Census Editing and the Art of Motorcycle Maintenance[1]

Michael J. Levin[2]

Harvard Center for Population and Development Studies
9 Bow St., Cambridge, MA 02138, michael.levin@yahoo.com

Introduction

Well-designed censuses and surveys are invaluable resources for a nation. But no census or survey data are perfect. Since data come in from the field with various invalid and inconsistent responses, some process must free them as much as possible of errors and inconsistencies to obtain accurate census or survey results. Countries have long recognized that data from censuses and surveys have these problems and have adopted approaches for dealing with data gaps and inconsistent responses. However, because of the long interval between censuses, offices often do not properly document the procedures used to edit the census data, so some countries have to reinvent the process used in earlier data collection activities for a new census or survey. Contemporary population and housing census editing is the procedure for detecting the errors in and between data records. These procedures occur during and after data collection and capture, and sometimes require adjusting individual items and groups of items to provide quality data for tabulation and dissemination.

During the late 1990s, I wrote the *Handbook on Population and Housing Census Editing* (2002) for the United Nations that many countries used during the first decade of the 21st century. Appendix A shows many of these countries. The handbook bridges the gap in census and survey knowledge about editing methodology. It also provides information for officials and staff in the use of various approaches to census editing. It also encourages countries to retain a history of their editing experiences, enhance communication between subject matter and data processing specialists, and document the activities carried out during the current census or survey in order to avoid duplication of effort in the future. The *Handbook* was to be a reference for both subject-matter specialists (demographers, social scientists, economists and others) and data processing specialists as they worked as teams to develop better communication for editing specifications and programs for censuses and surveys. It followed a "cookbook" approach, permitting countries to adopt the edits most appropriate for their own country's particular statistical situation. The present paper updates some aspects of the handbook, particularly in relation to new approaches needed for the increasing use of scanning rather than keying.

The Census Process

A population and housing census is the total process of collecting, compiling, evaluating, analyzing, and releasing demographic and/or housing, economic and social data pertaining to all persons and their living quarters (United Nations, 2007). Traditionally, censuses are conducted at specified times in an entire country or a well-delimited part of it. Recently, some countries have started carrying out continuing surveys to cover the whole country, using a "long" form, to provide complete coverage over time. In either scenario, the census provides a snapshot of the population and housing at a given point in time.

All censuses and surveys share certain major features that include (a) preparatory work; (b) enumeration; (c) data processing, including data entry (keying or scanning), editing and tabulating, (d) databases construction and dissemination of results; (e) evaluation of the results; and (f) analysis of the results. Although aspects of the census process are important, providing both feeding and feedback, we concentrate on only one of the aspects – the computer editing – in this paper.

---

[1] Why motorcycle maintenance? A census is like a precision instrument with inter-locking parts, all of which must work both independently and together. Without precision, the census because a lesser product. We always hope that no census becomes a clunker.

[2] The current paper abstracts from the UN Editing Handbook. I did much of the early work reported here while at the U.S. Census Bureau in cooperation with the United Nations Statistical Division. Neither of these organizations nor the Harvard Center for Population and Development Studies is responsible for any errors.

We classify the sources of error that census data suffer from, generally, as (1) coverage or structure errors and (2) content errors. We implement **Structure edits** to check and correct number of person records, sequencing, and the existence of duplicate persons. Coverage errors arise from omissions or duplications of persons or housing units in the census enumeration. The sources of coverage error include incomplete or inaccurate maps or lists of enumeration areas, failure by enumerators to canvass all the units in their assignment areas, duplicate counting, omission of persons who are not willing to be enumerated, erroneous treatment of certain categories of persons such as visitors or non-resident aliens and loss or destruction of census records after enumeration. Coverage errors should be resolved in the field. The office editing process eliminates actual duplicate records. However, the programs must determine whether these are duplicate persons or households. Twins, for example, may have identical information, except for sequence number. Hence, the editing rules applied during this process determine when to accept and when to reject seemingly duplicate information, and when to make changes through imputation.

Content Edits must correct errors arising from the incorrect reporting or recording of the characteristics of persons, households and housing units. Content errors may be caused by poorly designed questions or poor sequencing of the questions, or by poor communication between respondent and enumerator, as well as by mistakes in coding and data entry, errors in manual and computer editing, and erroneous tabulations of results. Edit trails (also known as audit trails) must be properly developed and stored at each stage of the process to ensure no loss of data. The following sections explain each of the above errors.

It would be foolish to report that errors only appear during computer editing. Errors occur at all stages of the census process, and these include: (1) Errors in questionnaire design, (2) Enumerator errors, (3) Respondent errors, (4) Coding errors, (5) Data entry errors (6) Errors in computer editing, (7) Errors in tabulation. The UN Editing Handbook discusses the contribution of each of these error types in detail, so we will not repeat them here. But, it is evident that the census process involves a number of sequential, interrelated operations, and errors may occur in any of them. Computer edits are part of a feedback system, with computer edits not only feeding forward to tabulations, but also linking backward to collection and field processing. The best way a national statistical office can prevent problems with the computer edit is to maximize the fieldwork and editing there. The national statistical office also needs to make sure that coding and data entry are accurate, and should have continuing feedback among all operations, including entry, editing, tabulations, and dissemination.

Editing in historical perspective

Before the advent of computers, most census operations hired large numbers of semi-skilled clerks to edit individual forms. However, because of the complexity of the relationships between even small number of items, simple checks could not begin to cover all of the likely inconsistencies in the data. Different clerks would interpret the rules in different ways, and even the same clerk could be inconsistent.

Census editing changed with the introduction of computers. Computers detected many more inconsistencies than manual editing. Editing specifications became increasingly sophisticated and complicated. Automated imputation became possible, with concomitant rules for the process (Nordbotten, 1963; Naus, 1975). At the same time, the process allowed for more and more contact with respondents, or at least with the completed questionnaires of these respondents. Many editing teams began to feel that "the more editing the better", and the more thorough the edit, the more accurate the results. Programs produced thousands of error messages, requiring manual examination of the original forms or, for some surveys, re-interviews of the respondents. Computers made it increasingly easy to make changes in the data set. Sometimes these changes corrected records or items, but many records passed through the computer multiple times, with errors and inconsistencies reviewed by different persons each time (Boucher, 1991; Granquist, 1997).

Several generalized census-editing packages came out of this whole process, with some of them still in use today[3]. Researchers initially developed the packages for mainframe computers: others modified them later for use on personal computers. During this period, Fellegi and Holt (1976) developed a new method for generalized editing and imputation,

[3] While SAS, SPSS, STATA, and other statistical packages can do edits, only specialized packages, like the DOS-based IMPS and ISSA, and the current Windows-based CSPro provide the types of error listings shown here. Also, CSPro is free software; unfortunately, it is not clear whether continued development and ready access for technical assistance will be available over the longer term.

which most countries did not put into practice, but which increasingly numbers of countries adopt today as national statistical offices become more sophisticated in their editing.

A major advance in census editing came in the 1980s when national statistical offices began to use personal computers to enter, edit, and tabulate their data. Suddenly, data processors could perform edits on-line at the data entry stage or soon after. For surveys and small country censuses, staff could develop programs to catch errors during collection or while entering data directly into the machine. Computer edits allowed more, continuous contact with respondents to resolve problems encountered in the editing process (Pierzchala, 1995). And, recently personal digital assistants (PDAs) help collect data in censuses and surveys, and edit immediately, even while the enumerator and the respondent are still together; in this case, the data should need little further editing.

In the early years, the process of making increasingly sophisticated and thorough checks on census and survey data seemed to be very successful. Editing teams created increasingly complicated editing specifications, and data processing specialists spent months developing flow charts or decision charts and program code. Analysts seldom evaluated the packages. It seemed that editing could correct any problems arising from earlier phases of data collection, coding, and keying. Nevertheless, it also became apparent to some analysts that in many cases, all of this extra editing harmed the data, or at the least, delayed the results or caused bias in the results. Sometimes the programs made so many passes through the data, correcting first one item, and then another item, that statistical offices obtained far different results from the initial, unedited data. As Granquist (1997) notes, many studies have shown that for much of this extra work, "the quality improvements are marginal, none or even negative; many types of serious systematic errors cannot be identified by editing". So, an issue that each national statistical office must face is what level of computer editing is appropriate for its purpose.

The editing team

As national statistical offices prepare for a census, they need to consider a variety of potential improvements to the quality of their work. One of these is the creation of an editing team. The editing process should be the responsibility of an editing team that includes census managers, subject-matter specialists, and data processors. The statistical office should ste this team up as soon as preparations for the census begin, preferably during the drafting of the questionnaire. The editing team is important from the beginning, and remains so throughout the editing process. Care in putting together the team and in developing and implementing the editing and imputation rules assures a census that is faster and more efficient.

Meetings between census officials and the user community concerning tabulations and other data products can provide insight into the needed edits. Developing the editing rules and the computer programs during a pretest or dress rehearsal makes it possible to test the programs for the tables as well and leads to faster turn-around times for various parts of the editing and imputation process. The editing team then ascertains the impact of these various processes and takes remedial action if necessary. As the subject matter and data-processing specialists work together on the editing and imputation rules, they will elaborate a nearly final edit strategy early in the census preparations.

The census editing team creates written sets of consistency rules and corrections. Communication is crucial at all stages of the editing process. In addition to developing the editing and imputation rules, the subject-matter and data processing specialists must work together at all stages of the census or survey, including during the analysis. The risk of doing too much editing is as great as the risk of doing too little editing and having unedited or spurious information in the dataset. Hence, both groups must take responsibility to maintain their metadata-bases properly. The editing team must also use available administrative sources and survey registers efficiently in order to improve the current and subsequent census or survey operations.

Editing Practices: Edited versus unedited data

No census editing can improve the quality of the enumeration. However, countries perform census edits to make the data and presentation more aesthetic. Data not completely edited can cause problems in interpretation for a particular census, and since microcomputers now permit trends analysis over time, between censuses (and censuses and surveys).

