its expected value by 11.3%. Finally, in the last stage, three more cases were resolved by imputing 'Mortgage' with a zero value. When all the 51 resolved records are compared the percent change is 10.3%. It appears that the imputation process was most efficient when the special edit rule was used.

Table 2. Impact of the Imputation Process on 'Mortgage' (C7).

Imputation stage	Number of cases resolved	Post-Imputation edit rules	% Change (cumulative total)
1	42 cases resolved	1. Edit rule for C7 and D1 2. C7 when imputed must be > 0	+2.6% (42 records)
2	6 additional cases resolved	1. C7 when imputed must be > 0	+11.3% (47 records)
3	3 additional cases resolved	1. C7 when imputed can be = 0	+10.3% (51 records)

To investigate the impact of matching fields on the imputation of 'Mortgage', a simulation was performed where donors were found using different set of matching fields (the actual database produced for this study was used). Two cases are presented: in the first case, the variables used as match fields are closely related to C7 (determine through cluster and correlation analysis), in the second case all other variables (30 variables) are used as match fields (Table 3). In this example, the recipient record has a missing value for C7. FAMEX imputed C7 with a value of 57,336.

Table 3. Effect of Matching Fields on Imputation of 'Mortgage' (C7).

Case	Matching fields	Donor Value for C7
1	CZ3, D1, D2 (closely related)	C7 = 55654 C7 = 63368
2	30 variables (most of them are uncorrelated)	C7 = 0 first choice
		C7 = 55654 one hundred and thirty first choice C7 = 63368 two hundred twenty first choice

Using closely related variables as matching fields, two donors are identified (they both had the same distance from the recipient). Since there is two donors, one of the donors would be selected at random. The imputed value for C7 would be either 55,654 or 64,368 (case 1, Table 3). Both values are comparable to the standard FAMEX value (57,336). When all 30 variables are used as matching fields, the closest donor has a value for C7 equal to zero (case 2, Table 3). It is interesting to note that the donors identified in the previous case, when using only correlated match fields, are now the one hundred and thirty first closest and the two hundred twenty first closest donors, when all 30 variables are used as matching fields. In GEIS, a particular donor's distance corresponds to the largest distance observed among its matching fields and the matching fields of the recipient. Hence potentially good donors for C7, such as those identified when only a few correlated match fields are used, are given a large distance due to the presence of at least one other variable (match field) that happens to differ greatly between the donor and the recipient.

Preliminary analysis of the data using an imputation procedure where all 31 variables are used as matching fields, and without any special edits for C7 and D1, resulted in an underestimation of 'Mortgage' by 52% for the same 51 cases. Clearly, the use of uncorrelated matching fields can reduced considerably the efficiency of the imputation procedure.

Further analysis of the standard FAMEX data revealed that the recipient balance sheets were characterized by a high frequency of larger 'Mortgage's. In the final FAMEX data corresponding to the 51 balance sheets where 'Mortgage' required imputation, 37% of the records (19 out of 51 records) had 'Mortgage's larger than $ 20,000. However, in the donor group (376 balance sheets), only 6% of the balance sheets had a 'Mortgage' larger than $ 20,000. After imputation with GEIS, using a special edit rule and a few correlated match fields, the percentage

of 'Mortgage' larger than $ 20,000 is 33% (17 records out of 51).

In summary, the specification of edit rules used in combination with a reduced number of matching fields contributed largely to the efficiency of the imputation procedure of 'Mortgage'. Simultaneous imputation has the advantage of preserving the relationships among variables since all imputed variables come from the same donor. However, the donor chosen, while meeting all requirements, may not have been picked if each variable would have been imputed independently (as it would be the case with 'block imputation'). It should be noted that block imputation is used in the FAMEX Edit and Imputation Process.

For the two following variables, C9 and BZ2, the efficiency of the imputation process was assessed (1) when only the variable under study was missing (this could be associated with a block imputation), and (2) when other variables were also missing in the balance sheet (which corresponds to a simultaneous imputation of many variables from the same donor).

3.2.2 Imputation of 'Amount-of-Interest-and-Principal-Paid' (C9)

Variable C9 required imputation in 23 balance sheets. No strong or moderate relationships with any clusters were identified for variable C9. However, since this variable was often missing it was tentatively matched with the closest cluster (R = 0.4). For analysis purposes the balance sheets were divided in two groups, (1) C9 was the only variable missing in the balance sheet (11 balance sheets), (2) other variables were also missing in the balance sheet (12 balance sheets).

Table 4. Impact of Different Set of Matching Fields on Imputation Process of 'Amount-of-Interest-and-Principal-Paid' (C9)*.

Matching fields defined	Number of balance sheet	Average GEIS value	Average FAMEX value	% change
Only closest cluster	11	3780	3568	5.94
Group A: Mortgage also missing	6	4410	3518	25.36
Group B: Other variables also missing	6	1764	3629	-51.39
Total	23	3418	3570	4.26

* The average values for C9 at the city level in the standard data (496 records) and the donor group (320 records) are 1267 and 1232 respectively.

The impact of linking this variable to the closest cluster was assessed by comparing the test and standard average values for the 11 balance sheets where only C9 required imputation. The average values for these 11 records were respectively 3,780 and 3,568 corresponding to an overestimation of the standard data by 5.9% (Table 4). The standard average value for C9 at the city level is roughly twice as small (1267). The recipient records appeared to be characterized by larger than average values and the use of match fields, even remotely related, contributed to successfully impute C9. In the absence of edit rules, the use of a block imputation method relying on a few correlated match fields could be given some consideration.

