Designing an Adaptable Database for Model-Based Research

Nancy Johnson and Mary Pritts
U.S. Census Bureau[1], 4600 Silver Hill Road, Washington, DC 20233

Abstract
In preparation for the 2020 Census, the U.S. Census Bureau is conducting research on building statistical, model-based simulations for a selective field canvassing operation and for producing estimates of coverage error. This type of research requires a flexible and adaptable database structure that incorporates data from various sources, allowing for different levels of analysis, across different time periods, in an ad hoc manner. The starting point for the database is the universe of over 155 million records from the 2010 Census Address Canvassing operation. It then incorporates data from the final 2010 Census and various Administrative Record sources covering different spans of time. The result is a multi-use, multi-source data set fully integrated at the level of analysis most appropriate for the research project.

Key Words: Census, Database, Address Canvassing, Administrative Records, Master Address File

1. Introduction

In this paper, we describe the data and database design used by Census Bureau statisticians performing the statistical modeling described in Young and Johnson (2014), Boies and Tomaszewski (2014), and Tomaszewski and Boies (2014). The statistical models being created will be used to predict coverage errors in the census address list and to predict what blocks will require field canvassing to improve the address list before the next census. A major goal for these projects is to save money in the next census by using a selective address canvassing approach to update the census address list. By predicting which areas of the country have the most change, we can choose where to send field representatives to update the address list. The field operation designed to update the address list before the census is known as Address Canvassing (AC). For more background on selective AC and some previous research results, see Boies (2012) and Tomaszewski and Shaw (2013).

The database constructed to support the statistical model development and evaluation consists of various data sources including multiple vintages of the Master Address File (MAF), 2000 and 2010 Census files, and geographic files. Currently, work is underway to incorporate data from a number of Administrative Record (AR) sources.

Some of the challenges we faced in developing our database include

[1] *The views expressed as those of the authors and not necessarily of the U.S. Census Bureau.*

- Records in some data sources are at the person level, some at the housing unit or group quarters[2] level, and some at the census block level.
- Some census blocks do not have living quarters.
- The geographic identifiers are not consistent across all files.
- Files from a single data source can come from multiple time periods, and each data source uses different time periods.

As our database was being created, various solutions were considered to address these challenges. Some limitations were also introduced to our database. More details on these issues are in the following pages.

2. Adaptable Database

To research address coverage error, statistical models are being used to estimate the coverage error in the address list at levels of geography down to the block level. To research what blocks will require field canvassing, other statistical models are being used to predict the probability of new addresses in a block. Because the MAF is the source of addresses used for the Census Bureau, it was the universe used for both modeling efforts. We started by creating a file consisting of records from the 2009 Pre-AC files as well as results from the 2010 AC file. This file was merged with other MAF extracts, as well as other data sources, including the decennial censuses. Our final file has many different vintages of data, and various variables were created during the database creation process to identify the vintage and source of the data.

As we created our database, we found that not every data source available to us was useable. One data source with promising information was the 2010 Census planning database. The planning database is a Census database with a range of housing, demographic, socioeconomic, and census operation data. The variables come from the 2010 Census and the 2006-2010 American Community Survey (ACS). Unfortunately, the data are at the block-group level of geography, which was too homogenous to be useful for our modeling efforts. The data sources currently in our database are all at the person, address, or block level.

The records we included in our database contain all addresses identified at some point as living quarters. This includes group quarters, as well as units that may have changed from a residential to nonresidential status. Addresses from the entire U.S., including Alaska, Hawaii, and Puerto Rico are in our files.

3. Data Sources

The primary data sources in the adaptable database design include:

- Master Address File
  - 2009 Pre-Address Canvassing Extract
- 2010 Address Canvassing Results
- Decennial Census Information from the 2000 and 2010 Census
- Geographic Files

[2] Group quarters are places where people live in a group arrangement managed by an organization, e.g., college dormitories, military barracks, and prisons.

- Administrative Records Files
- Field Test Results

Table 1 shows the vintages of the data sources we have already incorporated into our database and the number of records that came from each source.

Table 1: Number of Records from the Data Sources

	Count of Address Records
2009 MAF	181.8 mil
2013 MAF	195.5 mil
2009 Pre-AC Extract	144.9 mil
2010 AC Operations Results	163.6 mil
2000 Decennial Census	117.3 mil
2010 Decennial Census	133.3 mil
	Count of Block Records
2009 Geographic Files	11.2 mil
2014 Geographic Files	11.2 mil

The data sources are described below.

3.1 Master Address File

As the MAF is the source of all the Census address records, it is the natural universe for our studies. The MAF was developed by the Census Bureau to be a comprehensive, nationwide address list of every living quarters in the United States and its territories. The MAF is used as a frame for the decennial census activities as well as for many demographic and economic surveys.

As a comprehensive list of all living quarters, the MAF includes both currently occupied and vacant units. It includes multi-units such as apartment buildings, transitory locations such as trailer parks, and group quarters such as prisons. It also includes some nonresidential addresses, including addresses for businesses, schools, and churches. In addition, it contains addresses that were deleted in previous Census Bureau operations. These deleted address records are not removed from the MAF, but rather flagged as invalid.

