Chapter 8.
Accuracy of the Data

MASTER ADDRESS FILE AND ENUMERATION PROCEDURES

The majority of addresses in the United States are in what is known for census purposes as the mailout/mailback area, which in general consists of areas with predominantly city-style mailing addresses. The original source of addresses on the Master Address File (MAF) for the mailout/mailback areas was the 1990 Census address file, the Address Control File (ACF). The first update to the ACF addresses is a U.S. Postal Service (USPS) Delivery Sequence File (DSF) of addresses. The November 1997, September 1998, November 1999, and April 2000 DSFs were incorporated into the MAF.

Until shortly before the census, the ACF addresses and the November 1997 and September 1998 residential DSF addresses constituted the MAF. These addresses were tested against Census Bureau geographic information to determine their location at the census block level. The geographic information is maintained in the Census Bureau's Topologically Integrated Geographic Encoding Referencing (TIGER®) system. When an address on the MAF can be uniquely matched to the address range in TIGER® for a street segment that forms one of the boundaries of a particular block, the address is said to be geocoded to that block. Valid and geocoded addresses appeared on each address list used for a field operation.

The Block Canvass operation was the next major address list operation in the mailout/mailback areas for Census 2000, taking place in January through May 1999. There was a 100-percent canvass of every block. Every geocoded address was printed in a block-by-block address register, and Block Canvassing listers identified the addresses as verified as a housing unit (with possible corrections to the address); a delete (no such address); a duplicate, implying the unit exists elsewhere on the list with a different, unmatchable designation, such as a different street name or building name; uninhabitable; or nonresidential.

Occurring in approximately the same time frame as Block Canvassing was a cooperative address list check with local governmental units throughout the country, called Local Update of Census Addresses (LUCA) 98. In LUCA 98, the participating governmental units received an address list and were asked for input mostly on added units but also on deleted units and corrected street names or directionals. The outcome of this operation was similar to that of Block Canvassing; units were added to and deleted from blocks, and address corrections were made.

The Decennial Master Address File (DMAF) was created in July 1999. This was the file used for printing most of the Census 2000 questionnaires. In the mailout/mailback areas, the operations that had yielded housing units and their status before this initial printing stage were the ACF, the November 1997 DSF, the September 1998 DSF, LUCA 98, and Block Canvassing.

Following the creation of the initial DMAF, there were updates to the DMAF. Addresses were added by the November 1999, February 2000, and April 2000 DSFs. Address update operations that occurred subsequent to the creation of the initial DMAF were the LUCA 98 field verification and appeal processes. Units receiving a conflicting status from the Block Canvassing and the LUCA 98 operation were sent for field verification by the Census Bureau; the results of the field verification were sent to the governmental units. At this stage the governmental unit could appeal the Census Bureau's findings for particular units. At an appeal, the Census Bureau and the governmental unit submitted their evidence of the status of a housing unit for independent review, and a ruling was issued. Both the field verification and the appeal process had the potential to change the status of a housing unit.

A final operation in mailout/mailback areas that added addresses before Census Day was the New Construction operation, another cooperative effort with participating governmental units. This operation used governmental units' local knowledge to identify new housing units in February and March of 2000.

After mailout/mailback, the second most common method of questionnaire delivery was update/leave. The address list for update/leave areas was constructed during a Census Bureau field operation called Address Listing rather than from the ACF and DSF, because the addresses are primarily noncity-style. Census employees were sent to the field with maps of their assignment areas and were instructed to record the city-style address, noncity-style address or location description, or possibly some combination of the above, for every housing unit. In addition, the location of the unit was noted on the census map with what is known as a map spot. This operation took place in the fall of 1998.

At the completion of the processing of the address listing data, it was possible to tabulate the number of housing units in each block. Because the housing units in these areas may have nonstandard mailing addresses and may be recorded in census files solely with a location description, the governmental units participating in the local review operation in these areas were sent lists of housing unit counts by block. This operation was called LUCA 99. When the LUCA 99 participant disagreed with a Census block count, that block was sent out for LUCA 99 recanvassing, in which census employees were redeployed to make updates to the address list. There was also a LUCA 99 appeal process for settling housing unit status discrepancies, which has the potential to add units to the address list. The LUCA 99 recanvassing and LUCA 99 appeal process took place at various times during the updating of the DMAF. Most of the LUCA 99 entities had their recanvassing results processed before creation of the initial DMAF, but many did not. There were DMAF updates designed specifically for getting late recanvassing and appeal results added into the census files in time for USPS delivery of a questionnaire.