For 12 of the 23 cases, variable C9 was not the only missing variable in the balance sheet, hence different matching fields were selected to find the closest donor depending on the balance sheet. These 12 cases were partitioned in two groups, A and B. In group A 'Mortgage' was also missing, in group B variables other than 'Mortgage' were also missing. The values of the 6 cases that were resolved when 'Mortgage' was also missing were on average larger than their corresponding standard values (overestimation of standard value by 25.3%). However, when variables other than 'Mortgage' were also missing the imputed values were closer to the standard average value for C9 at the city level but underestimated the standard value by 51.4%.

The totals of the 23 resolved cases of variable C9 overestimated the standard total by only 4.3% (Table 4). It appears that the use of match fields, even remotely related, contributes to successfully impute C9. However, the overall success of the imputation process depends largely on which variables are missing. In this case it **happened to be successful**, since the impact of group A on the imputed values was compensated by the impact of group B (it could have been otherwise if other variables had been missing).

Furthermore, this case illustrates the impact that one variable can have on the choice of a 'suitable donor'. Whenever 'Mortgage' was also missing, the donor had to satisfy all the requirements related to 'Mortgage', the donors chosen, suitable for C7, systematically overestimated C9. Specific edit rules for C9 would have been required to overcome this problem.

3.2.3 Imputation of 'Other-Vehicle-Transportation-Cost' (BZ2)

Variable BZ2 was flagged for imputation in 9 balance sheets. No strong or moderate relationships with any clusters were identified for this variable. Again, it was matched with the closest cluster (R = .3). For analysis purposes the data were separated in two groups, 1) BZ2

was the only variable missing in the balance sheet, 2) other variables were also missing in the balance sheet.

For the three cases where only the closest cluster was used as match fields the average values between the standard and the test data differ by 3.3% (Table 5). For 5 of the 8 cases, other variables were also missing on the balance sheet. These five cases account for an overestimation of the corresponding standard data by 42.3% (Table 5). The totals of the 8 resolved cases of variable BZ2 overestimated the standard total by 26.4% (Table 5). It appears that the presence of other variables that also required imputation directed the system toward donors that systematically overestimated BZ2.

Table 5. Impact of Other Missing Variables on Imputation Process of 'Other-Vehicle-Transportation-Cost' (BZ2).

Matching fields defined	Number of cases	Average GEIS value	Average FAMEX value	% Change
Only Closest cluster	3	3941	3814	3.33
Other	5	4730	3324	42.32
Total	8	4434	3508	26.42

3.2.4 Imputation of 'Transportation-Cost' (AZ4 and AZ5).

Respectively 69% and 70% of the variation (R^2) of AZ4 and AZ5 could be explained by the linear relationship with the cluster composed of variables 'Household-Equipment-Expenditures' (A11), 'Clothing-and-Personal-Care' (AZ31), 'Running-the-Home-and-Food' (AZ32), AZ4 and AZ5. Overall, values for A11, AZ31 and AZ32 showed a tendency to increase linearly with AZ4 and AZ5. This relationship is particularly evident in figure 5 where the values for A11, AZ31 and AZ32 are plotted for the 10 smallest and largest values of AZ5. It also explains the relatively high correlation observed ($R = 0.8$). However, for a small increase in AZ4 or AZ5 values for A11, AZ31 and AZ32 do not vary linearly (see figure 5). Hence, donors with similar values for A11, AZ31 and AZ32 had values for AZ5 (or AZ4) that varied considerably.

The total of the imputed data for variable AZ5 differs by only 4.26% from its expected value. For variable AZ4 the total for the imputed value overestimated the expected total by

45.4% (Table 1). The larger difference observed between imputed and standard values for variables AZ4 gives an indication of the type of variability that can be expected when the variables used as matching fields show this type of relationship.