National statistical offices often face the dilemma of trying to serve multiple users. Some users (often demographers)

may want unknown entries included for analysis or research and others may want data with minimum noise for their planning or policy purposes. If the national statistical office disseminates an unedited table, such as that on the left side of table 1, both the analysts and the policy makers will have to make assumptions when using the results. Table 1 illustrates this point with only a small number of persons. It shows that 23 persons in this country reported no sex and 15 reported no age. Of these, two cases reported neither sex nor age. These omissions may be non-responses or keying errors.

TABLE 1. SAMPLE POPULATION BY 15-YEAR AGE GROUP AND SEX, USING UNEDITED AND EDITED DATA

	Unedited data				*Edited data*		
Age group	*Total*	*Male*	*Female*	*Not reported*	*Total*	*Male*	*Female*
Total	4,147	2,033	2,091	23	4,147	2,045	2,102
Less than 15 years	1,639	799	825	15	1,743	855	888
15 to 29 years	1,256	612	643	1	1,217	603	614
30 to 44 years	727	356	369	2	695	338	357
45 to 59 years	360	194	166	0	341	182	159
60 to 74 years	116	54	59	3	114	53	61
75 years and over	34	12	22	0	37	14	23
Not reported	15	6	7	2			

Most users would make their own decisions about what to do with the unknowns. A logical, possibly naïve, approach would be to distribute the unknowns in the same proportion as the known values. If the national statistical office chooses to impute the unknowns in the table, the editing team may decide to have 12 males and 11 females, figures that are about half-and-half. The results will then be consistent with the edited data shown on the right side of table 1. Another possibility, during analysis, would be to simply drop the unknowns and using percentage distributions.

When a country does not do computer editing using hot decks, it still imputes. But, it imputes at the end, when staff have no additional information available. They do not know whether the individuals had fertility reported; they do not know whether one spouse reported sex but the other didn't; they do know whether the person of unknown sex was the mother of someone in the household; and, they don't know the local conditions, if this was part of a nursing quarters or construction barracks. When statistical organizations impute during computer edit, they have information about the person, the household, the local community, and the larger geography that they lose later on, after they compile the data. The end – the bottom line – is not the best time to make these decisions.

An alternative imputation strategy would be to take one or more of these other variables into account. Or, the national statistical office could choose would be to base the imputation on the age distribution. For the sample population illustrated in table 1, 15 cases occurred with unreported age. The program could distribute these data in the same proportions as the known values, again, a logical strategy for imputation. Still, the editing team could probably obtain better results by considering other variables and combinations, such as the relative age of husband and wife, of parent and child or grandparent and grandchild, or the presence of school age children, retirees, and persons in the labor force.

In table 1, the edited data on the right are "cleaner" because the imputation suppressed the unknowns (see columns under "edited data"). This side of the table has no unknowns, since the program allocates them to other responses. Nevertheless, many demographers and other subject-matter specialists have traditionally wanted to have the unknowns shown in the tables, as in the unedited data of table 1. They believe that this procedure allows them to perform various kinds of evaluations on the figures to measure the effectiveness of census procedures or to assist in planning for future censuses and surveys. Both objectives can be accomplished—an edited table for substantive users and an unedited one for evaluation—by making tabulations both with and without unknowns. This procedure does not particularly help staff in the statistical office who must respond to user requests, with users wanting a single number when the call or visit the office, or look online for a particular statistic.

Statistical offices, then, must make every effort to maintain the original, collected data. The offices should archive a

complete set of the original, keyed data, both as part of the historical record, but also for reference if staffs make decisions about re-editing any part of the data set from the beginning. But, the programmers should keep original values of crucial items, like age, sex, and fertility, at the end on each record to allow demographers and others to analyze the results of the edits. As noted, unknowns all affect trend analyses, and so office need to keep original data for later analysis

The Basics of Editing

What is editing?
Editing is the systematic inspection of invalid and inconsistent responses, and subsequent manual or automatic correction (using "unknowns" or dynamic imputation) according to predetermined rules. Some editing operations involve manual corrections, which are hand-made corrections in the office. Other editing operations involve electronic corrections, using computers. Census publications are likely to contain a certain amount of meaningless data if national statistical offices do not edit the census or survey data. Editing reduces distorted estimates, facilitates processing, and increases user confidence. Further, according to Pullum, Harpham, and Ozsever, (1986) "The primary achievements of editing or cleaning are, first, to detect whether the various responses are consistent with one another and with the basic format of the survey instrument."

The raw data files in a census contain errors of many kinds. Data processing categorizes the errors into two types: those that may block further processing and those that produce invalid or inconsistent results without interrupting the logical flow of subsequent processing operations. As noted in *Principles and Recommendations for Population and Housing Censuses, Revision 1* (UN, **2007,para.195**), all errors of the first kind must be corrected and as many as possible of the second. The basic purpose of census editing at the processing stage, therefore, is to identify errors and make changes to the data set so that items are valid and consistent. Nevertheless, processing cannot correct all census errors, including questionnaire responses that are internally consistent but are in fact instances of misreporting on the part of respondents or mis-recording on the part of enumerators.

More and more evidence exists that no amount of computer editing can take the place of high-quality data collection. National statistical offices know that at some point computer editing is not only limiting, but also becomes counter-productive: the edit adds more errors to the data set than it corrects. Changing a census item is not the same as correcting it. Hence, the editing team must work together to determine the beginning, the middle, and the end of the editing process.

The main problem is determining how far to go to obtain a good quality dataset. As noted earlier, the advent of computers, first mainframe computers and then microcomputers, has allowed for virtually complete automation of the editing process. In many national statistical offices subject-matter specialists have in fact become editing enthusiasts. Hence, offices now perform many consistency tests that were difficult or impossible in the past, particularly those involving inter-record checking and inter-household checks. Unfortunately, this feature of microcomputers has also led to many problems, and the greatest of these is over-editing.

How over-editing is harmful
Over-editing reduces timeliness, while increasing cost, distorting true values, and giving a false sense of security regarding data quality:

Timeliness. The more editing a national statistical office does, the longer the total process will take. The major issue is to determine how much the added time adds to the value of the census product. Each editing team must evaluate, both on going and after the fact, the net benefits of the added time and resources for the overall census product. Often, the returns are so small in terms of the time invested that it is better to have small "glitches" in the data than deprive prime users of receiving the information on a timely basis.

Finances. The costs of the census process increase as time increases. Each national statistical office has to decide, as it increases the amount and complexity of its edits, whether the increases in costs are worth the added effort and whether it can afford these additional costs.

Distortion of true values. Although the intention of the editing process is to have a positive impact on the quality of the data, increases in the number and complexity of the edits may also have a negative impact. Sometimes, editing teams change items erroneously for a variety of reasons: mis-communication between subject-matter and data processing

specialists; mistakes in a very complicated, sophisticated program; or handling a census item many times in an edit. National statistical offices want to avoid this type of problem whenever necessary. Granquist and Kovar (1997) point out, for example, that imputing the age of a husband and wife using a set age difference between them can be useful, but may artificially skew the data when many such cases exist.

A false sense of security. Over-editing gives national statistical office staff and other users a false sense of security, especially when offices do not implement and document quality assurance measures. Furthermore, odd results will appear in census tabulations no matter how much editing the team does, so it is important to warn users that small errors may occur. This fact is especially true now that many countries release sample microdata. National statistical offices would not want to release data detrimental to the planning process, so staff must take great care to assure that they edit all crucial variables properly for later planning. For example, no national statistical office would want to release microdata or tabulations with unknowns for sex or age. On the other hand, variables such as disability or literacy work well with less editing. While some inconsistencies in the cross-tabulations may appear because national statistical offices cannot edit all pairs of variables, editing teams should check the most important combinations. When editing teams find inconsistencies, correction procedures should be available.

Treatment of unknowns

The editing team must decide early in census planning how to handle "not stated" or unknown cases. Columns or rows of unknowns in tables are neither informative, nor useful, so planners in most countries prefer to have these data imputed. Without treatment of unknowns, many users distribute the unknowns in the resulting tables in the same proportions as the known data, thus imputing the unknowns after the fact. So, the editing team needs to decide how to deal with the unknowns systematically.

Spurious changes

National statistical offices do not usually work with models when they develop their editing rules. Editing teams should develop rules that fit the actual population or housing characteristics. All data should pass the edit rules. For example, a set of rules may require that the child of a head of household should be at least 15 years younger than the head. However, a child of the head may actually be a social, rather than biological child: He or she might be the biological child of the spouse, but not the head. Hence, the difference in age might be less than 15 years. Since planners in most countries do not plan separately for children and stepchildren, if, under the above circumstances, the editing rules change the age of the child, inconsistencies in educational attainment, work force participation and other areas may develop. Therefore, the edit team should test this rule to see the results before full implementation. [Note here that the relationship should be changed, not the age since relationship is mainly collected to make sure everyone is counted, and offices don't usually plan selectively based on this item.]

Determining tolerances

The editing team must develop "tolerance levels" for each item, and sometimes for combinations of items. Tolerance levels indicate the number of invalid and inconsistent responses allowed before editing teams take remedial action. For most items in a census, for example, some small percentage of the respondents will not give "acceptable" responses, for whatever reason. For some items, like age and sex, used in combination with so many other items for planning, the tolerance level should be quite low. When the percentage of missing or inconsistent responses is low (less than one or 2 percent), any reasonable editing rules are not likely to affect the use of the data. When the percentage is high (5 to 10 per cent, or more, depending on the situation), simple, or even complex, imputation may distort the census results. In this case, the item has failed, and the edit can do little (or nothing) to save it. Proper pretesting usually helps. Some items, like disability or social programs, are more properly survey than census items, but have to be included to get full coverage.

To reduce missing responses to a minimum, national statistical offices should ensure that census workers make every effort to obtain the information in the field. If a given country decides that it does not need as much accuracy for some items, such as literacy or disability, the tolerance level for those items might be much higher. Sometimes editing teams can correct items that have too many errors, by returning enumerators to the field, by conducting telephone re-interviews, or by applying their knowledge of an area. Often, though, it is too costly to return to the field or do other follow-up operations, and the national statistical office may decide either not to use the item or to use it only with cautionary notes attached.

