Master Trace Sample Database - A Unique Opportunity for Census 2000 Research

Joan M. Hill, Jason D. Machowski
Planning, Research, and Evaluation Division, U.S. Census Bureau[1]

Abstract

The Master Trace Sample Database merges Census 2000 data from multiple sources to provide information about cases in various phases of data collection and processing. The objective of this effort is to support future methodological and operational analyses and decisions regarding the 2010 short form census. Over the last five years, the Census Bureau's Planning, Research, and Evaluation Division staff has worked with subject-matter and database experts to create a complex, relational database for research purposes. This prototype database merges Census 2000 address frame, collection, enumeration, capture, processing, response, and coverage files. These linkages allow quantitative insight into the relationship of key census processes through multi-variate analyses. In addition to being an innovative research tool, the Census 2000 Master Trace Sample Database is intended to serve as a model upon which we will improve in future censuses. This paper provides a detailed description of the content and structure of the database, as well as the motivations for the project, milestones in the creation process, description of special features, examples of recent applications, and discussion of limitations. In addition, the paper includes highlights of the results of a comprehensive database project assessment which was conducted to assist with future planning.

Keywords: Census 2000, research database, response data, enumerator

1. Background

In 1999, the National Research Council/National Academy of Sciences expressed renewed interest in the creation of a database that would allow users to conduct a broad range of analyses on the quality of decennial census content (National Research Council, 1999). The database they envisioned would contain a sample of census records selected in a way that users had the option of analyzing data at the block or interviewer level. In addition, the database would enable users to examine respondent data at various phases of data processing. This tool would be a resource throughout the decade as fresh ideas for examining data quality and the census process arose.

Census managers with experience from previous censuses also encouraged the idea of building a "trace" database using Census 2000 data. The possibilities of what census researchers could learn while planning the next census were appealing. The Census Bureau had attempted to create such a database in the previous decade based on a similar 1988 Council recommendation, but failed to do so due to operational difficulties and budget constraints. From 1999 to 2004, Census Bureau staff developed a research database called the Master Trace Sample (MTS) database.

2. Objectives

The MTS database merges Census 2000 data from multiple sources to provide information about cases in the various phases of data collection and processing. The objective of this effort is to support future methodological and operational analyses and decisions regarding the 2010 short form census. In the five years of development, the Census Bureau staff worked with subject-matter and database experts to create a complex, relational database for research purposes. This prototype database merges information from Census 2000 address frame, data collection, enumeration, data capture, data processing, and questionnaire response files. This merge yields a rich database which allows quantitative insight into the relationship of key census processes (for example, how information travels from data collection through data processing). Now, internal Census Bureau researchers are able to conduct explorations of operational and methodological issues aiming to benefit planners of the next decennial census that would not otherwise be possible. In addition to being an innovative research tool, the Census 2000 Master Trace Sample database is intended to serve as a model upon which we will improve in future censuses.

3. Research Applications

[1] This report is released to inform interested parties of research and to encourage discussion. The views expressed on statistical, methodological, technical, or operational issues are those of the authors and not necessarily those of the U.S. Census Bureau.

The MTS database contains a sample of Census 2000 housing unit records that allow Census Bureau researchers to trace questionnaire response (that is, data provided by respondents in Census 2000) and operational data, such as housing unit person count or field action codes, through stages of Census 2000 processing. These stages include address list development, data collection, data capture, and data processing. For the sample of housing unit records, the database contains all returns[2], which include housing unit level and person data.

The MTS database also contains data not typically analyzed in census evaluations. For example, data on each enumerator visit to a housing unit during the Nonresponse Followup (NRFU) operation (e.g., interview mode and outcome) are contained in this database. In addition, this database links micro-level data, such as enumerator production data and questionnaire response data, which are not traditionally linked in census evaluations.

One example of a research application involves determining the effectiveness of multiple NRFU contact attempts by an enumerator on Census response and data quality. At the Office of Management and Budget's request, Census Bureau staff conducted an analysis using the MTS database to determine if we could feasibly use fewer than six contact attempts to collect data from NRFU households without impacting Census response or increasing questionnaire item nonresponse. Since the NRFU is one of the most expensive operations in a decennial census, the number of data collection contacts that enumerators attempt for non-responding housing units is one area that could be examined in an effort to reduce costs. Census enumerators are trained to make up to six contact attempts to collect data at housing units from which no census questionnaires were received. MTS data were used to determine if field staff could feasibly use fewer than six attempts to collect data without impacting response or the quality of the data (Tancreto *et. al.*, 2004).