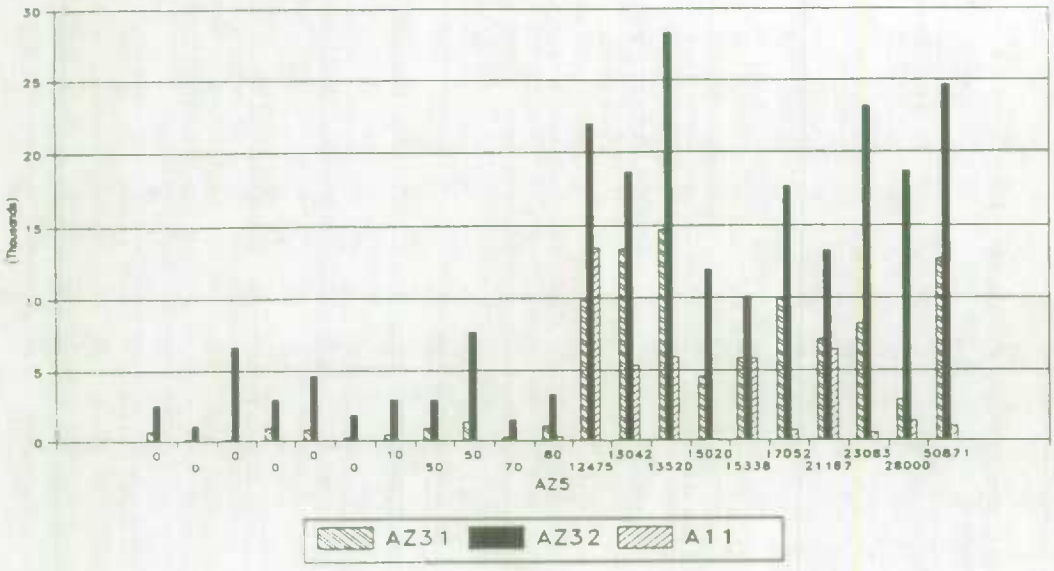


Figure 5. Relationships Between Variable AZ5 (Transportation cost) and Variables AZ31 (Clothing and Personal Care Expenditures), AZ32 (Running the Home and Food Expenditures) and A11 (Household Furnishing and Equipment Expenditures).

3.2.5 Imputation of 'Shelter-Expenses' (A5)

A moderate relationship (R = 0.7) was observed for 'Shelter-Expenses' (A5) with one cluster composed of variables 'Rent' (A4), 'Owned-Living-Quarters-Expenditures' (AZ1), 'Mortgage' (C7), 'Home-Purchase-Price' (CZ3), 'Amount-of-Money-Borrowed-or-Renewed' (D1), and 'Selling-Price-of-Home' (D2). The total of the 10 imputed data differed from their expected value in absolute value by only 3,713, however, due to the size of the estimates involved this difference translates to a overestimation of 45.4% (Table 1).

3.2.6 Imputation of 'Change-in-Loan-and-Debts' (CZ2D13)

Variable CZ2D13 was missing in 17 balance sheets. Matching fields could not be identified for CZ2D13. This variable was not linked to any cluster due to the absence of correlation with any defined clusters (the highest correlation observed was 0.07). Whenever

CZ2D13 was missing, other variables with defined matching fields were also missing.

Values for CZ2D13 were imputed from the closest donor determined from the different matching fields that were defined for the other variable(s) requiring imputation in each balance sheet. The estimate of total happens to differ by 10.5% from its expected value (Table 6).

3.2.7 Imputation of 'Change-of-Assets' (C1D4)

Variable C1D4 was never originally missing. However it was picked for imputation by the system in 10 balance sheets (see section on weights).

Matching fields were not identified for C1D4 due to a lack of a relationship with any defined clusters (highest correlation was 0.05). For 8 of the 9 resolved balance sheets, C1D4 was the only variable flagged for imputation. Consequently for these 8 resolved cases, the distance function could not be used: a donor was picked **randomly** among acceptable donors (see section on matching fields). Any donor that has a value for C1D4 that will make the balance sheet pass the balance edit is an acceptable donor. Values chosen may or may not be consistent with other variables present on the balance sheet. The total of the nine imputed entries overestimated the expected total by 31.9% (Table 6).

Table 6. Impact of Imputation Process for Variables Without Defined Matching fields.

	CZ2D13	C1D4
Correlation with closest cluster	0.07	-0.05
Number of fields requiring imputation	17	10
Number of fields resolved	16	9
% change with standard data for the resolved cases	-10.53	31.97

3.3 Comparison of Estimates at the City Level

The comparison of the estimates of totals at the city level may differ for two reasons: 1) The imputed values differ from their corresponding expected (standard) values, and 2) Some variables that were not modified by GEIS were modified by FAMEX between the red stage and the final stage. FAMEX may have edited these variable(s) to resolve the balance sheets or resolve any inconsistency.

Thirty one variables were considered in the imputation process. Twenty variables of the recipient balance sheets had at least one value imputed by GEIS. Six variables were never modified by GEIS but were modified by FAMEX. The impact of the imputation strategy for these variables is presented separately in section 3.3.1 and 3.3.2.

Five variables that were neither modified by GEIS or FAMEX (CZ1DZ1, A4, BZ1, D2, D3) were excluded from the analysis.

3.3.1 Imputed by GEIS

At the city level, the percent change between V_{i2}^{GEIS} and V_{i2}^{FAMEX} was less than 5% for 18 of the 20 variables requiring imputation (see table 7). Variables C1D4 and CZ2D13 showed a difference between the totals of 65.9% and 16.9% respectively[6]. More specific information would be required at the balance sheet level in order to improve the success of the imputation process of these two variables at the city level.

Some imputed values had a considerable impact on the estimates. For example, in one balance sheet the value for variable 'Withdrawals-Business-or-Farms' (DZ2) was not modified by GEIS but was modified by FAMEX causing a difference of 4.47% in the estimate of total at the city level (Table 7). This particular balance sheet was resolved by GEIS by imputing variable C7 instead of variable DZ2. The modification to variable DZ2 made by FAME6X could not be anticipated by studying the relationships among variables at the balance sheet level.

---

[6] Note that these two variables could take positive or negatives values. Hence, depending on the subgroup of balance sheets considered an overestimation or underestimation may be observed. For example, when considering only the 9 imputed values for C1D4, the corresponding standard total is overestimated by 31.97%, however, at the city level, when considering all 473 records, the standard estimate of total is underestimated by 65.9%.

Table 7. Comparison of Estimates at the City Level.