Learning from the editing process
As the team edits the data, they need to record detailed analyses of positive and negative feedback to improve the quality of the both current census or survey and future censuses and surveys. The editing team has to work constantly to determine what is working properly and what is not working. They must also determine whether those aspects of the process that are working properly can be improved and streamlined, so that the data can get to users even sooner. The earlier in the census process national statistical offices detect errors, the more likely they will be to correct them.

Quality assurance
Quality assurance is important in all census operations. Consequently, formal quality assurance mechanisms should certainly be in place to monitor the progress of the computer editing and imputation phase. Audit trails, performance measures, and diagnostic statistics are crucial for analysis of the quality of the edits and the rapidity of processing (Granquist and Kovar 1997; Statistics Canada, 1998).

Costs of editing
Editing activities take a disproportionate amount of time and funding, so each country must determine the return on its investment. Excessive editing can delay census results. While national census/survey staff may have only anecdotal evidence for such experience with censuses, a study by Pullum, Harpham, and Ozsever (1986) found that machine editing of the World Fertility Survey contributed to a delay in the publication of the results by about one year. National statistical offices might better spend their funding on obtaining a higher quality census or survey enumeration in the first place.

Imputation
Imputation is the process of resolving problems concerning missing, invalid, or inconsistent responses identified during editing. Imputation works by changing one or more of the responses or missing values in a record or several records to ensure that plausible, internally coherent records result. Contact with the respondent or manual study of the questionnaire eliminates some problems earlier in the process. However, it is generally impossible to resolve all problems at these early stages owing to concerns with response burden, cost, and timeliness. Imputation then handles the remaining edit failures, since it is desirable to produce a complete and consistent file containing imputed data. The members of the team with full access to the microdata and in possession of good auxiliary information do the best imputation.

- The imputed record should closely resemble the failed edit record. Imputing a minimum number of variables usually works best, and thereby preserves as much respondent data as possible. The underlying assumption (which is not always true in practice) is that a respondent is more likely to make only one or two errors rather than several;
- The imputed record should satisfy all edits;
- Editing teams should flag imputed values. They should clearly identify the methods and sources of imputation.
- The editing team should retain the un-imputed and imputed values of the record's fields to evaluate the degree and effects of imputation.

Archiving
Part of the quality assurance process of the census or survey is to document all processes and then to archive that documentation. National statistical offices need to preserve both the edited and unedited data files for later analysis. Some procedures, such as many forms of scanning, automatically keep the original image. Similarly, immediately after keying batches, the data should be concatenated and preserved for potential analysis. But, with either procedure, it is important to archive original copies of the non-edited files. In fact, statistical offices should keep copies of the unedited data in several places within the Statistics Office, as well in other parts of the country, and outside the country as well. The documentation should be complete enough for census or survey planners to be able to reconstruct the same processes later to assure compatibility with the census or survey under consideration. The processes and the results must be replicable. Finally, the unedited data as well as the edited data must be stored in several places, with appropriate measures to ensure their continued availability over time.

As noted below, part of the documentation involves the three types of edit reports. The first report provides the summary statistics giving numbers and percentages of errors (based on appropriate denominators, like total housing units, total population, working age population, adult females, etc.). The second report contains at least a sample of the "case" structure, with the unedited household or housing record, the listing of errors and their resolutions for the housing unit or individuals in the unit, and the edited housing unit or household. The third report provides frequencies of values for items

edited. The programmers should provide three sets of errors at logical geographic levels, certainly for the major civil divisions, but providing error listings at lower levels of geographic levels could assist in targeting problems in enumerator training, quality control, or other issues connected with the enumeration.

Editing Applications

Whether a census data set is scanned or keyed, a certain general flow pertains. The census edit team starts with the unedited data. Usually, enumerators or office staffs have pre-coded all data, so the data set is ready for the structure edit. Sometimes an operation converts the scanned data into another machine-readable form for the editing process, depending on the editing package used. Also, sometimes the scanned data require a second automated coding operation to fill in items like birthplace, industry, and occupation.

In either case, the unedited data should appear in a form allowing the computer programmers to develop the **structure edits**. The structure edit checks to make sure that all of the major civil divisions are presented in geographic or numerical order, and within each major civil division, each minor civil division occurs, and in geographic or numerical order. Then, within each minor civil division, each locality must appear, and within geographic or numerical order. This procedure continues down to the lowest geographic level. Appropriate procedures make sure that each housing unit appears once and only once in the data set.

Structure edit

The structure edit must also make sure that all record types are present when expected, and that the data set has no record types repeated when they should not be. So, for a population and housing census, either the population or housing records must come first, and then the whole data set must follow that convention. In most cases, only one housing record should be present, so programs must deal with surplus records, and programmers must supply housing records to households without them. Similarly, population records must be present for occupied housing units (usually defined as such on the housing record) and must be absent for vacant units.

After the structure is set, it is not really set. Inevitably, the structure edit will be re-visited during the content edit, and often beyond, as glitches appear during the various census processes; this revisiting is normal in census work and is expected. Time, personnel, and equipment requirements must take account of the need to maintain the structure.

Content edit

Then, the **content edit** begins. Content edits must consider each population and housing item alone and usually in combination to determine the validity of each item, and the best fit among the items.

When the team finishes the content editing, they should have a completely edited data set. The unedited data should be stored in several secure places, and the important unedited items (or all the unedited items) should also appear at the ends of the various types of records. Again, it is important to note that as the tables are developed, the content edits may have to re-visit the data as well to take care of any specific problems resulting from particular cross-tabulations.

The editing process works well when imputations deal with random omissions and inconsistencies. However, if systematic errors occur during data collection, editing cannot improve the quality of the data no matter how sophisticated the procedures. The choice of topics investigated is of central importance to the quality of the data obtained. When interviewed, respondents must be willing and able to provide adequate and appropriate information. Thus, censuses should avoid topics likely to arouse fear, local prejudices, or superstitions, as well as questions that are too complicated or difficult for the average respondent to answer easily. The exact phrasing for each question depends on national circumstances and must be well tested. National statistical offices must allocate sufficient resources to obtain the highest quality census data.

To implement the computer editing phase of the process the editing team prepares written editing instructions or specifications (pseudocode), decision tables, flow charts and pseudocode. Flow charts may help the subject-matter specialists to understand the various linkages among the variables and ease writing editing instructions. The subject-matter specialists write the editing instructions in collaboration with the computer specialists, describing the action for each data item. The editing instructions should be clear, concise, and unambiguous since they serve as the basis for the editing program package.

Coding considerations
Coding is the process of making machine-readable numbers and alpha-numerics**.** When developing a coding scheme, census and survey staff must consider the returns of each investment of time, energy, and funds. Coding considerations are reasonably insignificant for small countries or small surveys since the amount of processing is much less than for a census.

As national statistical offices develop lists of codes for the editing programs and for subsequent tabulations, they may wish to establish common codes for groups of items. For example, in many countries, place codes (birthplace, parental birthplace, previous residence, work place), language, ethnicity/race, and citizenship are very similar. A common coding scheme for "place" might be developed as three-digit codes with the first digit representing the continent, the second the region, and the third the specific country. National statistical offices can also use country numerical codes developed by international organizations such as the United Nations Statistics Division (United Nations, 1999). A set of common codes for closely related variables can reduce coding errors and assist the data processors during the edit. Common codes also allow data processors, where appropriate, to use an entry from one item to determine another.

The structure of coding can facilitate the coding process as well as later processing during editing, tabulation and analysis. For large countries with many immigrants or ethnic groups, codes based on continent, region and country, with different codes or digits assigned to each, would be preferable to a simple listing. If a group of items on a questionnaire is not independent of each other, national census/survey staff probably should not ask all of them. The editing team must decide, on a case-by-case basis, when to use other items directly for assignment, and when to use other available variables.

Manual versus automatic correction
Manual editing of a census may take months or years, presenting many possibilities for human error. Manual editing is a weak alternative to computer editing, partly because it is impossible to create or reconstruct an edit trail for the manual correction process. Computer, or automated, editing reduces the time required and decreases the introduction of human error. Both computer and manual editing check the validity of an entry by looking for an acceptable value, but computer programs also check the value of the entry against related entries for consistency. Finally, and most importantly, automated editing allows for the creation of an edit trail and is therefore reproducible, while manual editing is not.

In the early years of computer entry, no editing on entry was possible. That is, all correction had to be either manual as part of the coding and checking office operations, or as part of the computer operations, but after keying. Newer packages have built-in edit functions so that keyers cannot enter invalid entries, unless forced by the keyers, and the entry program can flag inconsistencies that the keyers or editors will correct manually or by computer. As scanning has become more prevalent, this sequence repeats; in the early years of scanning, no edit during entry was possible, but now programmers can build validity edits and data conversions into the scanning systems.[4]

When censuses and surveys collect large volumes of data, staff cannot always refer to the original documents to correct errors. Even if the original questionnaires are available, the data recorded on them may sometimes be wrong or inconsistent. A computer editing and imputation system corrects or changes erroneous data immediately and generates reports for all errors found and all changes made. Computer edits should be carefully planned to save staff time for other data processing activities. While running large quantities of data through a computer system can be time-consuming, it is not as time-consuming as manual correction.

Manual correction. Manual correction takes several forms. Consider a simple example of an error in the sex response: a supervisor checks an enumerator's work and finds an obvious error, such as assigning "male" to someone named "Mary". In changing the sex to "female," the supervisor performs a manual edit. If the supervisor does not correct the questionnaire, but instead sends it to the field office, the office workers there may observe the problem and manually correct it. At the central office, during coding, coders might see the mismatch between the name and the sex and make the manual correction then. Or, the coders might not observe the problem, but when the keyers are entering the data for

[4] Scanning does terrible things to data. A properly set up keying program will only allow 1 or 2 (and maybe 9) for sex, but, with scanning, almost anything can occur because of enumerator error or stray marks or lack of quality control in the scan. Hence, the editing program must account for many more possibilities.

the questionnaire, they may notice the mismatch between the name and the sex and make the manual correction before keying.