In another research study, Census Bureau staff conducted an analysis using the MTS database to determine whether the housing unit status (e.g., vacant) assigned by NRFU enumerators was consistently verified by enumerators during a supplemental field verification called Coverage Improvement Followup (CIFU). In Census 2000, almost a quarter of the housing units in the NRFU caseload were determined to be "vacant" and about one in seven had a status of "delete." Housing unit status was considered "questionable" if, for example, the status during NRFU was "vacant" and the status during CIFU was "occupied." Note that these operations used Census Day (April 1, 2000) as the reference date. The potential for mis-classification of housing unit status by enumerators is a concern among census data users. The goal of this study was to determine what enumerator and other characteristics are associated with discrepant housing unit status assignment during NRFU (Bentley *et. al.*, 2005).

Additionally, the Census Bureau's Population Division experts with Census Bureau support used the MTS database to determine whether proxy respondents caused erroneous clustering or "heaping" in the age distribution in Census 2000. If birth date is not provided, enumerators or other proxy respondents may tend to pick an age such as 25 or 50 when "guessing" the ages of non-respondents. This study used MTS data to test the hypothesis that heaping on ages with digits zero and five stems primarily from proxy reporting. The results are especially important since age heaping in the census age distribution results in error that will be carried forward for a decade, and every year there will be an overstatement of some cohorts and an understatement of others. Age heaping on ages ending in zero and five in 2000 would affect ages with end digits of one and six in 2001, of two and seven in 2002 and so on. Age heaping may always be present in the data, but it is possible to control and minimize the error through edit and quality assurance procedures. To address solutions to fix the problem, the cause needs to be known (West *et. al.*, 2004).

In general, the MTS database facilitates research on relationships among Census 2000 data and operations beyond the current Census 2000 Testing, Experimentation, and Evaluation Program.

4. Database Development

While developing the database, staff met with upper management in an attempt to meet the expectations of the Census Bureau and stakeholders (such as the National Academy of Sciences). Staff also met with Census subject-matter experts to identify their research interests. Based on these broad objectives and topics of interest, the MTS team developed a series of research requirements which were designed to cover a variety of research topics involving the relationship of key census processes. These research requirements were the basis for the design of the database. The requirements were used by the system designers, as well as software and

[2] A return is a response to the Census, such as a mail return short form, an enumerator-filled form, or a response collected via the Internet or telephone. More than one return can be associated with a given housing unit ID.

hardware experts, to develop the logical and physical database schema.

Upon completion of the database design, data loading, and content processing, the MTS database team implemented intensive system testing. This involved a technical quality review of each variable and linkage in the database, as well as comparisons to known data benchmarks and estimates. This was followed by acceptance testing which was conducted to ensure that the final MTS database satisfied the functional and research requirements. Additionally, an internal web page was created to assist users and provide information to managers regarding the MTS purpose, inputs, content, and analysis examples. A user interface was also developed but not included as part of the MTS database product due to access and other issues.

The final MTS product is an Oracle database which can be accessed using SAS® or SQL®, in addition to other tools. The MTS database contains over 25 data tables containing Census 2000 variables and recodes, with supporting indexes and scripts to improve efficiency.

5. Intended Uses and Targeted Users

With the wealth of data included in the MTS database, there is great potential for research in the following areas:

1. Modeling to identify and measure associations and relationships;
2. Tracing items, such as population count, through census processes; and
3. Investigating how to develop improved trace databases in future censuses.

For these research applications, the Census Bureau researcher can query the database or produce an extract for further analysis in statistical or database software.

The database is intended to be used to research hypotheses that involve relationships of various Census 2000 operations or systems. The MTS database is not intended to produce official totals or point estimates involving a single source file or program. The MTS estimates of descriptive statistics contain a level of uncertainty (that is, sampling error). However, descriptive results from single operations as reported for the full Census do not contain sampling error. Consult the Census Bureau's website for information on Census operations as reported in the Census 2000 Testing, Experimentation, and Evaluation Program[3]

The MTS database access is limited to internal Census Bureau use, in part, to protect confidentiality. Census Bureau researchers interested in pursuing studies that will help guide the planning of the 2010 short form census can develop research proposals for review and approval by senior staff, as well as planning groups guiding 2010 Census research.

6. Sample Design

The database contains a sample of approximately 1.5 million housing unit identifiers (IDs). A sample was chosen rather than the full Census 2000 data to make development easier for the prototype and reduce run time for the user. The sample is composed of two components: a Census 2000 housing unit ID systematic sample (over 600 thousand IDs) and a sample of block clusters (over 800 thousand IDs). Block clusters are a grouping of Census 2000 blocks which were created to support the Accuracy and Coverage Evaluation (A.C.E.). The block cluster sample contains all housing unit IDs within selected block clusters.