The last address list-building operation in the update/leave areas was the Update/Leave operation itself. This operation was responsible for having a census questionnaire hand-delivered at every housing unit. In the process the MAF and the maps were updated.

In the most remote areas of the United States, the housing units were listed at the time of Census 2000 as the persons within them were enumerated. These operations were called List/Enumerate and Remote Alaska enumeration. This was the only source of addresses in these areas. All housing units were map spotted at the time of enumeration.

For some other regions of the country, where the address list had already been created, it was thought that an enumeration of the population would be more successful than mailback of the forms. Here an update/enumerate operation was instituted. There are two types of update/enumerate areas. The urban areas had passed through all the mailout/mailback operations up through the point of the creation of the initial DMAF, and the rural areas had passed through Address Listing, and sometimes LUCA 99, by the time of the creation of the initial DMAF. Because of these separate paths taken, it was necessary to distinguish between the urban and rural update/enumerate areas.

Another special enumeration is urban update/leave, which took place in areas where mail delivery was considered to be problematic. The addresses had passed through all the operations of the mailout/mailback areas up through the creation of the initial DMAF, but the area was visited by enumerators during the census, and, therefore, additions, deletions, and corrections to the address list were made.

People who did not receive a questionnaire at their house could submit a Be Counted Form, or they could call Telephone Questionnaire Assistance and have their information collected over the phone. Addresses from these operations that did not match those already on the DMAF were visited in a Field Verification operation to determine if they exist. Verified addresses were added to the address list.

One more source of information about housing units listed on the DMAF is the Nonresponse Follow-up (NRFU) operation. During NRFU, enumerators follow up on units that had not returned a preaddressed census form. Units in NRFU can possibly be deleted or deemed vacant. At the same time, units that do not appear on the address list or maps could be added and enumerated concurrently. This operation occurs in mailout/mailback, update/leave, and urban update/leave areas.

SERVICE-BASED ENUMERATION

Service-Based Enumeration (SBE) was designed to account for persons without usual residence that use service facilities (i.e., shelters, soup kitchens, and mobile food vans). Only people using

the service facility on the interview day were enumerated. In addition, people enumerated in targeted nonshelter outdoor locations and persons without usual residence that filed Be-Counted Forms (BCF) augmented the SBE count. The final total was included in the total population. This component of the enumeration should *not* be interpreted as a complete count of the homeless population.

CONFIDENTIALITY OF THE DATA

The Census Bureau has modified some data in this data release to protect confidentiality. Title 13, United States Code, Section 9, prohibits the Census Bureau from publishing results in which an individual's data can be identified.

The Census Bureau's internal Disclosure Review Board sets the confidentiality rules for all data releases. A checklist approach is used to ensure that all potential risks to the confidentiality of the data are considered and addressed. Questions about confidentiality may be addressed to: webmaster@census.gov Attention Policy.

Title 13, United States Code

Title 13 of the United States Code authorizes the Census Bureau to conduct censuses and surveys. Section 9 of the same Title requires that any information collected from the public under the authority of Title 13 be maintained as confidential. Section 214 of Title 13 and Sections 3559 and 3571 of Title 18 of the United States Code provide for the imposition of penalties of up to 5 years in prison and up to $250,000 in fines for wrongful disclosure of confidential census information.

Disclosure Limitation

Disclosure limitation is the process for protecting the confidentiality of data. A disclosure of data occurs when someone can use published statistical information to identify an individual that has provided information under a pledge of confidentiality. Using disclosure limitation procedures, the Census Bureau modifies or removes the characteristics that put confidential information at risk for disclosure. Although it may appear that a table shows information about a specific individual, the Census Bureau has taken steps to disguise the original data while making sure the results are still useful.

