4. Data Editing and Imputation

This chapter describes the data editing and imputation procedures applied to data from the Survey of Income and Program Participation (SIPP) after completion of the interviews. Three different approaches are used for dealing with missing data in SIPP:

- Weighting adjustments are used for some types of noninterviews;
- Data editing (also referred to as logical imputation) is used for some types of item nonresponse; and
- Statistical (or stochastic) imputation is used for some types of unit nonresponse and some types of item nonresponse.

Weighting is discussed in Chapter 8.

The chapter begins with a brief discussion of the types of missing data and the goals of imputation in SIPP. It then presents an overview of the editing and imputation procedures used to deal with missing and inconsistent data. Next, the chapter provides a detailed description of each of the major steps used by the Census Bureau when creating its internal files and the files that are released for public use. Prior to 1996 the development of cross-sectional wave files involved mainly cross-sectional editing and imputation. The longitudinal files involved longitudinal editing. Beginning with the 1996 Panel, the processing procedures may also include methods that use prior wave information to edit and impute a current wave (after wave 1). The most common imputation technique, the hot-deck method, is still used in the 1996+ Panels. A new procedure allows donors, when appropriate, chosen on the basis of similarities in reported *prior* wave information when that reported information exists for certain variables. In panels prior to the 1996 Panel, the donors were chosen based only on current wave similarities.

The SIPP Web site (http://www.sipp.census.gov/sipp/) supplements the information in this chapter with detailed information about all variables on the public use files. To obtain more detailed information about imputation and editing procedures, contact the Demographic Surveys Division's (DSD) Income Programming Surveys Branch, 301-763-5244.

Types of Missing Data

As in all surveys, there are two general types of missing data in SIPP: (1) unit nonresponse and (2) item nonresponse. Unit nonresponse occurs in SIPP when one or more of the people residing at a sample address are not interviewed and no proxy interview is obtained. This can happen for a number of reasons, described in Chapter 2. Most types of unit nonresponse are dealt with through weighting adjustments (see Chapters 2 and 8). However, the data editing and statistical imputation procedures described in this chapter are used with one type of unit nonresponse: Type Z

noninterviews. Type Z noninterviews are cases where an interview was obtained from at least one Household member but interviews were *not* obtained from one or more other sample persons in that household.1 Prior to the 1996 Panel and in some instances in the 1996 Panel, the method used to adjust for person-level noninterviews in the core wave files is known as *Type Z imputation*, which is discussed below. Chapter 2 discusses person-level nonresponse, Type Z.

The other type of missing data is, item nonresponse. This occurs when a respondent completes most of the questionnaire but does not answer one or more individual questions. Item nonresponse data in SIPP occur under the following circumstances:

- Respondents refuse or are unable to provide requested information;
- Interviewers fail to ask a question or incorrectly record a response;
- A response is inconsistent with related responses or is incompatible with response categories; and
- Interviewers make an error when recording or keying in the data.[2]

Item nonresponse data are usually imputed for core items, as well as for many topical module items.

Goals of Imputation

Missing data cause a number of problems:

- Analyses of data sets with missing data are more problematic than analyses of complete data sets
- There is a lack of consistency among analyses because analysts compensate for missing data in different ways and their analyses may be based on different subsets of data
- In the presence of nonresponse that is unlikely to be completely random, estimates of population parameters are biased.

Because missing data are always present to some degree, analyses of survey data must be based on assumptions about patterns of missing data. When missing data are not imputed or otherwise accounted for in the model being estimated, the implicit assumption is that data are missing at random after controlling for other variables in the model. The imputation procedures used for SIPP are based on the assumption that data are missing at random within subgroups of the population (as

[1] That can happen because people refuse to be interviewed or they are unavailable and a proxy is not obtained.

[2] Prior to the 1996 Panel, errors could also occur when data-entry workers were keying in results from the paper survey.

defined by the cells of the imputation matrices described later in the chapter).

The statistical goal of imputation is to reduce the bias of survey estimates. This goal is achieved to the extent that systematic patterns of item nonresponse are correctly identified and modeled. In SIPP, the statistical goals of imputation are general, rather than specific. Instead of addressing the estimation of specific Parameters, SIPP procedures are designed to provide reasonable estimates for a variety of analytical purposes.

Data editing is generally preferred over statistical imputation, and it is used whenever a missing item can be logically inferred from other data that have been provided. The advantage of data editing is that it avoids the increase in variance that occurs when missing items on one record are imputed with nonmissing responses from other records.

Assessing the Influence of Imputed Data on Analysis

Users of SIPP data interested in assessing the influence of imputed data on their analyses should consider whether SIPP imputation procedures have properties that affect their specific analytical requirements. A general discussion of the treatment of missing data in sample surveys is given in Kalton and Kaspyrzyk (1986). Sedransk (1985), Little (1986), and Jinn and Sedransk (1987) discuss properties of commonly used imputation processes. An example of the impact of imputation procedures for the WIC program is discussed in CNSTAT, 2003. A report discussing sources of error for federal data collection programs is given in A *Statistical Policy Working Paper, 31, June 2001.*

An evaluation of the effects of imputed data should include a review of rates of unit nonresponse and an assessment of the extent of item nonresponse. Unit nonresponse tends to increase over the life of a panel, as does the likelihood that nonresponse is not a random effect. As the percentage of eligible sample members reinterviewed decreases, the pool from which donors[3] are selected shrinks accordingly. This smaller pool of donors leads to an increased likelihood that individual donors will be used more than once, which in turn increases the variance of an estimate.