The ID sample was selected for general research and modeling applications and represents the nation. The Block Cluster sample was selected to have geographically contiguous areas in the sample to facilitate the study of effects at the block or interviewer level, as recommended by the National Academy of Sciences' 2010 Panel on Research on Future Census Methods. Since the ID and Block Cluster samples were selected for different purposes and each weight up to the national level, they are never used in combination.

The initial ID sample is a systematic sample of housing unit addresses on the decennial address frame file as of January 2000 contained in mailback type of enumeration areas, including Puerto Rico. The sampling rate is 0.5 percent. A supplemental ID sample was selected from the mailback housing unit updates after January 2000, as well as a sample of housing unit IDs in the remaining enumeration areas. In this way, housing units that were added later in the census process were given a non-zero probability of selection.

For the block cluster component, we selected a sample

[3] For more information on the Census 2000 Testing, Experimentation, and Evaluation Program, go to the following website: http://www.census.gov/pred/www/

of the block clusters that were listed for the A.C.E.[4] All housing unit IDs contained in the sample block clusters at the time of the January 2000 selection, and any housing unit IDs added to these block clusters prior to the creation of the final decennial address frame file, were included in the MTS database. This sample included both block clusters selected for A.C.E. (approximately three-quarters of the MTS block cluster sample) and not selected for the A.C.E. The block cluster probability of selection was proportional to the total number of housing units in the block cluster as of January 2000. Note that block clusters containing zero housing units were separately sampled at random. Thus, these block clusters, which were empty in early 2000, had a non-zero probability of selection so that housing unit growth in these areas was represented.

For each of the ID and Block Cluster samples, a subsample of housing unit IDs and block clusters was designated to have questionnaire images retained.

7. Database Content

The MTS database links micro-level data from various stages of the Census 2000 process such as address frame development, data collection, data capture, data processing, and enumeration contact records. To facilitate research, data are linked at the following levels:

- Local Census Office (LCO),
- enumerator,
- housing unit,
- return (that is, census questionnaire),
- enumeration contact (personal visit or telephone interview), and
- person.

The MTS database is intended to address a wide variety of research requests that link decennial census questionnaire response, data collection, and processing information with enumeration characteristics. In general, the MTS database contains the following types of data:

- geography;

- census questionnaire response data at various stages of processing;

- enumeration characteristics (related to operations and enumerators);

- record of contact information from the NRFU and CIFU forms;

- data capture system evaluation information from a reconciled keyed-from-image data set;

- geographic coding error results from one of the Census 2000 evaluations; and

- Census Day housing unit status data (occupied/vacant/delete/unresolved) from NRFU, CIFU and the A.C.E.

In addition to geographic control data (e.g., state, county, and block) and geographic coding error results from address updates, the MTS database contains a substantial variety of information on data collection operations and enumerators. Operational data ranges from flags for interviews conducted using Census 2000 operation close out procedures to the time, date, mode, and outcome of each attempt made by an enumerator to contact a respondent. The database is also rich in information related to the enumerator such as cases completed per hour, number of days worked in a given field operation, full/part time work schedule, application test scores, educational attainment, ability to speak a language other than English, and previous computer experience.

Although housing unit status as measured in A.C.E. is included, the MTS database does not have Census 2000 person or housing unit coverage data from A.C.E. Coverage data represent cases that should have been added to or excluded from the Census based on the A.C.E. Since the intent of the database is to trace existing units/records through the census process, coverage data are not included.

The MTS database includes housing unit data but excludes data on group quarters. Millions of people in the United States live in group situations, collectively known as group quarters (e.g., nursing homes), although the vast majority of United States residents live in housing units. Unique operations were required in Census 2000 to compile the list of Group Quarters and unique enumeration activities were required to include residents of group quarters (Abramson, 2003). The MTS team initially considered including group quarters

[4] This level of geography was created for the A.C.E. sampling frame to form primary selection units of comparable size for enumerator workloads. Note that a portion of the MTS sampled block clusters was not involved with the A.C.E. survey work. The MTS used the A.C.E sampling frame which contained both A.C.E. and non-A.C.E. block clusters.

in the Census 2000 MTS database. However, to control the complexity of the task and reduce the number/types of sources to a more manageable level, the team decided to focus on the housing unit sample for the first prototype.

The questionnaire response data included in the MTS database are "100 percent" item data and do not include "sample data." The 100 percent data are the data collected on all Census 2000 questionnaires, such as sex, age, race, and Hispanic origin. The sample data refer to items that appear only on the Census 2000 long form, such as educational attainment and income.

8. Database Constraints

The following constraints apply to the overall MTS database:

- Any limitations present in the original Census 2000 files also are present in the MTS database, which contains data from numerous Census 2000 source files. Although the variables from these files went through testing to ensure that the data were properly extracted and merged, the values were not edited. [5]

- The database does not contain a comprehensive set of Census 2000 files. For example, the MTS database does not include the various Local Update of Census Addresses files or the preliminary versions of the Decennial Response File. The sources of data in the MTS database are intended to represent the major data collection, capture, and processing steps for Census 2000.