Variable	Processed with GEIS		Imputed by FAMEX but not by GEIS		Overall impact % change (ϕ_U)
	Number of imputed records	% change due to these records	Number of imputed records	% change due to these records	
C7	51	4.63	7	-0.79	3.84
C9	23	-0.58	0	0.00	-0.58
CZ2D13	16	16.85	0	0.00	16.85
AZ5	13	0.11	3	-0.15	-0.03
AZ4	11	1.38	13	-0.27	1.11
C1D4	9	-70.86	7	5.03	-65.93
A5	9	0.87	0	0.00	0.87
BZ2	8	3.05	0	0.00	3.05
AZ6	2	2.07	1	-0.06	2.01
AZ7	2	-0.86	1	1.63	0.77
D1	2	0.18	3	-0.24	-0.07
DZ2	2	0.00	1	4.47	4.47
BZ5	2	-0.50	3	-2.65	-3.16
A11	1	-1.44	10	-0.45	-1.89
A24	1	0.03	7	0.43	0.51
A27	1	-0.44	9	-0.44	-0.32
CZ3	1	0.00	0	0.00	0.00
C4	1	0.00	7	3.19	3.19
A23	1	0.00	0	0.00	0.00
DZ3	1	-1.52	3	0.88	-0.63

3.3.2 Imputed by FAMEX.

Six variables were never modified by GEIS but were modified by FAMEX (AZ1, AZ31, AZ32, BZ3, BZ4, B6). These variables were never missing and according to the preset criteria (see section on weights) were never selected by the imputation system. The impact of intentionally not modifying these variables was also measured by comparing the estimates at the city level. In this case, since no imputation took place, the original estimates created after the

red stage are compared with the final FAMEX estimates (Table 8). The criterion used to determine exclusion of five variables from the imputation process appears adequate; the difference between the red stage and the final stage for all 6 variables was less than 2% (Table 8).

Table 8. Percent Change ((Standard - Test / Standard) *100) at City Level for Variables that Were Never Modified by GEIS but Were Modified by FAMEX.

VARIABLES	AZ1	AZ31	AZ32	BZ3	BZ4	B6
Overall % change with standard data at the city level	0.83	-0.37	0.18	1.85	0.23	-0.07

4. CONCLUSIONS

The main results of this study on the application of GEIS to the FAMEX balance sheet are presented below as well as general comments . These comments are not intended for the possible use of GEIS for surveys other than FAMEX, and the conclusions presented here may not be applicable to them.

Imputation was successful for 97 of the 105 (92%) recipient balance sheets. Eight balance sheets remained unresolved. These balance sheets could not be resolved given the current imputation procedure (i.e. given the specific set of edit rules and values available in the donor group) and would have required manual intervention.

At the city level, estimates of total produced by GEIS and FAMEX differed by less than 5% for all 20 variables except two (C1D4, CZ2D13). No specific information was available for these two variables that would allow the elaboration of edit rules, furthermore they were not related to any other variables in the balance sheet. The imputation procedure used depended heavily on the exiting relationships among variables and was not adequate to treat variables where such relationships were nonexistent. It should be noted that in this application imputed values may not be consistent with non imputed values whenever edit rules involving the variables in question could not be determined.

In some cases, large differences were observed when comparing directly the values (for a particular variable) imputed by GEIS and their corresponding FAMEX value. For example,

the total of the 11 records imputed for variable 'Transportation-Cost' (AZ4) overestimated the corresponding FAMEX total by 50.7%. However, due to the small number of records involved, the impact of the 11 imputed values on the estimate of total at the city level is negligible. A larger discrepancy could be expected at the city level if the number of records requiring imputation for such variables were increased.

A specific edit rule, such as the edit rule used for Mortgage, was efficient in finding donors that were suitable for 'Mortgage' (C7). However, these donors did not necessarily yield acceptable values for the other imputed variables (for which no edit rules could be defined) that happened to be also missing when 'Mortgage' was missing. For example, whenever 'Amount-of-Principal-and-Interest-Paid' (C9) and 'Mortgage' (C7) were missing, C9 was found to be systematically overestimated (no edit rule was defined for C9). In the absence of edit rules, the use of a block imputation method relying on a few correlated match fields could be given some consideration.

Particular attention must be given to variables that contain a high percentage of valid zero values. For example, the modification of one field caused an overestimation of 4.5% at the city level (variable 'Withdrawals-Business-or-Farms' or DZ2). Edit rules defined at the record level may be necessary in these cases to specify different acceptable boundaries.

Six variables were intentionally not picked for imputation since (1) their relative contribution to the total budget was considered to be small, and (2) there was a small percent change in the estimate of total for these variables between the red stage and the final stage. The percent change between the red stage and the final stage for these 6 variables was less than 2%. This may indicate that imputation of these variables may not be necessary.

The imputation of the FAMEX balance sheet can not be considered a typical GEIS application since there was only one predefined edit rule at the balance sheet level that could be used during the imputation process. In GEIS, edit rules are used to ensure consistency among variables and are also used to determine which fields should be used in the distance function of the donor imputation module. In this study, due to the absence of edit rules, matching fields had to be determined by the user through cluster and correlation analysis. It was also necessary to perform extensive data manipulation in order to specify to the imputation system which matching fields should be used for a particular balance sheet.

Whether GEIS (or a slightly modified version) can be used to edit the FAMEX questionnaire can not be answered here as this question is outside the scope of this study.