The effects of imputation will likely be small for items with low rates of missing data as long as rates of item nonresponse are not high among important subclasses. Lepkowski et al. (1987), using data from a large federal survey, provide a framework for evaluating the effect of imputed values on analyses. This framework can be readily adapted to SIPP analyses.

Imputation Methods

The SIPP primarily uses two methods to impute missing data. The hot-deck method, used for item

[3] Cases with complete data that are the source of the imputed values placed on the records with missing data.

non-response, and the Type Z method, used for unit non-response. Item non-response refers to missing items within an interviewed case. Unit non-response refers to a non-interviewed case within an interviewed household. SIPP also uses another method in rare circumstances, logical imputation, where the imputation is logically derived.

The hot-deck method replaces individual missing data items with reported data from another person or household with similar characteristics. Initially, the input file is sorted by geographical keys: PSU, Segment, and Serial Number; this ensures that neighboring records represent geographically proximate units. Edits and imputations are then performed sequentially by unit for each topical section: demographics, household characteristics, labor force, assets, general income, health insurance, and program participation. Each section is processed completely before the next section is done. A hot deck array is created for each edited variable and is stratified by selected variables such as age, race, sex, etc.. Hot decks are first initialized with cold deck values then they are loaded with data provided by the respondent by passing through the data one time. The data are then passed a second time with good responses contributing to the hot deck and missing responses allocated from the hot deck. Each hot deck cell will contain exactly one value at any point in the edit: either the cold deck value or the most recently encountered good value meeting the same criteria for that cell - as defined by the stratifying variables. The hot deck imputation process as currently implemented is fully deterministic: subsequent re-processing using the same file and same edit program will result in identical imputations.

Type Z imputation method involves imputing an entire set of data from a single donor. This is used primarily for non-interviewed persons within an interviewed household. The Type Z procedure is based on a hierarchical sorting and matching operation based on a set of variables that are non-missing for both recipient and donor. The matching variables used are age, race, sex, marital status, household relationship, education, veteran status, parent/guardian status, and income and asset sources. The match is designed to progressively broaden ranges with the above match keys until a match is found. When a match is found, all data are transferred to the recipient record except for identification variables and other variables that may not be relevant within the recipient household. A second Type Z operation is used within the labor force edit to impute a set of labor force characteristics from a single donor (this is referred to as a *little type Z*). See the section, Type Z Imputation for Core Items in the Core Wave Files, for more information.

An Overview of the Process

The processing of SIPP data has traditionally been done cross-sectionally by wave and then longitudinally across waves once all waves were available. In 1996 this process was changed to apply selected longitudinal edits to individual wave files and in 2004, most longitudinal edits were discontinued.

For the pre-1996 panels, there are two phases to the processing of SIPP data. The first phase occurs at the conclusion of each wave of interviewing, then the data collected during that wave are processed, creating the core wave and topical module files. The second phase occurs at the conclusion of the final wave of interviews, core data from all waves are linked and a new set of edit and imputation procedures is applied to the resulting full panel file.

For the 1996+ panels, there are also two phases, however the second phase does not involve the creation of a full panel file. Waves 1-4 are edited cross-sectionally as they become available; this is phase one. Then, once wave 4 is complete, a longitudinal edit of selected demographic variables is done across the four waves. These variables are then placed on each of the individual wave files. For wave 5+, these variables are not re-edited, but are simply pulled forward from the previous wave. For 2004+ most longitudinal editing was abandoned. Previous wave data is still used to fill missing data, however no attempt is made to enforce consistency across waves.

Phase 1 - Summary

There are six steps in the first phase of SIPP data processing:

1. As each wave of interviewing is completed, core data collected during the wave are edited for internal consistency.

2. Following data editing, the statistical matching and hot-deck procedures described later in this chapter are used to impute missing data from the core wave file.

3. A public-use version of the core wave file is created from the internal core wave file. The public-use file is the same as the Census Bureau's internal file except that it has certain information suppressed or topcoded to protect the confidentiality of survey respondents (see sections on Topcoding and Suppression of Geographic Information, at the end of this chapter).

4. On a separate production track from the core data, data from the topical module file administered with the wave are edited for internal consistency. The extent of data editing varies across the topical modules, and some topical modules receive almost no editing.

5. Next, hot-deck procedures are used to impute missing data in the topical module. The extent of imputation varies across the topical modules; some topical modules have no missing data imputed.

6. A public-use version of the topical module file is created from the internal file. As with the public-use core wave files, the public-use topical module files have certain information suppressed to protect the confidentiality of survey respondents.

Figure 4-1 illustrates the steps that generate the Census Bureau's internal core wave and full panel files.