- Enumerator characteristic and production data have limitations. The association of enumerator data to a particular case is limited to the last enumerator who worked on the case. That enumerator is not necessarily associated with the full contact history of the case, if the case was worked by more than one enumerator. Another important limitation of these data is that they are "as reported" and "as keyed." Many of the limitations associated with enumerator files stem from the fact that the primary objective of these files was not evaluation or research needs, but rather real time information for operational monitoring and tracking.

- While the MTS team attempted to design a database to handle a variety of issues, the database cannot address every conceivable research question. For instance, Census Bureau researchers cannot trace every key measure through the census and similarly Census Bureau researchers cannot trace any one measure through all stages of data collection and processing. A set of 15 research requirements guided the design of the MTS database. This set represents anticipated research areas of high interest based on project goals and objectives.

9. Formal Assessment

After completion of the MTS database and associated products, a formal evaluation was implemented to assess both the usefulness of the database for research and the benefits to the Census Bureau of resulting products. This qualitative analysis assessed the development process, final product, and constraints, as well as the utility of the internal web documentation and Census Bureau researcher/customer satisfaction. Through interviews with staff and stakeholders closely involved in the MTS database project, the analysts identified several aspects of the development process that were problematic such as vague requirements, planning difficulty, schedule delays, competing staff priorities, and low customer usage due, in part, to limited database access.

One of the primary recommendations was to begin the project earlier in the decennial census cycle to improve planning and timing. This includes involving upper management/executive staff throughout the project to ensure a common understanding of the vision and scope of the database and its users. Another recommendation suggested integrating additional quality assurance earlier in the development process to prevent unexpected error in the later stages.

[5] One exception is the record of contact data. These data were edited based on expert review of the enumerator records.

This information is designed to guide the administration and implementation of any future trace database projects (Benton *et. al.*, 2005).

10. Additional Recommendations for Future Database Projects

In addition to recommendations gathered at the end of the project through the formal assessment, the MTS team received valuable input and suggestions from subject-matter and database experts throughout the database development process. Some of these ideas could not be incorporated into the prototype because of complexity and resource/timing limitations. These suggestions may greatly improve the usefulness of the MTS database and should be considered when designing future applications. The primary recommendations are provided below.

- The MTS could be expanded to include data on group quarters. Adding operational, questionnaire response, and evaluation data associated with group quarters to future database applications may prove valuable for planning the next decennial census.
- Expanding the next MTS database to include coverage measurement data associated with persons would provide an additional evaluation measure with regard to within household coverage. The A.C.E. final Census Day housing unit status is the only Census 2000 A.C.E. data included in the Census 2000 MTS database prototype.
- The MTS concept of linking valuable operational and questionnaire data into a comprehensive database for a specific survey or census, such as the American Community Survey or the Economic Census, may prove useful. The Census 2000 MTS database links enumerator data to questionnaire response data and quality measures for Census 2000. In the census, as in other major surveys, these key variables are traditionally located in separate systems or files and are not joined. Provided the Census 2000 Master Trace Sample database proves to be useful in planning the next decennial census, the Census Bureau may wish to consider building a 'trace' database specific to each of its major surveys, as well as the Economic Censuses.

Acknowledgments

The authors would like to thank and acknowledge the extraordinary contributions of the core Master Trace Sample Team: Mike Bentley, Julie Bouffard, Claude Jackson Jr., Lance Moore, Quang Phan, Cindy Rothhaas, Courtney Stapleton, George Train III., Frank Vitrano, and Henry Woltman. Special thanks to management, contractors, and the many individuals throughout the Census Bureau who made this project possible.

References

Abramson, F. (2003) "Special Place/Group Quarters Enumeration", Census 2000 Topic Report Number 5.

Bentley, M. and Tancreto, J. (2005) "Questionable Housing Unit Identification Analysis", Internal Census Bureau Draft Memorandum, April 21, 2005.

Benton, H., Nelson, L., and Tancreto, J. (2005) "Master Trace Sample Database Assessment", Internal Census Bureau Memorandum for Edison Gore, April 6, 2005.

National Research Council (1999) *Measuring A Changing Nation: Modern Methods for the 2000 Census.* National Academy Press, Washington, D.C.

Tancreto, J. and Bentley, M. (2004) "Enumerator Contact Study 2000", Internal Census Bureau Memorandum to Teresa Angueira, November 2, 2004.

West, K., Robinson, J.G., and Bentley, M. (2004) "Did Proxy Respondents Cause Age Heaping in the Census 2000?", Internal Census Bureau Memorandum for Teresa Angueira, December 6, 2004.