However, if no one notices the error, and the keyer enters the code for “male”, the editing programs may follow several different procedures at this point. For gender-related items such as the fertility block, the editing program might flag the fact that this is a male with fertility information and produce a message to that effect while the keyer is entering the data. The keyer could then look at the questionnaire, find that indeed this is a female and make the correction manually. Alternatively, if the national statistical office uses an editing program independent of the keying, the computer program might flag this person as a male with fertility information. Then, by using the geographical information, office workers can find the original questionnaire in the bins, pull it, and determine that the respondent, named “Mary”, was erroneously reported as "male" instead. At this point, the office staff can take this information back to the keyer, who can pull up the record and make the manual correction.

This example shows both the advantages and disadvantages of manual editing. At any of the steps outlined above, a census worker could note the error—the mismatch between the name and the sex—and make the correction. However, national statistical offices that use manual editing probably have staff checking for this relationship at every stage. An enormous amount of energy is expended in this activity, and the results are probably little different, particularly in the aggregate, than if the staff were instructed to do no manual editing.

Originally, the only way to make corrections in a dataset was to make this change manually. Some countries still do not feel comfortable using automatic correction, so they use manual correction at one of the stages described above. If the dataset is small, timing is not crucial or the work force is labor-intensive, then manual correction will work in most cases. The advantage is that if the information is both complete and accurate on the questionnaire, and the inconsistency can actually be resolved by looking at the form, the quality of the census or survey will probably improve marginally. The editing team has to assume, for example, that “Mary” is not “Gary”, and that if fertility appears, it was actually to be collected for this person – that the enumerator did not collect it erroneously. In fact, editing and imputation procedures rarely improve the quality of the data collection. They only change certain elements.

Sometimes, looking up a questionnaire for manual correction is fruitless. The information is not there, for whatever reason. Sometimes a person does not want to provide his or her age, so the item is blank on the questionnaire. In this case, examining the questionnaire will not resolve the issue. Then, the editing team must make a decision about how to handle the situation. For manual correction, the national statistical office must either assign “unknown” or use some set of values to assign the age item.

Automatic correction. Manual correction inevitably lowers quality and consistency unless the enumerator re-contacts the respondent. It takes more time, and it costs more. Computers do not tire and are faster; they do not have personal problems that might interfere with maintaining quality or consistency; and, in most cases, they make processing cheaper. Most countries now use some kind of automatic correction.

Missing and inconsistent responses reduce the quality of data and make it difficult to present easily understood census tables. Some users prefer to tabulate missing and inconsistent responses as a “not reported” category, while others prefer to distribute these cases proportionately among the reported consistent entries. Still others recommend rules for imputing *likely* answers for missing or inconsistent responses. The use of computers makes it feasible and efficient to impute responses based on other information in the questionnaire or on reported information for a person or housing unit with similar characteristics.

Since the computer can look at many characteristics at the same time, the editing process should take advantage of this feature. Thus, editing procedures involving many related characteristics may result in imputing more reasonable responses than a simple edit could produce; a poorly designed editing may lead to the production of poor census data. The editing team should be composed of experienced subject-matter specialists from different, relevant disciplines as well as data processors. The members of the editing team should carefully select the variables to examine in the tests for consistency in order to determine the editing and imputation specifications. The program outputs should include the percentage of responses that were changed or imputed. Analysts will then be in a better position to judge the quality of the data; a high percentage of imputations would be a warning to use the data with caution.

Finally, an edit, or audit, trail shows the changes made to each variable. The trail traces the history of the responses from the receipt of the data through the editing and imputation process.

Guidelines for correcting data

Whether performed manually or automatically, editing should make the data as nearly representative of the real-life situation as possible by eliminating omissions and invalid entries and by changing inconsistent entries. Consider a household with consistent relationships and sex entries: the head of household is male and has no fertility information; the spouse is female and has appropriate fertility information. In many instances, however, information is inconsistent. The following questions then arise: what should the editing process be for a household with inconsistent entries? For example, how should the editing team perform the edit, if both the head of household and spouse report as male? In the past, the typical editing rule would have assumed that the first person in a couple is male, particularly if that person is the head of household, and that the second person, or the spouse, is female.

But, if the head of household in this case happens to be the wife rather than the husband, then an editing rule assuming males always come first would be wrong and the national statistical office would end up with four errors:

(a) The head of household's sex would be wrong;
(b) The spouse's sex would be wrong;
(c) The head of household would lose her fertility information;
(d) And, the computer would erroneously assign fertility to the male spouse.

Clearly, this is not good editing procedure. In contrast, when a good edit finds that the head and spouse have the same sex, it then checks both persons for fertility. Since only the head has fertility, the head becomes the female. The editing rules for these items are then satisfied.

Top-down editing approach.

Top-down editing starts with the first item to be edited (the "top"), usually the first variable on the questionnaire, and then moves through the items in sequence, until completing the edit of all items. The usual approach is to first take into consideration the response rates and the relative importance of the various items. Because of their importance, particularly in dynamic imputation (hot deck), the edits usually start with sex and age. While the top-down approach does not completely preserve the relationships among the data items, it does provide an adequate framework to complete the edit.

Some edits change the value for an item more than once during the editing process. This type of edit is very dangerous, and can introduce errors into the dataset. An imputed value may be inconsistent with other data. Even when variables are dealt with sequentially, a particular variable should be edited against all other variables concurrently, if possible. For example, a child's age, imputed on the basis of the mother's age, may be inconsistent with the child's reported years of school or years lived in the district. In this instance, the age might be re-imputed until it is consistent. An imputed age is an intermediate variable until final assignment. In creating the edits, imputed intermediate variables should not be recorded as changes until the final assignment.

Although the editing program might accept a blank or "not reported" entry for a few items and conditions, related information can supply entries for most items left blank or having erroneous entries. Entries supplied in this manner may or may not be correct on an individual basis. However, the extensive capabilities and speed of the computer for comparing different stored values permit the determination of replacement entries that reasonably describe the situation. The resulting tabulations in most cases will be sometimes more consistent than those from unedited records or records in which imputation converts all unacceptable entries into "not reported".

Editing teams must avoid circular editing—making changes to an item or several items, and then, at some later point, changing them again or back to the way they were. Staff must make several runs to make sure they completely edit all items. They might create editing criteria that change the data during a first run, but that, when applied to the changed data during a second run, change it back to the original configuration. This procedure can continue through multiple runs. The editing team should avoid introducing such criteria into the editing process.

The editing program must perform structural checks, content edits for population items, for housing items, and, in most cases, create one or several recoded variables on population and housing records required for various tabulations.

Multiple-variable editing approach

The "top-down" approach may not always give the best results—those that come closest to the real distribution of the variables. The top-down approach, if applied without proper precautions, sometimes causes problems in the edit. Another approach is multiple-variable editing, based on the Fellegi-Holt system. This approach requires more computing expertise and computer power but probably obtains results that are closer to "reality". In the multiple-variable editing system, it is necessary to determine a set of positive statements to test the relationship between the variables. Then, the edit tests each statement against the data in the household to see whether all statements are true. For any false statement, the edit will keep track, on an item-by-item basis, of invalid entries or inconsistencies. After all tests, the editing and imputation system must assess how best to change the record so that it will pass all edits. Editing teams usually use a minimum-change approach and change the smallest possible number of variables to obtain an acceptable record.

Methods of Correcting and Imputing data

Blanks and invalid and inconsistent entries in data records from "not reported", "unknown" or otherwise missing information occur in all censuses and surveys. These entries occur from respondent, enumerator, or data entry mistakes. Methods of making corrections vary depending upon the item. In most instances, we can assign valid codes to data items with reasonable assurance that they are correct by using responses from other data items within the person or household record or from the records of other households or persons.

When imputation is not needed

Sometimes an edit can use the items on the questionnaire to resolve inconsistencies. For example, we can write the editing specifications for an edit as shown in figure 1. If fertility is complete for both, the edit will work. However, the edit is clearly not complete since it only takes care of the case in which fertility is complete and accurate for both the head of household and the spouse.

Figure 1. Sample editing specifications to correct sex variable, in pseudocode

```
If SEX of the HEAD OF HOUSEHOLD = SEX of the SPOUSE
    If FERTILITY of the HEAD OF HOUSEHOLD is not blank
       If FERTILITY of the SPOUSE is blank
          (if the SEX of the head of household is not already female) Make the SEX = female  endif
          (if the SEX of the spouse is not already male) Make the SEX = male endif
       else        Do something else because they have same sex and both have fertility !!!
             [The "something" could be using the sex of the previous head, or alternating the sex of the
               Head, or using ratios of sexes of all heads for an appropriate response, etc.]
       endif
    Endif
  Else          This is the case where the head of household's fertility is blank
    If FERTILITY of the SPOUSE is not blank
       (if the SEX of the head of household is not already male) Make the SEX = male endif
       (if the SEX of the spouse is not already female) Make the SEX = female  endif
    else          Do something else because BOTH have no fertility!!!
             [The "something" could be using the sex of the previous head, or alternating the sex of the
               Head, or using ratios of sexes of all heads for an appropriate response, etc.]
    endif
  Endif
Endif
```

However, in most cases, the items themselves are insufficient to resolve problems. This paper presents two computer techniques to correct faulty data. One is the static imputation or "cold deck" method, used mainly for missing or unknown items. The other is the dynamic imputation or "hot deck" method, used both for missing data as well as for inconsistent or invalid items.

Static imputation – the "cold deck" technique.
In static or cold deck imputation, the editing program assigns a particular response for a missing item from a predetermined set, a proportional basis from a distribution of valid responses imputes the item. In the cold deck method, the program does not update the original set of variables. The values do not change from those in the initial static matrix after processing records for the first, second, tenth or any other persons. The original values provide imputations for any missing data. Static imputation is a stochastic method, as is dynamic imputation, but the values do not change over time. Some static imputations come from items that provide values that are not "unknown"; for example, for some minimum age, say age 10, every person should be "never married", and that value will be assigned when something else, or nothing, appears in the data set.

Sometimes static imputation uses a ratio method, assigning responses based on predetermined proportions. As an example of the proportional distribution of responses, suppose we have a tabulation of valid data, that is, data from completed as opposed to missing items. We might have a distribution of time worked per week by males 33 years old employed in agriculture showing that 25 per cent worked 50 hours a week; 40 per cent worked 60 hours a week; and 35 per cent worked 70 hours a week. Missing or invalid responses for time worked for males 33 years old employed in agriculture would be replaced 25 per cent of the time by 50 hours, 40 per cent of the time by 60 hours, and 35 per cent of the time by 70 hours. However, unless reliable data are available from previous censuses, surveys or other sources, this technique requires pre-tabulation of valid responses from the current census, which may not be economically or operationally feasible.