These steps are repeated at the conclusion of each wave of interviews. Prior to the 1996 Panel, each wave was processed independently of other waves of data. Thus, when multiple core wave files are linked, apparent changes in a respondent's status could be due to different applications of data edits and imputations to the files being combined (file linkage is the subject of Chapter 13). With the 1996 data, the hot-deck procedure was redesigned to rely on historical

information reported in prior waves. In addition, other forms of longitudinal imputation, such as carryover methods, were adapted**.**

Figure 4-1. Sequence of Cross-Sectional Imputation and Longitudinal Editing Procedures

<table>
<tr><td>Imputation of Sample Unit Characteristics (Tenure, etc.)</td><td rowspan="2">Imputation of Item Missing Data for Sample Unit Characteristics and Personal Demographic Characteristics</td><td rowspan="8">Sequence is repeated for each wave in a panel</td></tr>
<tr><td>Imputation of Personal Demographic Characteristics (Age, Race, Marital Status)</td></tr>
<tr><td>Type Z Imputations</td><td>Imputation of Person-Level Noninterviews</td></tr>
<tr><td>Imputation of Labor Force Items and Recipiency of Income and Assets</td><td rowspan="5">Imputation of Item Nonresponse in Core Questions</td></tr>
<tr><td>Imputation for Item Nonresponse in Records for others Cash Income</td></tr>
<tr><td>Imputation for Item Nonresponse in Self-Employment Identification Sections</td></tr>
<tr><td>Imputation for Item Nonresponse in Asset Sections (Property Income)</td></tr>
<tr><td>Imputation for Item Nonresponse for Household Program Information</td></tr>
<tr><td>Editing for Demographic and Household Variables, Employment Variables, General Amount Variables**,** and Other Variables</td><td>Editing of Longitudinal Record</td></tr>
</table>

For 1996+ panels, type Z records only handled in a separate process if no previous wave data are available.

The imputation procedure for SIPP Panels 1996+ allows for item imputation from previous wave's data if the previous wave's data had valid data regardless if went through a hot deck procedure. In these situations, an allocation flag of 3 was assigned. One advantage of using prior wave data instead of using a hot deck procedure to impute is that the data are more consistent from wave to wave. A disadvantage is that a particular donor has potential to be an influence each wave thereafter.

Phase 2 Summary - Pre 1996 panels

At the conclusion of the panel, the Census Bureau creates a full panel file containing core data from all waves. There are four steps to this process.

1. Core data from all waves are linked. Those data have already been subjected to the Phase 1 edit and imputation procedures.

2. A series of longitudinal edits are applied to the full panel file. Unlike the core wave edit procedures, these edits are designed to create longitudinally consistent records for each person.

Both reported values and values that were imputed during the first phase of processing are subject to change. Thus, the data in a full panel file may differ from the data in the core wave files from which the full panel file was constructed.

3. A missing wave imputation procedure is then applied. Data are imputed when a sample member was absent for one wave but was present for the two adjacent waves. Data for the missing wave are interpolated on the basis of information from the fourth month of the prior wave and the first month of the subsequent wave. The missing wave imputation procedure was introduced with the 1991 Panel. Earlier panels were not subjected to this procedure.

4. A public-use version of the full panel file is created from the internal file. The public use file has certain information suppressed to protect the confidentiality of survey respondents.

Phase 2 Processing – 1996 to 2001 Panels

1. Core data from waves 1- 4 are linked. Those data have already been subjected to the Phase 1 edit and imputation procedures.

2. Demographic and household composition variables are edited to ensure consistency across the 4 waves. Waves 1-4 are re-processed using the longitudinally edited values.

3. Waves 5+ are processed cross-sectionally as they become available; demographic and household composition variables are pulled forward from longitudinally edited values in the previous wave.

Note that no full panel files are created for the 1996+ panels.

Phase 2 Processing – 2004 Panel

Only cross-sectional edits and imputation procedures were applied. There were no longitudinal edits of demographic and household composition variables.

The balance of this chapter describes in greater detail the full sequence of data edit and imputation procedures applied to SIPP data files. Most of the material contained in this chapter is taken from Pennell (1993).

The data processing sequence for each wave is detailed below.

Data Entry and Initial Editing

Beginning with the 1996 Panel (Chapter 2), all of the data entry and some of the initial data editing are performed by computer-assisted interviewing while the interview is in progress. Before the 1996 Panel, the first stages of data processing involved editing the paper questionnaires for completeness, reasonableness, and consistency. Those data checks were conducted first by field representatives before they submitted their questionnaires to the regional offices and then by the regional and central offices of the Census Bureau. The next step was data entry, in which clerks keyed in the information from control cards and questionnaires. Edits were built into the data-entry program to ensure that the data were keyed in the proper sequence and that certain key identifiers, such as control number, name, and relationship to householder, were present. Following this step, the data files were transmitted electronically to Census Bureau headquarters.

Imputation for Sample Unit Characteristics and Personal Demographic Characteristics

Items in this category, including housing tenure (owned or rented), age, race, marital status, and so forth, must be present for any further data processing to take place. If these values cannot be logically derived, they are imputed. The imputation procedure is a modified version of the sequential hot-deck procedure described below.