The information on the MAF comes from a variety of sources. The primary source used to maintain the file over time is the United States Postal Service (USPS) Delivery Sequence File (DSF). This file includes every mail delivery point recognized by the USPS. It is updated regularly, and the Census Bureau receives these DSF updates semiannually. The DSF is a good source of “city style addresses”, that is, addresses with

street names and house numbers. U.S. Census Bureau operations help to provide addresses for noncity-style cases (e.g., addresses with a post office box or rural route). During some census field operations, field staff list addresses in areas determined to have many noncity-style addresses. Additionally, the ACS and the Demographic Area Address Listing (DAAL) both provide updates to the MAF throughout the decade, in selected locations. Finally, many units of local government also contribute addresses through the Local Update of Census Addresses (LUCA) program.

Since the 2000 Census, the MAF has been updated twice per year with information from the DSF. In addition, census field operations are constantly producing address updates that are incorporated into the MAF. In particular, AC took place before the 2010 Census to ensure the MAF and maps were as up-to-date as possible. During AC, the MAF's inventory was verified and addresses were added, corrected, or flagged as deletes as necessary. (Bainter, 2008).

3.2 Pre-Address Canvassing Extract

The Pre-AC extract of the MAF contains the initial universe of addresses that were considered potentially valid at the time of the 2010 AC operation. The addresses for the extract are selected from the MAF based on a set of rules called a filter. The filter excludes some records on the MAF that do not meet certain address requirements. For example, one category of addresses excluded from the Pre-AC extract is ungeocoded addresses (addresses for which we do not have a census block code). Some records that are included in the extract are addresses that were enumerated in Census 2000 and have not been deleted by subsequent operations. Addresses added by the DSF after Census 2000 are also included in the extract. The 2009 Pre-AC filter resulted in a total of approximately 144.9 million addresses (U.S. Census Bureau, 2012) and 6.6 million tabulation blocks.

3.3 Address Canvassing Results

The file with the 2010 AC results contains a record for every address that came back from the field operation with an action code. This includes every address that was on the 2009 Pre-AC list, as well as additional units that were found during the field operation. These adds consisted of addresses completely new to the MAF (true adds) and adds that matched to an address already on the MAF that had been excluded based on the filter (reinstated adds). Reinstated adds are mostly made up of addresses that were ungeocoded before AC. The action codes from AC, in particular the adds and deletes, were used to create the dependent variables in both modeling efforts. The total number of addresses with action codes from AC was 163.6 million for the U.S. and PR in this file.

3.4 Decennial Census Files

The Census Bureau has a series of files for the U.S. and Puerto Rico containing the results from the 2000 Census and the 2010 Census enumerations. These files have demographic characteristics such as age, sex, and ethnicity for each person residing in living quarters. To incorporate the person-level characteristics into our model database, we collapsed values into categories where appropriate and summed to the living quarter level in order to create an address-level file. Then, we summed and averaged these to the 2010 tabulation block level. The 2000 Census file contained 117.3 million living quarters and the 2010 Census file contained 133.3 million (Mazur and Wilson, 2011).

3.5 Geographic Files

Several files, both internal and external, contain characteristics of the 2010 tabulation blocks. Examples of some characteristics include Bureau of Land Management (BLM) areas, block distance from urban areas, distance from university landmark features, and National Land Cover (such as an area covered by forest or areas that are developed). We are particularly interested in these files as a source of information for empty blocks. Empty blocks have been excluded from the main statistical models because our independent variables largely depend on information from the known housing units and persons residing in the block. We are currently planning to create a separate empty block model which would use the block's geographic information to predict how likely there are to be adds in the block. In addition to contributing to an empty block model, the files will tell us which blocks are likely to never have housing units (e.g., blocks that are very small, are 100% BLM owned, or are very mountainous) and can therefore be dropped out of all our models.

3.6 Administrative Records Files

In addition to the DSF, the Census Bureau has obtained several AR files from commercial and federal sources. Administrative records, by definition, are not collected for statistical use, but are normally used for housekeeping purposes. Therefore, they inherently present us with several challenges when incorporating them into our modeling universe. These challenges generally fall into four categories: coverage, content, spatial assignment, and record linkage.

Coverage issues include incomplete coverage of the population of interest and of the geographic area within the universe. Often, the administrative files will come from individual states, and we do not have the file for every state. Other files that are for very specific programs, such as Housing and Urban Development, will only have coverage of their small subgroup of relevant blocks.

Content issues with the AR files refer to the tendency of the files to contain variables that are not directly suitable for our database, inconsistencies in the variable definitions across the files, inaccuracies, incompleteness, and duplication of data items, and missing time references. Additional time was needed to review data dictionaries, recode variables into more meaningful categories, and find and remove duplicate addresses.

Spatial assignment issues relate to the geo-spatial characteristics of administrative records. In the AR files, it is hard to know how accurate and current the address information is. It is possible for some records to have recently updated information while other records may be very outdated. Accurate address and spatial information is needed so each record can be correctly geocoded to a census block. Using misgeocoded records in our models could affect our results, while ungeocoded records cannot be used in our block-level models at all.