Dynamic imputation – the "Hot Deck" technique.
Another method of ridding the data of unknowns is the dynamic imputation or hot deck technique, which allocates values for unavailable, unknown, incorrect, or inconsistent entries. United States Census Bureau originally developed the method, and it and other agencies have since added refinements. Dynamic imputation uses one or more variables to estimate the likely response when an unknown (or, in some circumstances, more than one unknown) appears in the dataset. Dynamic imputation has become increasingly popular for census edits because it is easy and produces clean, replicable results. In addition, by eliminating unknowns, trends between censuses and surveys are easier to obtain since the analyst does not have to deal with the unknowns on a case-by-case basis.

For dynamic imputation, known data about individuals with similar characteristics determine the most appropriate information to be used when some piece (or pieces) of information for another individual is unknown. These characteristics include sex, age, relationship to head of household, economic status, and education. The imputation matrix itself is a set of values, similar to the cards in a deck. These matrices store, and then provide, information used when encountering unknowns. The deck constantly changes by updating (putting good values in cells) and/or by logically "shuffling the deck", so that response imputations change during data processing: hence the term "hot deck."

The values stored in the hot deck represent information about the "nearest neighbors" with similar information. Note that the nearest neighbor is usually the nearest *previous* neighbor because, especially in the top-down approach, housing units and people in those units are only considered once, and then the program moves on. So, within a village for example, when a person's maternal orphanhood is unknown, for example, the hot deck will contain information about the most recent person encountered with the same sex and age and valid maternal orphanhood. This approach is particularly important in countries having relatively large migration movements or HIV/AIDS or other unusual statistical activity. Housing characteristics are more likely to be similar within a compound or a village than to those in other parts of the country.

Hot deck – geographical considerations. If the editing program uses dynamic imputation to impute missing values, it should attempt to use data sorted by the smallest geographically defined area. This procedure should increase the probability of obtaining a correct answer, since people living in the same small geographical area are usually somewhat homogeneous with respect to their demographic, housing, and other characteristics. Where the population is not homogeneous, no correlation will exist, so the editing team must look at variables on a case-by-case basis. Also, some areas should never have certain variables – like central heating in very warm places – and the edit should consider this.

Hot deck – use of related items. Before using dynamic imputation to obtain missing values, the editing team should use related items to assign a value that is likely to be correct. For instance, if the marital status of a person is missing, the editing program will determine whether the person has a spouse in the household. If so, the program will assign the code

for married without using an imputation matrix. However, when no such evidence is present, the program may have to rely on an imputation matrix value.

Hot deck – How the order of the variables affects the matrices. National statistical offices that use imputation matrices need to determine the variables for the hot decks as they develop the order of their edits. For population items, the offices will want to edit sex and age at the beginning, so they can use these in the other imputation matrices. The overall edit should not use unedited variables in imputation matrices, although most computer packages will accept "unknown" rows or columns. Response rates and distribution of attributes within variables will assist in determining the best variables, and the most useful attributes within those variables, to assist in developing the hot decks. Subsequent imputation matrices can use the data items after editing. However, whenever possible, statistical offices should consider excluding imputed data from the imputation matrix.

For example, if the edit imputes age based on sex and relationship, the program should not update the cells in the array for this imputation matrix (sex by relationship), after imputing either the sex or the relationship. As a rule, only when age, sex, and relationship are all valid and consistent should the editing package enter age in the cell for the appropriate sex and relationship. However, sometimes the use of edited data is unavoidable because of other factors; most countries actually do impute from previously imputed values.

Hot deck – complexity of the imputation matrices. The national statistical office increases the probability of obtaining a consistent, "correct" imputation matrix value by making the imputation matrix more detailed. For example, the program could impute marital status using relationship alone. However, the likelihood of widowhood or divorce increases with age. Therefore, it makes sense to impute marital status by age and relationship. Using the age and relationship of the current person, the editing program takes the value for marital status from a person with the same characteristics in the immediately preceding valid record stored in the imputation matrix.

Nonetheless, the procedure described above can create new problems. The national statistical office usually edits questionnaire items in a fixed sequence, with age edited after marital status in a top-down approach. If this is the case, when both marital status and age are missing from a record, it is impossible to take the value for marital status from the immediately preceding record with the same age and relationship values. As a result, the program may not be able to determine the age category for this record. Another solution would be for the imputation array to have a row or column for "not reported" items. This procedure would allow the program to assign a value for marital status using the marital status category from the immediately preceding record with the same relationship and age "not reported". Two factors, however, argue against this approach. One is that "not reported" cases in the same combination are so few that it would be difficult to update the imputation array for the missing item. Secondly, it is essentially impossible to obtain proper cold deck, that is, initial values for these combinations of "unknown" values for a hot deck since they do not exist in the "real" world.

The solution to the problem described above creates more work for the data processor but results in a cleaner product. The editing program first tests to determine whether the items have valid codes. If the record for the current person does not have a valid code for the item, the imputation matrix does not use the item for this record. Data processors can facilitate the process by creating a simpler imputation array. To continue the earlier example, if the program must impute marital status because the value is missing, the imputation array will ordinarily have two-dimensions: age and relationship. If, after testing, the program finds no valid code for age, it will impute marital status by relationship alone. Because the edit for relationship comes before marital status, the relationship code will be valid. The program uses these same principles for all dynamic imputation procedures.

Hot deck – Imputation matrix development. The subject-matter staff, in collaboration with the data processors, should prepare the appropriate imputation matrices. (Some editing teams use multiple imputation matrices for the same variables, depending on already edited variables). Only valid responses update the imputation matrices; editing teams do not use allocated or imputed values. Both subject-matter specialists and data processors must check editing specifications and hot decks for consistency and completeness.

Considerable time and thought should go into the development of an imputation matrix, including research into data available from administrative records and the results of previous censuses or surveys, particularly for cold deck (initial) values. Even after research and development, editors should not apply imputation matrices randomly. When imputation

matrices are not internally consistent, considerable effort is required to reconcile them. When imputation matrices do not use standard conventions, staff must consider each one separately.

Although the "normal" procedure is to have one value for each cell in the imputation matrices, some editing teams use more than one possibility for each cell. Imagine this as beginning with a two dimensional matrix, and then adding a third dimension, like going back into a blackboard. These cells provide an extra dimension. To illustrate, if the ages of all the children in a family are unknown, as for example, in a family with four male children, the computer will not assign the same value for age four times, creating quadruplets. Instead, the program will assign four different ages. However, the program could assign the same value more than once, depending on what is stored in the matrices.

Hot deck – Standardized imputation matrices. Standardized imputation matrices can streamline the editing process. Imputation matrices with standard dimensions for various social and economic variables, such as age groups and sex, can be tested and applied quickly.

For example, the national statistical office may want to develop an imputation matrix to determine a code for language when none is given. The first place for the editing program to look will almost certainly be within the household for another person reported as speaking a given language. Failing that, the program can select the language of a previous person of the same sex and age group (having updated the imputation matrix when all three items were valid). This procedure will give a likely language, since persons speaking the same or similar languages are usually located geographically close to each other.

Or, the editing team may decide that when no language appears for anyone in the household, the program must do something else. First, for example, the edit might look for other variables to give an indirect estimate of the language used. Sometimes race, ethnicity, or birthplace gives an indication of the appropriate language to impute. If such an identifier is available, then the editing team might choose to use that to determine the language for the head of household. If not, the edit can use age and sex for imputation.

If the team decides to impute, the program assigns the head of household a language based on age group and sex. In this case, the entries in the imputation matrix will be for previous heads of household only, since all other persons in a given household receive the same language code as the head of household. At this point, if the household still has no one who reports speaking a defined language, the editing program uses the imputation matrix to assign a language to the head of household based on the head of household's age and sex. The language assigned is the most recent one in the data file spoken by another head of household of the same age and sex. Since the imputation matrix is "updated" continuously as acceptable cases are encountered, the assigned language is likely to be a language spoken in the general community.

Exceptions to the editing rules will occur at the very beginning of an edit run. Staff must be careful to take note of language changes that may occur when they move from one geographical area to another. Some countries must also be concerned with localized mixtures of language speakers. However, even in this case, unless selective under-reporting for certain languages exists, the percentage of allocated and unallocated values resulting from the imputation should be about the same. Similar procedures can impute many of the economic characteristics, such as labor force participation, time worked last week, or weeks and time worked last year, using similar characteristics. By using similar imputation matrices, the editing program can quickly check the value for the characteristics of the variables, and the editing process should move faster overall.

Seeding the deck. Sometimes it is difficult to obtain appropriately edited characteristics for the first imputation matrices in a series to seed the hot deck. Usually a statistical office does not want to include unedited items as dimensions for an imputation matrix; the edit would not use either sex or age as imputation matrix dimensions if not edited. Hence, the first few imputation matrices will use different variables that need no editing or those that cannot change in value. For the very first imputation matrix for population items, the edit might use the number of persons in the housing unit including a zero for vacant units. For housing edits in general, the first imputation matrix might also use the number of persons in housing units as the initial dimension, but the editing team might modify actions for housing items to account for vacant units. For example, if the first housing edit is for "construction material of outer walls" or "type of walls", the initial values might be based on the number of persons in the housing unit, including a value for when the unit is vacant.

After the initial use of this imputation matrix, the editing team might then want to switch to some other housing characteristics, such as “type of roof” or “tenure”. Whatever is selected must distinguish clearly between units and provide enough diversity that the same attribute will not be selected repeatedly. Recurring selection of the same attribute can give quasi-cold-deck rather than dynamic imputation (hot deck) values. Using dynamic imputation, for instance, in an army barracks “group quarters” might cause the same value to be used repeatedly if the only characteristics selected are age and sex. In this case, all of the residents would probably be male, and most would be within a limited age range. Hence, that particular matrix might not give the best results. If “tenure” has sufficient diversity, with sufficient percentages of owners and renters, this variable could work. Otherwise, the country could use different types of roof. In general, many editing teams find that by using comparable dimensions for imputation matrices, they do less checking, get their results more quickly and probably get them more accurately.