Type Z Imputation for Core Items in the Core Wave Files

Pre-1996 Panels. Type Z imputation was the method used in the pre-1996 panels to impute core items for person-level noninterviews. There are two categories of person-level noninterviews subject to imputation for the core questions. The first category includes individuals 15 years of age and older who were members of interviewed households at the beginning of the 4-month reference period but were not original sample members or members of any SIPP-interviewed household on the date of the interview that is, people not interviewed because they moved out of the sample household between the beginning of the reference period and the interview date. Had these people been original sample members, they would be interviewed at their new address.

Rather, these are all people who entered the SIPP sample after the first wave and were in the sample because at some point they were living with an original sample member.

The second category of imputed noninterview includes people 15 years of age or older who were members of SIPP-interviewed households on the date of the interview and during all or a portion of the 4-month reference period but who were not interviewed because they refused to cooperate or were unavailable for the interview and a proxy interview was not obtained.

The Type Z imputation procedure is based on a hierarchical sorting and merging operation that matches noninterviews with respondents on socioeconomic characteristics available for both. The

variables used to match noninterviews with respondents are age, race, gender, marital status, household relationship, education, veteran status, parent/guardian status, and income and asset sources. Pennell (1993, Figure C-1) provides a table of variables used to match recipients with donors. The Type Z imputation procedure is designed to always find a match. Type Z noninterviews are imputed by assigning values from the matching donor to the noninterview record. The donor values are assigned in full, except for identification variables or other variables not relevant for the household in which the noninterview occurred. Pennell (1993) gives a complete account of Type Z imputation, including detailed descriptions of matching operations.

For 1996+ Panels, the Type Z procedure is only used where Type Z persons do not have an interview record available in the previous wave, i.e. in wave 1, for new respondents in wave 2+, or where person was a non-interview in the previous wave. For all others, general imputation procedure (the sequential hot-deck procedure described in the following pages) is used to impute core items for most person-level noninterviews.

Imputation of Item Nonresponse in Core Questions

SIPP core items are imputed in the following order:

1. Labor force participation, recipiency of income, and asset holdings;

2. Other cash income;

3. Wage, salary, and self-employment income amounts;

4. Asset income amounts; and

5. Program participation and benefits.

The Sequential Hot -Deck Imputation Procedure

The statistical imputation method used to impute missing items from the core questions and topical modules is known as a *sequential hot-deck procedure*.45 In a general sense, the sequential hot-deck procedure, like the Type Z imputation procedure, matches a record with missing data to that of a donor with similar background characteristics and uses the donor 's values. This procedure differs from data editing, which replaces missing data with inferred values based on nonmissing data from the *same* case.

[4] The hot-deck procedure used in SIPP for the core questions and topical module items is sequential because the selection of replacement values is implemented one record at a time from an ordered file.

The sequential hot-deck procedure used in SIPP involves five key steps:

1. Specifying cold-deck or initial donor values;

2. Sorting the sample cases;

3. Identifying records with no item nonresponse and updating hot-deck values;

4. Classifying cases into subclasses of the population, referred to as imputation classes or adjustment cells, according to values on a set of classification or auxiliary variables that are nonmissing for all cases (this step is omitted in the initial processing of the key demographic items' race, gender, etc.); and

5. Selecting replacement values from donor cases to impute item-missing data on recipient records.

Two types of sequential hot-deck imputation are used to provide values for missing items. In Wave 1 and for each sample member who is new to a subsequent wave, the hot deck is cross-sectional; only values from current wave responses are used in the definition of the hot-deck cells. Beginning with Wave 2, previous wave values are included in the definition of the hot deck cells. In both instances, however, only current wave values from selected donors are used to replace missing items (with several exceptions, described below). Longitudinal (or previous wave) hot-deck imputation was not performed prior to the 1996 Panel. Each wave received only the cross-sectional hot-deck imputation. For example, the item indicating whether a person worked part-time in the reference period for the wave (a dichotomous item) uses the longitudinal hot deck for old sample members and the cross-sectional hot deck for new sample members. The 1996 Panel cross-sectional hot-deck imputation is based on a cell structure with 288 cells that are based on cross-classifications of sex (two categories), race (two categories), age (six categories), marital status (three categories), disability status (two categories), and presence of own children (two categories). On the basis of his or her current wave values for those categories, each new sample member in any later wave is assigned to a cell; then the donor 's value in that cell is used to impute a value to the new sample member.

The longitudinal hot-deck imputation for the part-time work item for old sample members in Waves 2+ is based on a cell structure with 576 cells that are based on the same categories described above with one extra category: whether or not the person worked part-time in the previous wave. A donor is selected from that cell, and that value is imputed. The actual item is imputed from a donor 's value of the item in the current wave; the previous wave value is used only in the assignment of the cell. That procedure guarantees that the sample member is matched to the donor who had the same value for the item in the previous wave. Therefore, sample members who worked part-time in the previous wave will be matched only to donors who also worked part-time in the previous wave. However, the actual hot-deck imputation comes from the donor 's value in the current wave, which may or may not include part-time work.