However, the determination of relationship edit rules and deterministic edit rules remain a key factor in the success of any imputation process. Special attention should be given to this aspect during the redesign of the FAMEX questionnaire.

ACKNOWLEDGEMENTS

The author would like to thank M. Brodeur, F. Hardy, S. Kumar, F. Mayda, U. Nevraumont and E. Wilson of Statistics Canada for their constructive suggestions during the progress of the work, and valuable comments that helped improve the quality of this paper.

REFERENCES

1. Cotton, C. (1991), Functional description of the Generalized Edit and Imputation System. Statistics Canada Technical manual.
2. Legault S., Roumelis, P., (1992), The use of the Generalized Edit and Imputation System (GEIS) for the 1991 Census of Agriculture. Statistic Canada, Methodology Branch Working Paper No. BSMD 92-010 E.
3. Statistics Canada, Family Expenditure in Canada, Catalogue 62-555, June, 1992.
4. Silver, C., Chen, E. (1992), The colour Pencil Study: An Investigation into incidence and Impact of Edit and Imputation Activities in the 1990 Survey of Family Expenditures. Statistic Canada, Methodology Branch Working Paper No. SSMD 92-005 E.
5. Kovar, J.G., Mac Millan, J. and Whitridge, P. (1988). Overview and strategy of the Generalized Edit and Imputation System. Statistics Canada, Methodology Branch Working Paper No. BSMD 88-007 E/F.
6. Harman, H., (1970), Modern factor analysis. University of Chicago Press, 474 p.

APPENDIX 1. DESCRIPTION OF VARIABLES

VARIABLE	QUESTIONNAIRE SECTION(CELLS)	CELLS USED TO DETERMINE IF THE VARIABLE SHOULD BE CONSIDERED MISSING	DESCRIPTION
A11	C09(8)	8	SHELTER EXPENSES (OWNERS AND RENTERS): RENTAL OF HEATING
A11	D01(12-14-EVEN... -32,35,36)	12-14-EVEN...-32,35,36	HOUSEHOLD FURNISHING AND EQUIPMENT: REFRIGERATORS...-AIR CONDITIONER, ATTACHMENT, MAINTENANCE
A11	D03(1-3-ODD...-29, 30-40)	31,34-36	HOUSEHOLD FURNISHING AND EQUIPMENT: FURNITURE
A11	D04(1-5,7-ODD...-1 9,25,27-30)	1	HOUSEHOLD FURNISHING AND EQUIPMENT: ANTIQUES, TABLEWARE AND FLATWARE, SMALL ELECTRICAL AND NON ELECTRICAL APPLIANCES
A11	D05(1-23)	8,13	HOUSEHOLD FURNISHING AND EQUIPMENT: LAWN..., TOOLS..., OTHER...
A23	M01(7,9)	7,9	PERSONAL INCOME: INCOME IN KIND NON-FARM AND FARM
A24	N01(2)	2	PERSONAL TAXES: INCOME TAX ON REFERENCE YEAR
A27	N01(5-10)	5-10	SECURITY AND EMPLOYMENT RELATED PAYMENTS: INSURANCE, PENSION PLAN...
A4	C08(2,3,4,7,17)	2,7	RENTED LIVING QUARTERS, RENT, ADDITIONS-RENOVATIONS-ALTER ATIONS, OTHER (SECURITY DEPOSIT), INSURANCE, PARKING
A5	C09(1-7,9)	1-7,9	SHELTER EXPENSES (OWNERS AND RENTERS): WATER-GAS-ELECTRICITY,HEAT
AZ1	C01(2,3,5)	2,3,5	OWNED LIVING QUARTERS: TAXES, INSURANCE, CONDO 'S CHARGES
AZ1	C02(1-40)	3,11-20,25,27,28,35,40	OWNED LIVING QUARTERS: ADDITIONS, RENOVATIONS AND ALTERATIONS , INSTALLATIONS AND REPLACEMENTS

APPENDIX 1. DESCRIPTION OF VARIABLES

VARIABLE	QUESTIONNAIRE SECTION(CELLS)	CELLS USED TO DETERMINE IF THE VARIABLE SHOULD BE CONSIDERED MISSING	DESCRIPTION
AZ1	C07(8,14,15)	8,14,15	OWNED LIVING QUARTERS: TRANSFER TAXES, LEGAL CHARGES, OTHER(SURVEYING..)
AZ31	G01-G02-G03(ALL)		CLOTHING EXPENDITURES NOT INCLUDING INFANTS NOT YET BORN
AZ31	D04(21,23)		SMALL ELECTRICAL APPLIANCES PERSONAL CARE
AZ31	H01(1-24)		PERSONAL CARE EXPENDITURES
AZ31	I01(2-22)	2-5,6-10,13,14,18-20	MEDICAL AND HEALTH CARE EXPENSES
AZ32	E01(1-22)	1-4,6,8,11,12,14,15,19	RUNNING THE HOME: COMMUNICATION, CHILD CARE EXPENSES, DOMESTIC SERVICES , GARDEN SUPPLIES, PET EXPENSES
AZ32	E02(4-27)		RUNNING THE HOME: CLEANING SUPPLIES, STATIONARY PRODUCTS, OTHER...
AZ32	F01(1-3,5-19)	1-5,5-8,11-19	FOOD AND ALCOHOL EXPENSES: FROM STORE AND RESTAURANT
AZ32	F02(3,4)		FOOD AND ALCOHOL EXPENSES: FOR BOARD
AZ32	L01(1-4)	1,4	TOBACCO AND SMOKERS' SUPPLIES
AZ4	J01(10,12)	10,12	TRANSPORTATION: LEASING COST,PURCHASE PRICE
AZ4	J02(1-16,161-168)	1-4,7-16,161-168	TRANSPORTATION: OPERATION OWNED AND RENTED CARS AND TRUCKS
AZ4	J03(1-3)	1	TRANSPORTATION: DRIVING LICENCES, TESTS, LESSONS
AZ4	J04(1-11)	1,3,6,7,9	TRANSPORTATION: TRANSPORTATION SERVICES LOCAL, COMMUTER AND INNER CITY
AZ5	D02(8-EVEN...-16, 15-ODD...-27,24-26 -28,30-36)	8,10,12,17,19,21,23,30-33, 35	HOUSEHOLD FURNISHING AND EQUIPMENT: HOME ENTERTAINMENT EQUIPMENT