Building the edit logically

If the editing team decides to impute all or most of its items, it should develop a strategy for building the edit in a logical way. For population items, the edit should begin by considering all items potentially having unknowns. Editing teams should use information from surveys and administrative records, earlier censuses, the pilot for the census under consideration, and other information available to help determine each item’s inclusion in the first, and subsequent, imputation matrices. While development of the details of imputation matrices is very country-specific, all national statistical offices are likely to have some information available for this purpose. Testing of various sets of variables in the hot decks will assist in getting the most appropriate set for the particular country.

Many editing software packages keep track of the number of persons in the housing unit as they go along. An imputation matrix for unknown sex, for example, could allow for assignment of male or female depending on the number of occupants in the housing unit. Hence, the initial value to be selected for a person of unknown or invalid sex for a one-person house might be male. For a two-person house, the initial value might be female. For a three-person house, the value would be male and so on. The matrix would be used only as a last resort after all consistency edits, such as the sex of the head of household and the spouse and the presence of fertility information, had been tested and resolved.

How big should the imputation matrices be?

Most computer packages can accept multidimensional imputation matrices. The following points should be taken into consideration before setting up the imputation matrices:

Problems that arise when the imputation matrix is too big. One of the biggest problems that some national statistical offices have as the team of subject matter and data processing specialists work together is that of over-eager editors. It is easy to be carried away in developing the editing packages so that the programming takes much longer than necessary and slows the census or survey processing. The editing team may decide, for example, that in order to determine age, in addition to “sex”, “educational attainment” and “labor force participation”, “number of children ever born” must also be included for females. The addition of “number of children ever born” may provide a slightly better age estimate, but the increased complexity of the programming may not justify it. Editing teams have to decide how many imputation matrix dimensions will give the best results, in terms of both accuracy and efficiency. Imputation matrices that are too big (with too many cells) cannot be updated thoroughly, and cold deck values may inappropriately be used instead.

Understanding what the imputation matrix is doing. In addition to imputation matrices that are too big, paths may be confusing. It is important to make sure that the subject-matter personnel as well as the data processors are able to follow all the paths. Together, they must make sure that the imputation matrix is performing its intended task. Again, the subject-matter persons and data processors must work together to verify that each variable or dimension of the imputation matrix is implemented properly. Moreover, they must ensure that all of the combinations are working properly.

Problems that arise when the imputation matrix is too small. The imputation matrix is too small if it has too few dimensions or if, because of groupings (such as too few age groups or educational levels), the same imputation matrix value is used repeatedly before being updated. For example, without a dimension for sex in an age array, all children in a family are more likely to receive the same age when age is unknown. Subject-matter personnel should work with the data processors to test the imputation matrices for all of the different combinations and should ensure that none occurs too frequently.

Items that are difficult for imputation matrices. Some items, such as "occupation" and "industry" have proven notoriously difficult to edit. While separate imputation matrices for occupation and industry may produce inconsistent results, an effort to crosscheck all pairs of occupation and industry entries can be costly and difficult. For example, if barbers or hairdressers are found working in fish processing plants, some other type of edit is needed. In addition, the large number of occupations and industry categories can make dynamic imputation very difficult. For some items the editing team may decide that editing is counter-productive and, instead, opt to use "not stated" or "not reported." Otherwise, use of a static imputation (cold deck) approach may suffice.

Checking the edits

The basic structure of the imputation matrix in an editing software package should look something like the display in Figure 2. Editing specifications must identify the arrays used for the imputation and use cold deck values for the initial set of values.

Setting up the initial static matrix. The procedure outlined below updates the imputation matrix each time it finds a person with valid values in all three items—in this case, "relationship", "sex", and "age". However, when the editing program finds an invalid (or blank) sex, the imputation matrix selects a value based on valid relationship and sex codes (variables that have already been edited).

Figure 2. Sample set of values for a cold deck array and sample imputation code

```
…
22   A01-AGE-FM-SEXRL (2,6)
 23.   Head of household   Spouse   Child   Other relative   Parent   Not reported   .Sex
 24.          40             40       10          20           65          20        .Male
 25.          40             40       10          20           65          20        .Female
…
40 if AGE = 0:98
41   let A01-AGE-FM-SEXRL (SEX,RELATIONSHIP) = AGE
42 else
43   message 'Age is unknown, so imputed' AGE
44   write ' Age is unknown, so imputed, Age = ' AGE
45   impute AGE = A01-AGE-FM-SEXRL (SEX,RELATIONSHIP)
46   message 'AGE is now known' AGE
47 endif
```

Messages for error listings. Editing packages should provide several methods to make certain that they implement edits and imputations properly. Two of these features, message commands and write commands, are reviewed here. One source of information is the display of a message, as seen above in figure 2. This command generates specific messages and summary counts (the total number of times the message occurs) for levels of geography (e.g., enumeration area, minor civil division, major civil division) as well as for each questionnaire. For all of the questionnaires, a summary report might look something like figure 3:

Figure 3. Example of a summary report for number of imputations per error

Count	*Error number*	*Message*	*Line number*
-	14-1	Too many children per woman	2629
-	14-2	Too many children per woman	2645
2	14-3	Boys present not stated	2669
2	14-4	Girls present not stated	2678
33	14-5	Month last birth not stated	2723
7	15-6	No children ever born; age difference between mother & child OK	2892

NOTE: Here "14" simply refers to item 14 in a given series; errors are numbered sequentially.

The following example, from the Malawi 2008 Census, shows part of a listing for determining that one and only head is in the household[5]. Since the number of errors is low, even if the edit were not done, the data could still be easily used for planning and policy formation. However, the edit provides an aesthetic data set for consistent tabling and trends analysis.

Figure 4. Example of a listing summary for Malawi 2008 Census

```
1718  336574    -    ******************************** ...                    -
1719  336574    -    *******  Age & Head    ********* ...                    -
1720  336574    -    ******************************** ...                    -
1805    1546   0.1  *P00-1* Head is not first person, is %2d...        1748490
1823     877   0.1  *P00-2* No head of household, first person 14+... 1748490
1835      62   0.0  *P00-3* No head 14+, first person becomes head... 1748490
1850    5074   0.3  *P00-4* Too many heads of household - 1 ...        1748490
1860    5238   0.3  *P00-5* Remaining heads made other RELATIONSHI... 1748490
1874     939   0.1  *P00-6* After head edit, not one and only one ... 1748490
1889    2301   0.1  *P00-6a* Spouses too young made other relative... 1748490
1909    1062   0.1  *P00-6ax* Multiple spouses for unmarried head...  1748490
1911    1062   0.1  *P00-6ax* Multiple spouses for unmarried head...  1748490
1929      44   0.0  *P00-6a1* Crazy case where spouse is visitor a... 1748490
1949      89   0.0  *P00-6a3* Crazy case where spouse is visitor a... 1748490
1998      12   0.0  *P00-6a1* Extra spouses who are visitors...       1748490
2017    1483   0.1  *P00-6a2* Extra spouses not married...            1748490
```

The following example is from the 2006 Lesotho Census with a non-representative sample of part of the listing showing sisterhood characteristics; collection of sisterhood characteristics provides information on maternal mortality. The percentages shown here, some over 3 percent, are large, showing either that the enumerators did not ask the questions or the respondents did not always respond. These are the last items on the questionnaire, so enumerator and respondent fatigue may have come into play.

Figure 5. Example of a listing summary for Lesotho 2006 Census

```
4388   21471    -    ...                                                   -
4389   21471    -    *******  Sisterhood Characteristics *********...      -
4390   21471    -    ...                                                   -
4401    1449   1.2  *G45-1* Total sisters out of range [%2d] illeg...  124839
4410    2897   2.3  *G45-2* Dead sisters out of range [%2d] illega...  124839
4419    3791   3.0  *G45-3* Pregnant sisters [%2d] illegal...          124839
4426    3895   3.1  *G45-4* At birth sisters [%2d] illegal...          124839
4433    4908   3.9  *G45-5* Week 6 sisters [%2d] illegal...            124839
4440     103   0.1  *G45-6* Sum of Dead Sisters [%2d][%2d][%2d] gr...  124839
4453       8   0.0  *G45-7* Sum of Dead Sisters [%2d][%2d][%2d] gr...  124839
4461     616   0.5  *G45-8* Dead Sisters [%2d] greater than total ...  124839
```

A report organized by questionnaire (figure 6) might give the questionnaire number, including all of the specified geographical codes.[6] The report could then list the errors found in the program, by item (in this case age), and by line number in the software program, seen below on the right. In this example, the age was blank, but the imputation matrix provided the age of 48, based on the relationship and sex of this person. For this case, the specific age was unknown, but the message command could also write that information, also, if desired.

Of course, it makes sense to list all individual errors on sample tests or small, selected data sets, for even mid-size countries. But, the amount of output in production runs could be so large and cumbersome (and leading to meaningless after a while), that a trigger might be set to turn off all or parts of the individual questionnaire problems for the complete census. The summary statistics would remain, of course.

[5] All of the examples shown here come from recent CSPro edits. UNFPA is planning to make both the completed edits available online (without microdata), as well as prototype edits based on the UN Principles and Recommendations, Revision 2.

[6] Both IMPS and CSPro provide vertical listings as shown in Figure 6. Many programmers prefer to use these since they are built into the system. However, the WRITE statement provides horizontal listings, as shown in the text, and subject matter specialists usually find these easier to use since they look more like the item layouts.

Figure 6. Sample report for errors in a questionnaire

Questionnaire ID: 01 01 017		*Line number*
AGE (1) =	Age is unknown, so imputed	#46
AGE (1) = 48	Age is now known	

Custom-made error listings – WRITE statements. The software might also provide another command, allowing for a more detailed analysis of the editing specifications and edit flow. The command may be used to show the information before a change is made, and then all of the changes made. Finally it shows the record or records again, with the changes made. In this way, the analyst can make certain that the edit follows all paths properly. The results may be as shown in figure 7. The first line of the output gives the variables (e.g., province, relationship, sex, age). Then, the incoming data are shown, followed by the error (in this case, no age), and then the data after the change was made.