Data Swapping

Data swapping is a method of disclosure limitation designed to protect confidentiality in tables of frequency data (the number or percentage of the population with certain characteristics). Data swapping is done by editing the source data or exchanging records for a sample of cases when creating a table. A sample of households is selected and matched on a set of selected key variables with households in neighboring geographic areas that have similar characteristics (such as the same number of adults and same number of children). Because the swap often occurs within a neighboring area, there is no effect on the marginal totals for the area or for totals that include data from multiple areas. Because of data swapping, users should not assume that tables with cells having a value of one or two reveal information about specific individuals.

NONSAMPLING ERROR

In any large-scale statistical operation, such as Census 2000, human- and computer-related errors occur. These errors are commonly referred to as nonsampling errors. Such errors include not enumerating every household or every person in the population, not obtaining all required information from the respondents, obtaining incorrect or inconsistent information, and recording information incorrectly. In addition, errors can occur during the field review of the enumerators' work, during clerical handling of the census questionnaires, or during the electronic processing of the questionnaires.

While it is impossible to completely eliminate nonsampling error from an operation as large and complex as the decennial census, the Census Bureau attempts to control the sources of such error during the collection and processing operations. Described below are the primary sources of nonsampling error and the programs instituted to control this error in Census 2000. The success of

these programs, however, was contingent upon how well the instructions actually were carried out during the census. As part of the Census 2000 evaluation program, both the effects of these programs and the amount of error remaining after their application will be evaluated.

Types of Nonsampling Error

Nonresponse. Nonresponse to particular questions on the census questionnaire or the failure to obtain any information for a housing unit allows for the introduction of bias into the data because the characteristics of the nonrespondents have not been observed and may differ from those reported by respondents. As a result, any imputation procedure using respondent data may not completely reflect these differences either at the elemental level (individual person or housing unit) or on the average. Some protection against the introduction of large biases is afforded by minimizing nonresponse. Characteristics for the nonresponses were imputed by using reported data for a person or housing unit with similar characteristics.

Respondent and enumerator error. The person answering the mail questionnaire for a household or responding to the questions posed by an enumerator could serve as a source of error. Although the question wording was extensively tested in several experimental studies prior to the census, the mail respondent may overlook or misunderstand a question, or answer a question in a way that cannot be interpreted correctly by the data capture system. The enumerator may also misinterpret or otherwise incorrectly record information given by a respondent, may fail to collect some of the information for a person or household, or may collect data for households that were not designated as part of the sample. To control problems such as these with the field enumeration, the work of enumerators was monitored carefully. Field staff were prepared for their tasks by using standardized training packages that included hands-on experience in using census materials. A sample of the households interviewed by each enumerator was reinterviewed to control for the possibility of fabricated data being submitted by an enumerator.

Processing error. The many phases involved in processing the census data represent potential sources for the introduction of nonsampling error. The processing of the census questionnaires completed by enumerators included field review by the crew leader, check-in, and transmittal of completed questionnaires. No field reviews were done on the mail return questionnaires for this census. Error may also be introduced by the misinterpretation of data by the data capture system or the failure to capture all the information that the respondents or enumerators provided on the forms. Write-in entries go through coding operations, which may also be a source of processing error in the data. Many of the various field, coding, and computer operations undergo a number of quality assurance and quality control checks to help ensure their accurate application.

Reduction of Nonsampling Error

To reduce various types of nonsampling errors, a number of techniques were implemented during the planning, development of the mailing address list, data collection, and data processing activities. Quality assurance methods were used throughout the data collection and processing phases of the census to improve the quality of the data. A reinterview program was implemented to minimize the errors in the data collection phase for enumerator-filled questionnaires.

Several coverage improvement programs were implemented during the development of the census address list and census enumeration and processing to minimize undercoverage of the population and housing units. These programs were developed based on experience from the 1990 census and results from the Census 2000 testing cycle.

- Be Counted questionnaires, unaddressed forms requesting all short form items, plus a few additional items were available in public locations for people who believed they were not otherwise counted.

- An introductory letter was sent to all mailout/mailback addresses and many addresses in update/leave areas prior to the mailing of the census form. A reminder postcard was also sent to these addresses.

- Forms in Spanish or other languages were mailed to those who requested them by returning the introductory letter.
- A well-publicized, toll-free telephone number was available to answer questions about the forms. Also, responses of households who had received a short form could be taken over the phone.
- Under the Local Update of Census Addresses (LUCA) program, many local governments had the opportunity to address specific concerns about the accuracy and completeness of the Master Address File before mailings began.