APPENDIX 1. DESCRIPTION OF VARIABLES

VARIABLE	QUESTIONNAIRE SECTION(CELLS)	CELLS USED TO DETERMINE IF THE VARIABLE SHOULD BE CONSIDERED MISSING	DESCRIPTION
AZ5	J03(4-6,9-15,19)	9-15	TRANSPORTATION: RENTALS OTHER THAN CARS, PURCHASE OTHER VEHICLE (NOT CARS), SALE OF ...
AZ5	J04(14,18,22,25)	14,18,22,25	TRANSPORTATION: TRAVEL TOURS PACKAGES
AZ5	K01(1-ODD...35,36 -38)	25-odd...-35,38	RECREATION EQUIPMENT: SPORTS, CAMPING, PHOTOS, MUSICAL INSTRUMENTS
AZ5	K02(1-ODD..-21,22 ,24,26-34)		RECREATION EQUIPMENT (OTHER) AND SERVICES
AZ5	K03(1-31)	2,6-9,11-15,25-27,29,30	RECREATION SERVICES, READING MATERIAL, EDUCATION
AZ5	C11(22,26)		SHELTER EXPENSES: INTEREST PAID, TAXES ON OTHER PROPERTY USED FOR PART OF THE YEAR
AZ5	L01(5-13)	5-9,11-13	MISCELLANEOUS EXPENSES: FINANCIAL SERVICES...
AZ5	N01(11)	11	SECURITY AND EMPLOYMENT RELATED PAYMENTS: DUES TO UNION
AZ6	N01(3,4,12-16)	3,4,12-16	PERSONAL TAXES, SECURITY, GIFTS: INCOME TAX RECEIVED BEFORE REFERENCE YEAR, OTHER TAXES REFERENCE YEAR, MONEY GIVEN (GIFTS)
AZ6	G03(2-11)		CLOTHING EXPENDITURES: FOR INFANTS NOT YET BORN
AZ7	C09(13-23)	18,19	SHELTER EXPENSES (OWNERS AND RENTERS): ACCOMMODATION(HOTELS...,
AZ7	C11(4,5,7,8,12-16)		OWNED VACATION HOME: ADDITIONS-RENOVATIONS-INSTA LLATIONS, REPAIRS-MAINTENANCE, MORTGAGE PAYMENT,INTEREST PAID, TAXES-INSURANCE -ELECTRICITY-FUEL

APPENDIX 1. DESCRIPTION OF VARIABLES

VARIABLE	QUESTIONNAIRE SECTION(CELLS)	CELLS USED TO DETERMINE IF THE VARIABLE SHOULD BE CONSIDERED MISSING	DESCRIPTION
B6	M01(4,5)	4	PERSONAL INCOME: GROSS WAGES AND SALARIES, MILITARY PAY AND ALLOWANCE
BZ1	C08(5,19)	19	SHELTER EXPENSES: RENT RETURNED, RENT FOR BUSINESS
BZ2	J03(17,19)	17,19	TRANSPORTATION (OTHER VEHICLES): OPERATING COST CHARGED TO BUSINESS, AMOUNT RECEIVED FOR SALE OF VEHICLE
BZ2	J01(14,18)	14,18	TRANSPORTATION (AUTOMOBILES AND TRUCK): AMOUNT RECEIVED FOR SALE OF VEHICLE, OPERATION COST CHARGED TO BUSINESS
BZ2	D01(34)	34	HOUSEHOLD FURNISHING AND EQUIPMENT: AMOUNT RECEIVED FOR SOLD APPLIANCES
BZ3	M01(6,8,11,12-15)	6,8,11,12-15	PERSONAL INCOME: SELF EMPLOYMENT INCOME NON-FARM AND FARM, INCOME FROM BOARDERS, INTEREST, DIVIDENDS, INVESTMENTS
BZ4	M01(16)	16	PERSONAL INCOME: FAMILY ALLOWANCE
BZ4	M01(17)	17	PERSONAL INCOME: OLD AGE SECURITY
BZ4	M01(18)	18	PERSONAL INCOME: PENSION PLANS
BZ4	M01(19)	19	PERSONAL INCOME: UIC
BZ4	M01(20)	20	PERSONAL INCOME: SOCIAL ASSISTANCE
BZ4	M01(21)	21	PERSONAL INCOME: OTHER INCOME FROM GOVERNMENT SOURCES...
BZ4	M01(22,23)	22,23	PERSONAL INCOME: RETIREMENT PENSIONS, ...OTHER
BZ5	M01(24-25)	24,25	PERSONAL INCOME: OTHER MONEY RECEIPTS (GIFTS, INHERITANCE)