Figure 7. Example of supplementary error listing by questionnaire including multiple variables

	Province	*District*	*Head of household*	*Relation*	*Sex*	*Age*
Incoming data	01	01	17	1	1	
Error		Age is unknown, so imputed age = BLANK				
Edited data	01	01	17	1	1	48

The following example comes from the 2007 Ethiopia census, and shows errors being corrected for persons 3, 4, and 5 in the unit. The five people in the house are displayed with all of their captured data before the edit, and then what the edit is finding to be corrected, and, finally, the corrected household.

Figure 8. Example of a write listing for Ethiopia 2007 Census

```
BARCODE REGION  ZONE  WEREDA TOWN SUB_CITY  SA KEBELE   EA    HHNO   HUNO
-----------------------------------------------------------------------

PN RS RH SX AG RL MT ET DS 1 2 3 4 5 6 7 8 9 0 1 2 3 CS YR PR ZN MO FA LT SC HG WL RS LY ES MS MH FH MA FA MD FD LB
01 01 01 01 31 01 05 67 02                              08 01 01       01 97 12 01    01 07 01
02 01 06 01 34 01 05 67 02                              08 01 01       01 97 17 01    01 03 01
03 01 09 02 30 01 05 05 02                              07 02 02          97 05 01    01 05 04 00 00 00 00 00 00 00
04 01 09 02 20 01 05 05 02                              03 02 02       02 98    03 01 03    01 01 00 00 00 00 00 00
05 01 09 01 01 01 05 05 02                              08                      03 07 08    01
P18-3 No literacy   , but schooling 97, so literate, PN =  3
P20-20 Unable to read and write 98 because never attended school   , PN =  4
P16-1 Mother's vital status invalid =         PN =  5
P17-1 Father's vital status invalid =         PN =  5
PN RS RH SX AG RL MT ET DS 1 2 3 4 5 6 7 8 9 0 1 2 3 CS YR PR ZN MO FA LT SC HG WL RS LY ES MS MH FH MA FA MD FD LB
01 01 01 01 31 01 05 67 02                              08 01 01       01 97 12 01    01 07 01
02 01 06 01 34 01 05 67 02                              08 01 01       01 97 17 01    01 03 01
03 01 09 02 30 01 05 05 02                              07 02 02       01 97 05 01    01 05 04 00 00 00 00 00 00 00
04 01 09 02 20 01 05 05 02                              03 02 02       02 98 00 03 01 03    01 01 00 00 00 00 00 00
05 01 09 01 01 01 05 05 02                              08 01 01 01 01
```

The following is an example for housing information from the 2007 Fiji Census:

Figure 9. Example of a write listing for Fiji 2007 Census

```
EAREA    BN      GN LQ WL CN RM WS DR EL LG CF TL TN LD SH FS FS MO TS PS RT CS NO OT LC BR BT RV SE BY WO CR CR MT GN
BC SL WT
010100600 4  999999 01 03 01 02 01    01 01    01 01 01 00 01 00 00 00 00 00 00 00 00 00 00 01 00 00 00 00    00 60
00 00 00
*H4-7* - Ever dry up from TENURE, adry = 02, dryup =
*H07-1* Cooking Fuel from electric not current used, cooking fule =  , elec = 1, light = 1
*H13-1*, Car unknown =
*H13-3*, Outboard motor unknown = 60
*H13-4*, Generator unknown =
EAREA    BN      GN LQ WL CN RM WS DR EL LG CF TL TN LD SH FS FS MO TS PS RT CS NO OT LC BR BT RV SE BY WO CR CR MT GN
BC SL WT
010100600 4  999999 01 03 01 02 01 02 01 01 03 01 01 01 00 01 00 00 00 00 00 00 00 00 00 01 00 00 00 00 01 00 00 00
00 00 00
```

This procedure assists the editing team in determining whether the edit is taking the proper paths. Testing is an important part of census and survey editing. The following method represents one possible way of testing editing procedures. The process might begin by having specialists perform the analysis systematically by creating a "perfect" household. A perfect household is one that is a complete household—head of household, spouse, children, other relatives, and non-relatives—with all their characteristics (which are also 'perfect' in the sense that they are internally consistent and consistent with other household members). The perfect household must pass all of the edits without any errors. Then, the unit is duplicated over and over again in a single file. The procedure continues as outlined below:

(a) The data processors introduce a single error into each household, in sequence, to correspond to the sequence of the editing specifications and the editing program;
(b) The analyst then checks all of the paths early in the editing process;
(c) Once the edit follows all paths properly, data processors run a sample of the whole data set, looking for idiosyncrasies in the actual data set and making modifications as necessary;
(d) Finally, the data processors run the whole dataset.

The data processor may decide to turn them off for lower levels (like for each questionnaire) when satisfied that the messages are working properly and the appropriate modifications have been made. If large countries were to run their whole data sets with message statements left in for each questionnaire, the resulting quantity of lines and paper would be prohibitive. However, the summary report for these messages should continue because it gives useful information for the various levels of geography. The output will look something like that in figure 6.

Computer edits usually include a safeguard procedure. The edit trail shows all data changes and tallies for cases of changes and substituted values. Reference to the edit trail will determine whether the number of changes is sufficiently low for the group of records to be accepted. If a particular item has too many errors, the item may not have been adequately pretested, either on its own, or in relation to other items, indicating that enumerators or respondents did not understand the item. Sometimes enumerators get confused, for example, and collect fertility information only from male adults and not from females. If this type of data collection is systematic, the editing team might have the programmers move the fertility data from the males to the females in a married couple. Otherwise, the editing team can do little at this stage to correct the error.

Usually the editing program needs to look at several different (or sequential) files to cover all situations. In addition, the data processors will need to make changes because of faulty syntax or logic. Even the most experienced data processing specialists occasionally key a "greater than" sign in place of a "less than" sign, and the error is found only after several runs are made since the particular problem may not be immediately apparent. Similarly, small flaws in logic may not be apparent at first. Again, the subject matter and data processing specialists need to work together to resolve these issues early in the editing process, if possible.

Frequency distributions

Besides the listings and "before-and-after" displays, a third type of error listing – frequency distributions – assists in determining whether the edits are working as they should. Frequency distributions can take

several forms. They can show the distribution of values before edit when the program determines that the value has to change; or they can show the edited values; or, in some cases, they can show both. But, the frequencies show, subject matter specialists show test, using Excel or some other spreadsheet, that the unedited distribution, the editing distribution, and the frequencies of the changes all show what the analysts expect to see. That is, edits for sex tend to over-allocate females when presence of fertility information (unless the analysts are very careful in their specifications), so a surplus of females in the frequencies should not be surprising.

The following example shows a frequency distribution from early runs on part of the 2008 Sudan census processing. Note that the @17 shows an illegal code which the edit must address:

Figure 10. Example of a frequency distribution for Sudan 2008 Census

Imputed Item Q18_ATTAINMENT: Education Attainment - all occurrences

Categories	Frequency	CumFreq	%	Cum %	Net %	cNet %
1 No Qualification	105	105	2.2	2.2	2.4	2.4
2 Incomplete Primary	1564	1669	33.5	35.7	35.3	37.7
3 Primary 4	529	2198	11.3	47.0	11.9	49.6
4 Primary 6	492	2690	10.5	57.6	11.1	60.7
5 Primary 8	302	2992	6.5	64.0	6.8	67.5
6 Junior 3	251	3243	5.4	69.4	5.7	73.2
7 Junior 4	58	3301	1.2	70.7	1.3	74.5
8 Secondary 3	95	3396	2.0	72.7	2.1	76.6
9 Secondary 4	5	3401	0.1	72.8	0.1	76.7
10 Post Secondary Diploma	2	3403	0.0	72.8	0.0	76.8
11 University Degree	154	3557	3.3	76.1	3.5	80.3
12 Post Graduate Diploma	10	3567	0.2	76.3	0.2	80.5
13 Master	52	3619	1.1	77.5	1.2	81.7
14 Ph.D	1	3620	0.0	77.5	0.0	81.7
15 Khalwa	1	3621	0.0	77.5	0.0	81.7
@17	144	3765	3.1	80.6	3.2	85.0
@98	667	4432	14.3	94.9	15.0	100.0
NotAppl	240	4672	5.1	100.0		
TOTAL	4672	4672	100.0	100.0		

The following example shows the frequency distribution for number of rooms in the house during the rehabilitation of the 1990 Zambia census in preparation for trends analysis:

Figure 11. Example of a frequency distribution for additional edit for Zambia 1990 Census

Input: 1IN100.DAT Program: ZAMHOUSE

ROOMS

Values Imputed	Number of Imputations	Percent	Cum. Percent
< 1	1,415	37.21	37.21
1	2,185	57.45	94.66
2	121	3.18	97.84
3	22	0.58	98.42
4	16	0.42	98.84
5	23	0.60	99.45
6	21	0.55	100.00
> 6	-	-	-
	3,803		

How many times to run the edit

As noted, as soon as the questionnaire is set, development and testing of edit specifications and programs should begin. Individual items should be developed separately when a top-down approach is used, but even when several variables are to be edited at the same time, edits for individual items will need to be tested on small parts of the whole data set. The edit specifications should be developed by the subject matter specialists, and then individual edit programs implemented by the programmers. The total edit can then be built, and run on larger and large parts of the data set, refined along the way. In general, for both the parts of the program and the whole program, it is a good idea to run an editing program three times:

The first edit run supplies the imputation matrices with real values rather than the values created in the initial static matrix. Some countries use data from other sources— either a previous census or survey or administrative records— to supply cold deck values for an array. The data processor runs the complete dataset, or a large part of it, to supply values for the imputation matrix. Cold deck values from the actual dataset are more likely to be accurate and current. The edits use only about two percent of this initial static matrix: the rest are dynamic imputation values. Some newer packages allow filing the static matrix from the real data in the census; here, a run through the whole data set provides real values.

The second edit run performs the actual editing. The second edit run may consist of a single run of all edits or several repeat runs may be needed in order to cover all situations. At this time, the data processors will need to make changes in order to correct errors resulting from faulty syntax or logic. In addition, even the most experienced data processing specialists may make mistakes and, since the particular problem may not be immediately apparent, the error may be found only after a few runs. Similarly, small flaws in logic may not be apparent at first.