Imputed values for the sample member are allowed in assigning the cell for some items. If a sample member had an imputation for part-time work in the previous wave, that imputation is used to define

the cell for the longitudinal hot-deck imputation, even though it is an imputation itself. That is not done for other items, such as asset items. Only a nonimputed or logically imputed value counts toward the longitudinal hot deck for those items.

The part-time item is dichotomous; the previous wave imputation matrix was essentially the current wave imputation matrix with the previous wave's value of the item added to the matrix. In many cases, the differences between the two imputation matrices will be more pronounced, especially for items with several categories of answers. An example of this is the item reasons why person worked less than 35 hours in the reference period there are 12 categories for that item. The previous wave imputation matrix uses the following characteristics to define cells:

Previous wave value for item (12 categories);

- Sex (two categories);
- Race (two categories);
- Age (six categories);

The current wave imputation matrix uses the following characteristics to define cells.

- Sex (two categories);
- Race (two categories);
- Age (six categories);
- Marital status (three categories);
- Disability status (two categories);
- Presence of own children (two categories).

A different type of example is the item gross pay in the first month of the reference period. For new SIPP sample members, cross-sectional hot-deck imputation is carried out by using the following characteristics to generate cells:

- Industry and occupation category (16 categories);
- Sex (two categories);
- Hours worked (three categories); and

- Education level (three categories)

For old sample members, a longitudinal hot-deck imputation is carried out by using the previous wave value for the item gross pay in the fourth month of the preceding wave 's reference period.[6] This continuous value is divided into 138 categories, starting from $1 to $100, to over $50,000. Sample members are matched to donors by using the previous wave values of those categories.

For labor force items, the Census Bureau uses the following special imputation procedures when a person has no current wave information indicating whether or not he or she worked during the reference period. If the Census Bureau can infer from what it knows about the previous reference period whether the person had a job or business at the start of the current period, the Census Bureau carries out the following procedure:

1. If the person was working at the end of the prior wave, then labor force participation is imputed from a single donor for the complete current wave.

2. The Census Bureau then projects job characteristics for the person from the person's prior wave through the current wave.

3. Finally, the Census Bureau edits the job characteristics for consistency with the imputed labor force participation variables.

This procedure is known as an EPPFLAG imputation, after the name of the variable that indicates its use.

If a person was a nonworker in the prior wave or the Census Bureau cannot infer work status on the basis of prior wave data, then the person 's work status is imputed. If the person is imputed as a worker in the reference period, the Census Bureau imputes the complete set of job/business characteristics variables and labor force participation variables to the person from one donor, in order to maintain consistency among the fields. That procedure is called a little Type Z imputation.

For some items in some cases, a direct logical or carryover imputation is made. The carryover imputation takes the previous wave's value for the item for the sample member and imputes it to the current wave. That imputation is done particularly for items that rarely (or never) change for a sample member across waves (such as sex and race) or for items that change in predictable ways (such as age).

SIPP hot-deck procedures are designed to preserve the univariate distribution of each variable subjected to imputation. These procedures do not, in general, preserve the covariances among variables. Although some of those interrelationships might be preserved to a certain extent, that is

[6] The second month of the reference period actually uses as the "previous wave value" the first month value, with the third month using the second month, and so forth, so that these imputations are really previous month rather than previous wave

not the primary intent of the hot-deck imputation procedures used by the Census Bureau. One consequence is that imputation can introduce inconsistencies into the data. For example, if a respondent has reported program participation, but his or her income is too high for that program, it is possible that the income data have been imputed. Whenever users detect inconsistencies, it is wise to check the allocation (imputation) flag to see if the inconsistent data might have been imputed. The discussion of allocation (imputation) flags later in this chapter provides more information.

Starting or Cold -Deck Values

In other surveys, cold-deck values in a sequential hot-deck procedure historically served as the initial set of replacement values for missing items in the first record processed; missing items in subsequent records typically received replacement (hot-deck) values from the current data set. In SIPP, however, cold-deck values are seldom used as replacement values for either the first or subsequent records processed. During later stages of processing, as the cold-deck values are replaced with information from the current wave, the array of cells is referred to as the hot-deck matrix. The cells in the matrix are defined by the cross-classification of auxiliary variables (Pennell, 1993, Figure 3.3). Each cell in the matrix corresponds to respondent cases with the same set of values on the classification variables. Many different matrices are defined in SIPP, and each matrix corresponds to one or more variables subject to imputation.

Sorting the Sample Cases

The records in the sample file are sorted by three geographic variables prior to imputing item-missing data. The three geographic sort variables are primary sampling unit, segment number, and serial number. The cases are sorted prior to processing and are not re-sorted at any other time during the imputation process. The sorting operation creates a file in which neighboring records represent geographically proximate households.

Preprocessing the Sample File: Initial Updating of Cold-Deck Values

Once the cases have been sorted, they are processed through a series of programs. During the first pass against the programs, the cold-deck values are updated with information from the current wave; missing data are not imputed. The initial processing is done separately for each of the five groups of related core variables listed above. During the first pass, the first record in the sorted file with consistent and nonmissing data for a particular group of variables is identified and the values from that case replace the cold-deck values for that section in the matrix. The values for each subsequent record with consistent and nonmissing information update the previous set of consistent and nonmissing values written to the matrix. The checking and updating operation continues until all records in the data file have been processed. The last values written to the matrix serve as the starting values in the subsequent sequential hot-deck procedure. In this way, cold-deck values are rarely used as replacement values in SIPP because the initial processing usually replaces all starting values with values from the current wave of data.