APPENDIX 1. DESCRIPTION OF VARIABLES

VARIABLE	QUESTIONNAIRE SECTION(CELLS)	CELLS USED TO DETERMINE IF THE VARIABLE SHOULD BE CONSIDERED MISSING	DESCRIPTION
BZ5	M01(26)	26	PERSONAL INCOME: PREPAYMENT OF CHILD TAX CREDITS
BZ5	M01(27)	27	PERSONAL INCOME: GST CREDIT
BZ5	M01(28)	28	PERSONAL INCOME: TAX REFUND
C1	P01(1,3,5,7)	1,3	CHANGE IN HOUSEHOLD FINANCIAL POSITION: NET CHANGE OF ASSETS IF A NET INCREASE
C4	P01(18,19)	18,19	CHANGE IN HOUSEHOLD FINANCIAL POSITION (BUSINESS OR FARM): REPAYMENTS ON PRINCIPAL OF MORTGAGE, PURCHASE PRICE OF ASSETS
C7	C06(1-3)	3	SHELTER EXPENSES: MORTGAGE ON OWNED LIVING QUARTERS - INTEREST AND PRINCIPAL PAID - PREMIUM PAID ON LIFE INSURANCE
C9	P02(10-11)	10,11	CHANGE IN HOUSEHOLD FINANCIAL POSITION (LOANS AND OTHER DEBTS): AMOUNT OF INTEREST AND PRINCIPAL PAID
CZ1	P01(9)		CHANGE IN HOUSEHOLD FINANCIAL POSITION (CONTRIBUTIONS): RRSP
CZ1	P01(11,13,15)	11,13,15	CHANGE IN HOUSEHOLD FINANCIAL POSITION (CONTRIBUTIONS): BONDS, STOCKS, SHARES
CZ2	P02(121-126)	121-126	CHANGE IN HOUSEHOLD FINANCIAL POSITION (LOANS AND OTHER DEBTS): DIFFERENCE IN MONEY OWED BETWEEN JAN AND DEC- IF JAN LARGER
CZ2	P02(133-138)	133-138	CHANGE IN HOUSEHOLD FINANCIAL POSITION (LOANS AND OTHER DEBTS): AMOUNT OF INTEREST CHARGES

APPENDIX 1. DESCRIPTION OF VARIABLES

VARIABLE	QUESTIONNAIRE SECTION(CELLS)	CELLS USED TO DETERMINE IF THE VARIABLE SHOULD BE CONSIDERED MISSING	DESCRIPTION
CZ3	C07(3)		SHELTER EXPENSES: PURCHASE PRICE OF HOME BOUGHT IN REFERENCE YEAR
CZ3	C07(11)	11	SHELTER EXPENSES: REAL ESTATE COMMISSIONS ON HOME SOLD IN REFERENCE YEAR
CZ3	C11(3,18,19,21)		SHELTER EXPENSES: OWNED VACATION HOME - PURCHASE PRICE; OTHER - PURCHASE PRICE, ALTERATIONS, MORTGAGE
D1	C05(7)	7	SHELTER EXPENSES: AMOUNT OF PRINCIPAL BORROWED OR RENE WED
D13	P02(127-132)	127-132	CHANGE IN HOUSEHOLD FINANCIAL POSITION (LOANS AND OTHER DEBTS): DIFFERENCE IN MONEY OWED BETWEEN JAN AND DEC-IF DEC LARGER
D2	C07(10)	10	SHELTER EXPENSES: SELLING PRICE OF HOME SOLD IN REFERENCE YEAR
D3	C11(6,11,20,25)		SHELTER EXPENSES (OWNED VACATION HOME AND OTHER): - AMOUNT BORROWED, AMOUNT RECEIVED FROM SALE
D4	P01(2,4,6,8)	2,4,6,8	CHANGE IN HOUSEHOLD FINANCIAL POSITION: NET CHANGE OF ASSETS IF A NET DECREASE
DZ1	P01(10,12,14,16)	10,12,14,16	CHANGE IN HOUSEHOLD FINANCIAL POSITION (WITHDRAWALS): RRSP, BONDS, STOCKS, SHARES
DZ1	P01(17)	17	CHANGE IN HOUSEHOLD FINANCIAL POSITION (WITHDRAWALS): SALES OF OTHER PERSONAL PROPERTY
DZ2	P01(20)	20	CHANGE IN HOUSEHOLD FINANCIAL POSITION (WITHDRAWALS BUSINESS OR FARM): MONEY BORROWED