The third edit run makes certain (1) that no errors remain in the data set, and (2) that the editing program did not introduce new errors. When the processors run the edit this last time, no errors should appear in the error listings. If errors remain, the logic of the edit is probably faulty, so the data processor needs to modify it. In addition, this run usually tells the data processor if the edit accidentally introduced new errors by the logic of the edit.

Saving original responses

The original data set obtained before computer edit, whether keyed or scanned, should be kept and archived. As new demographic and other analytical techniques become available, statistical offices may want to revisit their unedited data to test these. Also, the original responses for key variables, like relationship, age, sex, marital status, orphanhood, and fertility and mortality information, should be kept on the individual records. While these take up some room, and make the data set marginally larger, many demographic methods use the unedited data to test various hypotheses. Another series of values – the imputation flags – may also appear on the final data set, as described below.

Imputation flags

Imputation flags are one method used to retain information about unedited data. Many editing teams are concerned about the loss of potential information when unedited responses are changed. In cases where a value is changed because of an inconsistency, the editing teams may wish to save the original value or values in order to carry out further demographic or error analysis after the census. Both subject-matter specialists and programmers will want to analyze various aspects of the missing, invalid, or inconsistent data. Members of the editing team need to make sure that the imputed and un-imputed distributions are consistent, to see if any systematic error appears in the editing and imputation plan. For example, sometimes data processing specialists accidentally use only cold deck values because the program neglects to update the imputation matrix. If the country conducted a census pretest, the editing team may need to investigate the relationships between some of the variables after the pretest in order to finalize the questionnaire. In prior censuses, before microcomputers with large hard disks were common, many statistical offices did not have the space on their tapes or other storage media to maintain extra data; however, these days, for most countries, keeping information about unedited data is no longer a problem.

Some countries choose to maintain a simple, binary accounting variable as a flag for each item. This method is simple and takes up a single byte for each variable. For example, the United States Census Bureau places imputation flags for each variable at the end of each record, for both housing and population records. For each housing variable, for example, the variable for the flag was initially “0”, but was changed to “1” if the original item is changed in any way. The program does not retain the original value, although offices sometimes compile these, either for each record or taken together.

Other methods are available to save unedited responses. In the example in figure 12, the national statistical office has changed a spouse’s age from 70 to 40 using an imputation matrix. The national statistical office can easily put the pre-imputation value, in this case 70, in the area reserved for imputation flags and reserve the variable used for published tabulations for the allocated value, in this case 40. In order to examine

changes in the data set, the statistical office can make frequency distributions or cross-tabulations of the allocated and the unallocated values. If, following this analysis of the effects of the edits on the data set, the tabulations based on the edit appear suspicious or anomalous; the editing teams might want to consider changing the edit or part of the edit flow. And, because hard disk capacities have increased so much in recent years, all initial values can be stored on the records for later use. Offices will probably want to maintain at least two files since a file of all edited data is likely to run slightly faster.

Figure 12. Sample population records with flags for imputed values

Person	*Sex*	*Age*	*Children ever born (CEB)*	*Sex flag*	*Age flag*	*CEB flag*
1	1	40	BLANK			1
2	2	40	7		70	

One problem in the use of imputation flags is that the procedure just described takes up considerable space in the computer. When the flags repeat each variable, the edited data set will be approximately twice as large as the unedited data set. For many countries, this would be unacceptable for long-term storage. However, the original data and the edits could be stored for later reconstruction.

Countries with very large populations might prefer to use imputation flags on a sample basis for research purposes. For example, a country might want to create a data set with every 100th housing unit. Then the edit would run with imputation flags on this smaller set, helping to evaluate how the edit affects the quality of the data and determine what differences exist between the unedited and edited data.

Discussion and Conclusions

This paper describes the use of top-down census and survey computer editing methods. A few countries implement another, more complicated, procedure for computer editing, known as multiple-variable editing, as discussed earlier. Fellegi and Holt (1976) were the first to develop these procedures, which are usually applied to the most important variables in a census or survey: age, sex, relationship, and marital status. However, this method can be applied to any group of variables, or all of the variables on a census or survey questionnaire. In the method, the edit program looks at responses to these items simultaneously for one person or for all of the persons in a household in order to identify missing or inconsistent responses. When unknown (blank), invalid, or inconsistent entries are found, a series of tests determine which of the selected items is most in error, and that one is changed first. Then, the tests are repeated to determine that no invalids and inconsistencies remain; if they do, an edit changes the item with the most remaining problems. The procedures are repeated until no errors remain.

Statistics Canada developed the Fellegi-Holt approach and used it for Canadian censuses from 1976 to 1991. For the 1996 Canada Census, this approach was refined and called the New Imputation Methodology (NIM). It permitted for the first time, “minimum-change imputation of numeric and qualitative variables simultaneously for large [editing and imputation] problems” (Bankier, Houle, and Luc, n.d.). The New Imputation Methodology uses donors for items, with the hope that all missing or inconsistent information can come from a single donor or a few donors. In order to obtain all or most of the information from a single donor, whole data records must be stored in the computer’s memory. Then, when both age and sex are unknown or invalid, the same, stored variable provides values for both items.

If the editing process is carried out using traditional dynamic imputation or hot deck method, the imputation information for a series of questionnaire items may come from many different individuals, depending on the information used to update the imputation matrix. For example, if person A’s sex, relationship and marital status are correct, these values will update the appropriate imputation matrices. If A’s age is missing or invalid, it will, of course, not be used to update imputation matrices. In fact, other items will update that value. So, if the next person has an inconsistent sex and “sex” is imputed, person A will donate the sex. If the age is also unknown, the editing program will use some other person’s age.

The objectives of an automated hot deck imputation methodology should be as follows:

(a) The imputed household should closely resemble the failed edit household;
(b) The imputed data for a household should come from a single donor, if possible, rather than two or more donors. In addition, the imputed household should closely resemble that single donor;
(c) Equally good imputation actions, based on the available donors, should have a similar chance of being selected to avoid falsely inflating the size of small but important groups in the population (Bankier, Houle, and Luc, n.d).

This paper discussed the role of computer editing as part of the total census process, as the data go from the field, through capture, editing, tabulation, and dissemination. Computer editing can be very simple, with the use of unknowns for invalid or inconsistent data, or can be very complex, using dynamic (hot deck) imputation or nearest-neighbor assignment. Unfortunately, these types of edits remain in the realm of art rather than science. Statistical offices use many different approaches to resolving invalid and inconsistent entries, but because of time and financial constraints, and the general nature of data collected from humans, cannot know whether they have used the best procedures. But, as we continue to apply these methods to more and more censuses and surveys, we are deriving sets of best practices. While these methods currently focus on censuses, we can also apply them to surveys, particularly intercensal surveys requiring comparability with the country's censuses. The methods can also be used with administrative records, like births and deaths, immigrants and emigrants, etc., to obtain data for estimates and projections. And, the methods may also prove useful in other types of data collection activities.

References

Bankier, M., A.-M. Houle and M. Luc (n. d.). Canadian census demographic variables imputation. Manuscript.

Boucher, L. (1991). Micro-editing for the annual salary of manufacturers: what is the value added? In Proceedings of the Annual Research Conference. Washington D.C.: United.States Bureau of the Census, pp. 765-781.

Fellegi, I.P, and D.Holt, (1976). A systematic approach to automatic edit and imputation. Journal of the American Statistical Association, vol. 71, No. 353 (March), pp. 17-35.

Granquist, L. (1997). The new view on editing. International Statistical Review, Vol. 65, No. 3, New York: Academic Press, pp. 381-387.

Granquist, L. and Kovar, J.G. (1997). Editing of survey data: how much is enough? In Survey Measurement and Process Quality, Lyberg et al (eds), New York: Wiley and Sons pp. 415-435.

Naus, J.I. (1975). Data Quality Control and Editing. New York: Marcel Dekker.

Nordbotten, S. (1963). Automatic editing of individual statistical observations. Conference of European Statisticians, Statistical Standards and Studies, No 2. New York: United Nations.

Pierzchala, M. (1995). Editing systems and software. In Business Survey Methods. B.G. Cox, and others, eds. New York: John Wiley & Sons, pp. 425-411.

Pullum, T.W., T. Harpham and N. Ozsever (1986). The machine editing of large-sample surveys: the experience of the World Fertility Survey. International Statistical Review, Vol. 54, 311-326.

Statistics Canada (1998). Statistics Canada Quality Guidelines, 3rd Edition. Ottawa: Statistics Canada.

United Nations (1998). Principles and Recommendations for Population and Housing Censuses, Revision 1. Statistical Papers, Series M, No. 67/Rev.1. Sales No. E.98.XVII.8.

United Nations (1999). Standard Country or Area Codes for Statistical Use, Statistical Papers, Series M, No. 49/Rev.4. Sales No. M.98.XVII.9.

United Nations (2002) Handbook for Editing Censuses Statistical Papers, Series F, No. 82/Rev.1. ST/ESA/STAT/SER.F/82

Appendix A Censuses where some of these methods were applied	
Country	Census Years
American Samoa	1974, 1980, 1990, 2000
Ethiopia	2007
Fiji	1996, 2007
Ghana	1984, 2000, 2010
Grenada	2001
Guam	1980, 1990, 2000
Indonesia	1980, 2010
Kenya	1999
Kiribati	2005
Lesotho	1996, 2006
Malawi	1998, 2008
Maldives	2006
Marshall Islands	1973, 1980, 1988
Micronesia	1973, 1980, 1994, 2000
Northern Marianas	1973, 1980, 1990, 1995, 2000
Palau	1973, 1980, 1990, 1995, 2000, 2005
Papua New-Guinea	1990
Samoa	2001
Sierra Leone	2004
Solomon Islands	1999
South Africa	2001
Sudan	2008
Tanzania	2002
Timor Leste	2004
Tonga	1996, 2006
Uganda	1991, 2002
US Virgin Islands	1980, 1990, 2000
Vanuatu	1989
Zambia	2000
Note: For some, processing occurred during the census, for others it was during preparation or during analysis (including own children estimation).	