Resolving Multiple Responses

With multiple ways for people to initiate their enumeration, as well as the field follow-up operations, it was very likely that some people would be enumerated more than once. A special computer process was implemented to control the extent of this type of nonsampling error by resolving situations where more than one form was received from an address. The process consisted of several steps. Addresses that had more than one viable return were analyzed. Housing data from one form were chosen as the housing data to use in subsequent census processing. Within each of these addresses, comparisons of the person records on each return were made against the person records on the other returns at the same address. People found to have been included on two or more different returns were marked as such, and only one of the person records was used in subsequent processing.

IMPUTING HOUSING UNIT STATUS AND POPULATION COUNTS

Following the completion of all data collection activities for Census 2000, a computer file of census housing units was created. For some housing units, information about whether the housing unit was occupied, vacant, or nonexistent was not available. These housing units were defined as "unclassified." Unclassified housing units were assigned a housing unit status of occupied, vacant, or nonexistent by assigning the status of a nearby housing unit to the unclassified unit. Additionally, the number of persons living in some housing units known to be occupied was unknown. Housing units with unknown population were assigned the population count of a nearby occupied housing unit. All other data for these housing units was assigned via substitution or allocation during the editing of unacceptable data described in the next section.

EDITING OF UNACCEPTABLE DATA

The objective of the processing operation was to produce a set of data that describes the population as accurately and clearly as possible. In a major change from past practice, the information on Census 2000 questionnaires generally was not edited during field data collection nor during data capture operations for consistency, completeness, and acceptability. Enumerator-filled questionnaires were reviewed by census crew leaders and local office clerks for adherence to specified procedures. No clerical review of mail return questionnaires was done to ensure that the information on the form could be data captured, nor were households contacted as in previous censuses to collect data that were missing from census returns.

Most census questionnaires received by mail from respondents as well as those filled by enumerators were processed through a new contractor-built image scanning system that used optical mark and character recognition to convert the responses into computer files. The optical character recognition, or OCR, process used several pattern and context checks to estimate accuracy thresholds for each write-in field. The system also used "soft edits" on most interpreted numeric write-in responses to decide whether the field values read by the machine interpretation were acceptable. If the value read had a lower than acceptable accuracy threshold or was outside of the soft edit range, the image of the item was displayed to a keyer, who then entered the response.

To control the creation of possibly erroneous people from questionnaires completed incorrectly or containing stray marks, an edit on the number of people indicated on each mail return and enumerator-filled questionnaire was implemented as part of the data capture system. Failure of this edit resulted in the review of the questionnaire image at a workstation by an operator, that identified erroneous person records and corrected OCR interpretation errors in the population count field.

At Census Bureau headquarters, the mail response data records were subjected to a computer edit that identified households exhibiting a possible coverage problem and those with more than six household members—the maximum number of persons who could be enumerated on a mail questionnaire. Attempts were made to contact these households on the telephone to correct the count inconsistency and to collect the census data for those people for whom there was no room on the questionnaire.

Incomplete or inconsistent information on the questionnaire data records was assigned acceptable values using imputation procedures during the final automated edit of the collected data. Imputations, or computer assignments of acceptable codes in place of unacceptable entries or blanks, are needed most often when an entry for a given item is lacking or when the information reported for a person on that item is inconsistent with other information for that person. This process is known as allocation. As in previous censuses, the general procedure for changing unacceptable entries was to assign an entry for a person that was consistent with entries for persons with similar characteristics. The assignment of acceptable codes in place of blanks or unacceptable entries enhances the usefulness of the data. Allocation rates for census items are made available with the published census data.

Another way corrections were made during the computer editing process was through substitution; that is, the assignment of a full set of characteristics for people in a household. When there was an indication that a household was occupied by a specified number of people, but the questionnaire contained no information for the people within the household or the occupants were not listed on the questionnaire, a previously accepted household of the same size was selected as a substitute, and the full set of characteristics for the substitute was duplicated. Housing characteristics are not substituted. Matrix H18, Occupied Housing Units Substituted, represents a count of occupied housing units into which all persons have been substituted.