Allocating Cases into Imputation Classes

In the next step of the imputation procedure, each respondent record or noninterview record in the sorted file is allocated to one of the imputation classes or adjustment cells according to its values on the set of classification, or auxiliary, variables.[7]

1. The auxiliary variables are chosen for each item or set of related items on the basis of their level of correlation with the item receiving the imputation (i.e., classification variables are chosen on the basis of their ability to explain the variability of the item or set of related items); Census Bureau researchers assign different sets of classification variables to different sets of items.

2. The auxiliary variables are either dichotomous or polychotomous categorical variables (e.g., sex, race); if they are continuous, they are categorized into a parsimonious number of levels (e.g., income, asset levels)

3. The level of the auxiliary variables then define a matrix, with the number of cells in this matrix being the product of the number of levels for each auxiliary variable. For example, an imputation defined by five variables, each with three levels, has a total of 243 cells. Any given item or set of related items may have imputation matrices with the numbers of cells ranging from under 100 to well over 1,000, depending on the matrix.

Auxiliary variables such as sex, race, and categorizations of age (with different categorizations for different items) are used frequently in the matrices, as are more specialized auxiliary variables that are relevant for particular items (such as industry and occupation category for the monthly gross pay item). Pennell (1993) gives examples of the different sets of classification variables for previous panel years.

The allocation of sample cases into imputation classes (also known as subclasses or strata) according to a set of classification variables serves several purposes. Ideally, the set of classification variables should account for a large proportion of the variance in the variable being imputed and should be associated with variations in response rates. To the extent that this is accomplished, the classification procedure creates homogeneous adjustment cells containing similar cases. In this way, donors and recipients are similar under the assumption that the nonresponse mechanism within the imputation class is not related to the item being imputed; *that is, an underlying assumption is made that item nonresponse data are distributed randomly within the subclass defined by the cross-classification of the auxiliary variables*. The selection of classification variables may also place bounds on the range of values that can be imputed and implicitly satisfy edit constraints. The implicit stratification created by the sort order of the file further improves the opportunity for better imputation to the extent that nearby cases are more similar to each other than cases that are farther apart in the file.

[7] This step is omitted for the imputation of the primary demographic values that are imputed before the person-level noninterviews.

Imputing for Missing Data and Updating of Hot-Deck Values

The selection of replacement values for missing items is restricted to donor and recipient records within each particular cell; that is, records allocated to one cell never donate information to records in another cell with missing items. As the file is processed through the set of programs the second time, the imputations are performed and the set of hot-deck values is updated once again.

The records are processed sequentially, according to the sort order of the file. A missing item is given the value of the last corresponding item that is nonmissing from a record in that imputation class. If the value of an item in the current record is nonmissing, it replaces the previous hot-deck value for that imputation class. In this way, the hot-deck value for each imputation class is constantly being updated with the value of the last nonmissing case.

The updating is done item by item. Missing items in one record receive the current set of replacement values. Then the nonmissing values in that record are used to update the hot deck in preparation for the next record. At any point during the process, the donated values in the hot deck likely come from many different respondents, even within imputation classes. That is why this imputation procedure does not preserve covariances among the variables being imputed.

Allocation (Imputation) Flags

An allocation (imputation) flag is associated with each core item subject to imputation. When an item has been imputed, an allocation (imputation) flag for that item is set. Beginning with the 1996 Panel, allocation flags denoting either data edits or statistical imputations for all variables are included on the core wave files. For core wave files from earlier panels, imputation flags are included for most items subject to imputation.

One type of variable that does not have an allocation flag, are the recode variables. SIPP produces recodes that combines variables to produce one estimate. Recoded variables do not have imputation flags. These variables assist users who are interested in summarizing related questions. For example, the Total Household Income variable, is a recode variable of more than sixty possible income sources. Using recodes cuts down the amount of programming significantly.

An allocation (imputation) flag with the value 0 indicates no imputation, a value of 1 indicates a hot-deck imputation that uses only current wave values, a value of 2 indicates a cold deck value, a value of 3 indicates a logical imputation, and for panels 1996+, a value of 3 may also indicate data from the previous wave was carried over to the current wave, and finally, a value of 4 indicates a dependent imputation. This last category includes imputations in which data have been carried over from the sample unit's previous wave data and imputations in which previous wave data are used as control variables. For detailed documentation about the coding of allocation (imputation) flags for specific variables, analysts can refer to the data dictionary for the data file with which they are working.

For items that receive Type Z imputations (in both the pre-1996 panels and the 1996+ Panels) and items receiving EPPFLAG and little Type Z imputations in the 1996+ Panels, the allocation (imputation) flag for a particular imputed item will not indicate by itself the imputation status of the item. For Type Z imputations, the EPPINTVW field in the 1996+ Panels and the person-level INTVW field in the pre-1996 panels will indicate whether the Type Z procedure was used to impute all items for the sample person (in these cases, EPPINTVW = 3 or 4 or INTVW = 3 or 4).[8,9] The individual imputation flag for each item indicates whether or not that item was imputed during the processing of the donor's fields.