APPENDIX 1. DESCRIPTION OF VARIABLES

VARIABLE	QUESTIONNAIRE SECTION(CELLS)	CELLS USED TO DETERMINE IF THE VARIABLE SHOULD BE CONSIDERED MISSING	DESCRIPTION
DZ2	P01(21)	21	CHANGE IN HOUSEHOLD FINANCIAL POSITION (WITHDRAWALS BUSINESS OR FARM): SELLING PRICE OF ASSETS
DZ2	P01(22)	22	CHANGE IN HOUSEHOLD FINANCIAL POSITION (WITHDRAWALS BUSINESS OR FARM): CAPITAL COST ALLOWANCE
DZ3	P02(2)	2	CHANGE IN HOUSEHOLD FINANCIAL POSITION (LOANS AND OTHER DEBTS): ORIGINAL PRINCIPAL BALANCE
DZ3	P02(4)	4	CHANGE IN HOUSEHOLD FINANCIAL POSITION (LOANS AND OTHER DEBTS): ADDITIONAL MONEY BORROWED ON THE LOAN

APPENDIX 2. LIST OF EDITS

FIRST STAGE	LIST OF EDITS MODIFIED IN THE SECOND STAGE	LIST OF EDITS MODIFIED IN THE THIRD STAGE
BALANCE EDIT 0.85(BZ1 + BZ2 + BZ3 + BZ4 + BZ5 + B6) + DZ2 + DZ3 + D1 + D2 + D3 < = + AZ1 + AZ31 + AZ32 + AZ4 + AZ5 + AZ7 + A11 + A23 + A24 + A27 + A4 + A5 + CZ1DZ1 + CZ2D13 + CZ3 + C1D 4 + C4 + C7 + C9 < = 1.176(BZ1 + BZ2 + BZ3 + BZ4 + BZ5 + B6) + DZ2 + DZ3 + D1 + D2 + D3		
0 < A11 < = 81822		
0 < = A24 < 75000		
0 < A5 < = 5715	0 < = A5 < = 5715	
0 < AZ4 < = 48498	0 < = AZ4 < = 48498	
0 < = AZ5 < = 50871		
0 < = BZ2 < = 22000	0 < = BZ2 < = 22000	
0 < = BZ3 < = 159617		
-85000 < = C1D4 < = 455489		
0 < C7 < 121239		0 < = C7 < 121239 *
0 < = C9 < = 17090		
-33000 < = CZ2D13 < = 9500		
0 < = D1 < = 160000		
D1 RECIPIENT = 60% OF D1 DONOR	EDIT REMOVED	
0 < = DZ2 < = 562000		

Edits defining upper and lower bounds were defined based on the maximum and minimum value for each variable observed in the final FAMEX data (standard database). This is of particular importance when all the records are treated simultaneously during the error localization module (see GEIS manual). It allows the system to determine if the missing variables alone could theoretically make the record pass the balance edit given that the missing values can only vary within a predefined range. If it is not the case, one or more variables will be flagged for imputation.

APPENDIX 3. WEIGHTS

Weights were defined based on two criteria: 1) the percent change of a particular variable from the red stage to the final stage, and 2) how much this change affected the total budget (sum of all variables). Since the goal is to find a solution that involves the minimum change (the minimum number of variables involved) twice as much importance was given to the second criterion. As well, considering the nature of the balance edit, weights were designed in order to allow imputation of variables that represent either inflows or outflows of money.

GROUP	THE BALANCE EDIT IS OF THE FORM : (DISBURSEMENT + CREDITS) - (RECEIPTS + DEBITS)				WEIGHT
	DISBURSEMENTS	CREDITS	RECEIPTS	DEBITS	
1	AZ4	C7, C9, C1		D4, D1	1
2	AZ5, A11, A24		BZ3	DZ2	3.1
3		C4, CZ2	BZ2, BZ4, BZ5	D13	6.3
4	A4, A5, A23, A27, AZ1, AZ31, AZ32, AZ6, AZ7	CZ1, CZ3	B6, BZ1	D2, D3, DZ1, DZ3	6.4

These are the non missing variable that were flagged for imputation by the system according to the predefined weights.

VARIABLE AGGREGATE	VARIABLE RANGE	NUMBER OF TIMES THAT IT WAS PICKED FOR IMPUTATION BY THE SYSTEM
C1D4	-85,000 to 455,489	10
AZ4	0 to 48,498	5
C7	0 to 121,239	4
C9	0 to 17,090	3
D1	0 to 160,000	3

APPENDIX 4. CLUSTER ANALYSIS

VARIABLE(S) REQUIRING IMPUTATION (NUMBER OF TIMES)	MATCH FIELDS
A11 (1), AZ4 (14), AZ5 (14)	A11, AZ4, AZ5, AZ31, AND AZ32
AZ6 (2) OR AZ7 (2)	A11, AZ4, AZ5, AZ31 AND AZ32, AZ6, AZ7
A5 (11), C7 (55), CZ3 (1), D1 (3)	A4, A5, AZ1, C7, CZ3, D1, D2
A24 (1), A27 (2)	A24, A27, B6
C4 (1), DZ2 (2)	BZ1, BZ3, C4, AND DZ2
C9* (24)	A24, A27, B6
BZ2* (9)	A11, AZ4, AZ5, AZ31, AND AZ32
A23 (1)	NONE IDENTIFIED
BZ5 (2)	NONE IDENTIFIED
CZ2D13 (17)	NONE IDENTIFIED
C1D4 (10)	NONE IDENTIFIED
DZ3 (1)	NONE IDENTIFIED

* Matching fields were not identified for this variable, however the matching fields defined in the closest cluster were used in the distance calculation when that variable required imputation.