For EPPFLAG imputations, the EPPFLAG field will equal 1. When this is true, all labor force participation and job/business characteristics fields are imputed via the EPPFLAG procedure, whether or not the individual items indicate an imputation. As with the Type Z procedure, an allocation (imputation) flag with a value greater than zero for any of the labor force participation items means that the values of these items are not the original values from the donor but are processed values that are consistent with the sample person's demographics and household composition; for the job/business characteristics fields, an allocation flag with a value = 4 indicates that the sample person's values in these fields have been projected forward from the person's values for these fields in the previous wave.

To find little Type Z imputations, check the allocation (imputation) flag of the variable EPDJBTHN. If (a) EPDJBTHN = 1 (indicating that the person was a worker), (b) this item's allocation (imputation) flag is 1 or 4, and (c) EPPFLAG is not 1, then a little Type Z imputation has taken place for all of the labor force participation and job/business characteristics fields. As with the Type Z procedures, the allocation (imputation) flag for an individual item only indicates whether the item was imputed when the donor 's fields were processed.

The full panel files carry only a subset of the allocation (imputation) flags carried on the core wave files. The value of an allocation (imputation) flag is set during wave processing, and, usually, it is not modified to reflect any changes in value resulting from the longitudinal editing discussed below. The Census Bureau does reset the values of some allocation flags to indicate that a longitudinal imputation has occurred.

[8] The codes for EPPINTVW and INTVW differ. In the 1996+ Panels, EPPINTVW is coded as follows: 1= Interview (self), 2 = Interview (proxy), 3 = Noninterview Type Z, 4 = Noninterview pseudo Type Z (left sample during the reference period), and 5 = Children under 15 during the reference period. In the pre-1996 panels, INTVW for person is coded as follows: 0 = Not applicable (children under 15), 1= Interview (self), 2 = Interview (proxy), 3 = Noninterview Type Z refusal, and 4 = Noninterview Type Z other.

[9] Note that for the 1990-1993 Panels, INTVW can equal 5 on the core wave files (this value is not documented in the codebook). A value of 5 denotes persons in the sample early in the wave who were not in the sample at the time of interview. Such persons are processed as if they are a Type Z nonrespondent. Prior to the 1990 Panel, such persons are identified as those with PP-MIS5 $\neq$ 1 but PP-MISj = 1 for j = 1, 2, 3, or 4.

Topical Module Imputation Procedures

When item-missing data in topical modules are imputed, the same sequential hot-deck procedure used to impute item-missing data in the SIPP core is used. Topical module data for Type Z noninterviews are also imputed item by item with the sequential hot deck. Those cases are *not* subjected to the Type Z imputation procedure that was used for core items in the pre-1996 panels.

Phase 2: Data Editing Procedures for Full Panel Files – Pre - 1996 Panels only

At the conclusion of each SIPP panel, core data from all waves are assembled into the full panel file. That assembly is done after all waves have been processed separately, producing the core wave files. Once all waves are linked, longitudinal edits are applied to the SIPP full panel files to ensure that the data for each respondent are consistent over time. Although the core wave files are edited for consistency, some types of inconsistencies become apparent only when looking at the data over multiple waves. Starting with the 1996 Panel, some longitudinal editing has been built into the CAI instrument. The ability to carry data across waves in the CAI environment is expected to result in better cross-wave consistency in the core wave files and in less need for subsequent longitudinal editing.[10]

Pre–1996 Full Panel Files

The following discussion refers only to pre-1996 procedures. Longitudinal edits in the pre-1996 panels were applied for selected variables. The edits were designed (1) to correct crosswave inconsistencies, which become apparent only when multiple waves are examined together, and (2) to honor the preference to replace imputed values from one wave with reported values from another wave.

Unlike the hot-deck imputation procedures used with the core wave files, the longitudinal edits in the pre-1996 files did not replace missing data for one person with reported data from another person. When a data value was modified during longitudinal editing, the replacement value was obtained from the same record either directly (by copying a reported value from a different month) or indirectly (using some form of interpolation or extrapolation from reported values in other months). Those procedures could cause modifications both in reported and imputed values. When a data value was modified during longitudinal editing, the associated imputation flag was not changed. In

[10] Prior to CAI, a *control file* was developed at Wave 1 that contained a unique identifier for each sample person, as well as that person's age, sex, and race. In subsequent waves, the control file provided a means of detecting inconsistencies in age, sex, and race across waves. As each wave of data was received, the reported age, sex, and race of the sample person were checked against the control file and corrections were made. Also prior to CAI, income recipiency was brought forward to the subsequent wave.

addition, the core wave files were not revised to reflect changes made during longitudinal editing. Thus, the data for any given respondent may differ between the core wave files and the full panel file, and estimates based on the full panel file may differ from those based on the core wave files.

The longitudinal edits in the pre-1996 files were performed independently on four groups of variables:

1. Demographic and household composition variables;

2. Earned income variables;

3. Other Income variables, Food Stamp variables, WIC variables, and program coverage variables; and

4. Medical insurance variables.

In most cases, the values reported during Wave 1 were used as the standard against which inconsistencies were judged. Pennell (1993) provides detailed information about longitudinal consistency edits for specific variables.

Missing Wave Imputation

There are many instances in which data are missing for a person in one wave but are present for that same person in the two adjacent waves. For example, a person may be missing in Wave 5 but have complete data for Waves 4 and 6. Beginning with the 1991 Panel, the Census Bureau began imputing those missing waves in the full panel files. Missing wave imputation is performed only when a missing wave is bounded on both sides by waves in which the sample member was present. If a respondent has missing data for more than one consecutive wave, the imputation is not performed.

For missing waves that are bounded on each side by interviewed waves, data are interpolated using a *random carryover* procedure. A value *r* is randomly assigned to each nonrespondent's household for each missing wave, where $r = 0, 1, 2, 3,$ or 4. The first *r* reference months within the missing wave receive their imputed values from the fourth month of the preceding wave, and the remaining 4-*r* reference months receive their imputed amounts from the first month of the subsequent wave.

Although this procedure results in data conducive to many analytic purposes, the random carryover forces stability in responses for wave nonrespondents. That stability could result in underestimation of between-wave changes. The procedure also results in imputed waves that do *not* exhibit the seam effect common to waves of reported data (Chapter 6). Williams and Bailey (1996) provide a complete account of the handling of missing wave data in SIPP.

Phase 2: 1996 & 2001 Panels

The 1996 & 2001 panels use waves 1-4 to inform the values of selected demographic and household composition variables for the entire panel. Waves 1-4 are linked longitudinally and made consistent across the four waves. Where a disagreement exists, data from a later wave takes precedence over data from an earlier. Once these data are edited, the cross-sectional edits are re-run for each wave 1-4 with the original demographics replaced with the longitudinally edited values. For waves 5 plus, each wave is only run cross-sectionally with the reported demographics replaced with the longitudinal values from waves 1-4.

Phase 2: 2004+ Panels

There is no phase 2 for the 2004+ panels. Each wave is edited cross-sectionally with no attempt to make the data consistent across waves - except for the normal phase 1 editing procedures which can use previous wave data to supplement missing data in the current wave.

Mapping: 2004+ Panels

The SIPP data collection instrument was extensively modified for the 2004 panel. The intentions of the changes were to decrease the respondent burden and to increase the accuracy of the data collected, while keeping the scope and content of the survey essentially unchanged. Due to the complex nature of the edit programs it was determined that they would not be re-written. Instead, a mapping operation was inserted into the beginning of the processing stream which (wherever possible) translated the data from the format of the new instrument to the format of the 2001 instrument. This allowed us to run the remaining processing steps with limited changes and to release public use files for 2004+ in the same format and with the same variable names as the 1996 & 2001 panels. Analysis of the affect of these changes on data quality and/or respondent burden is beyond the scope of this guide.

Confidentiality Procedures for the Public Use Files

All of the editing and imputation procedures described in the preceding sections are part of the process of preparing the data for internal Census Bureau use. Before the files are released for public use, they undergo additional editing to protect the confidentiality of respondents. Two procedures are used: topcoding of selected variables (income, assets, and age) and suppression of geographic information. As a result of these procedures, estimates based on data from the public use files will differ slightly from the Census Bureau 's published estimates.

Topcoding

One piece of information that might reveal a respondent's identity is a very high income. For that reason, the Census Bureau *topcodes* income before making that information publicly available, recoding any income amounts over a certain maximum value to that maximum. In other words, income on the public use data files has a ceiling value. Although income is the primary variable that is topcoded, other variables that may disclose a respondent's identity, such as age, are also topcoded. A few variables, such as starting dates for employment, may be *bottom coded* if they pose a disclosure risk. Chapter 10 and Appendix B provide a thorough discussion of top coding methods and procedures in SIPP.

Suppression of Geographic Information

Geographic information that can be used to directly identify survey respondents, such as an address, is removed from the public use files. In addition, states and metropolitan areas with populations less than 250,000 are not identified. Specific nonmetropolitan areas (such as counties outside of metropolitan areas) are never identified. In certain states, when the nonmetropolitan population is small enough to present a disclosure risk, a fraction of that state 's metropolitan sample is recoded to nonmetropolitan status. For that reason, the SIPP data cannot be used to estimate characteristics of the population residing outside metropolitan areas. Chapter 10 provides details.

For the 1996 & 2001 Panels, state-level geography is shown for 45 states and the District of Columbia. The remaining five states are combined as follows:

1. Maine, Vermont; and

2. North Dakota, South Dakota, Wyoming.

For the 1984 through 1993 Panels, state-level geography is shown for 41 individual states and the District of Columbia; the nine other states are combined into three groups:

1. Maine, Vermont;

2. Iowa, North Dakota, South Dakota; and

3. Alaska, Idaho, Montana, Wyoming.

All States are identified for the 2004+ Panels.