Finally, the accurate linkage of records across different data sets is a significant challenge, especially when data are inconsistent, missing or erroneous. Other data sets directly created by the Census Bureau are automatically given unique identification numbers, called MAFIDs. The process of matching addresses in order to link records from the MAF to AR files requires additional programming, and often additional duplicates would be created which have to be found and removed.

We believe that using AR files in our statistical models has the potential to be very

fruitful; however, these challenges require additional processing efforts. Therefore, for modeling we have only started exploring the use of these files.

3.7 Field Test Files

In the fall of 2014, we will conduct a field test designed to provide "ground truth" data to help validate the statistical models. The field test will mimic the 2010 AC operation and will provide action codes for every address in our sample. The field test sample was selected from the July 2013 ACS MAF extract in time to meet the schedule for files going out to the field. The selected sample includes 10,000 blocks with known housing units and has an estimated listing workload of 1,037,363 addresses. One hundred additional blocks with zero known housing units will also be included, for a total of 10,100 blocks. (Pritts and Snodgrass, 2014).

4. Data Incorporation Methods

To deal with some of the development challenges and to define a universe of records flexible enough for research and analysis, we integrated a set of data files that resembles a "string of pearls." Figure 1 illustrates this design.

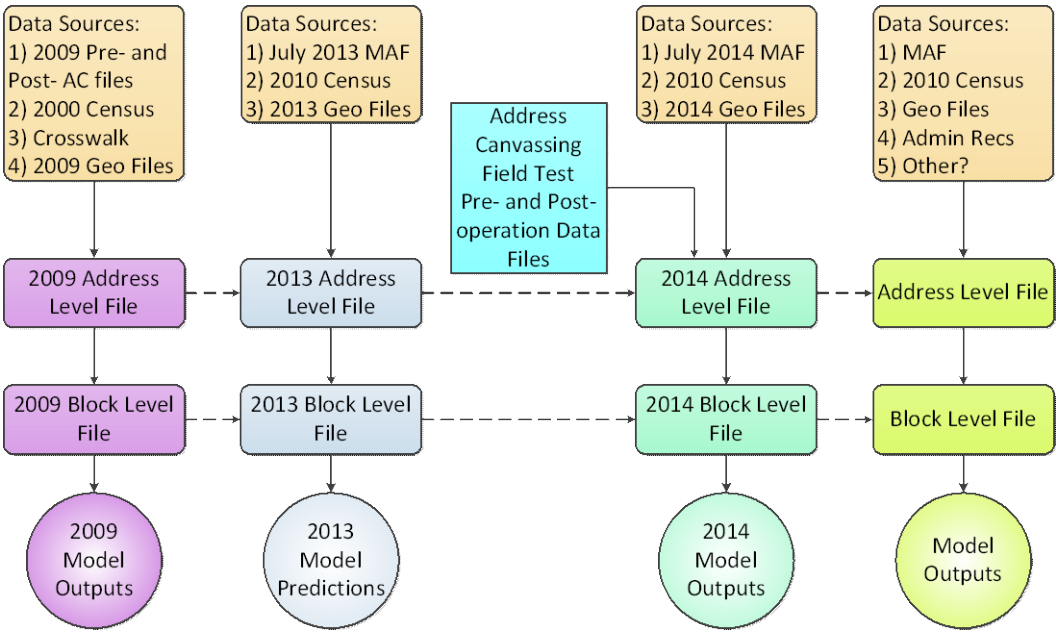

Figure 1: Adaptable Database Design Concept

Figure 1 demonstrates how the different data sources become incorporated into the database to be used later for the statistical models. The initial file we created consisted of address records from the 2009 pre-AC files as well as results from the 2010 AC operation. Variables from the 2000 Census, a crosswalk file for converting between different types of geography, the 2009 Geographic Reference File, and other geographic files were also included. In this address-level file, we recoded many of the variables (and values) into new dichotomous or dummy variables. To create a block level file, these dummy variables were summed and averaged at the 2010 tabulation block level. This file included all the blocks in the MAF before or after AC. Because it is possible for some blocks to be empty before AC and then to have living quarters found in them during AC,

and for other blocks to have living quarters in them before AC and then to have them be empty during AC, we included all the blocks that contained living quarters either before or after AC. For the other sets of files, we used the same process – we recoded many of the variables (and values) into new binary variables and summed and averaged these address records at the 2010 tabulation block level.

One of the challenges already mentioned is that the files come from different time periods. For example, the MAF is updated twice per year, but other files are updated yearly, and some only exist for one point in time. Because of this, we use only the appropriate variables for each statistical model. For the 2009 statistical models, we used the 2009 pre-AC information to predict the outcomes (action codes) in the 2010 Census AC operation. For the 2013 set of models, we do not have updated action codes so we do not have dependent variables for these sets of models. Instead, we will use the parameter estimates from the 2009 set of models with 2013 data to see what blocks are predicted to have certain action codes. For the 2014 set of models, we will use pre-2014 data to predict the results of the field test. Each set of files can be linked together by MAFID at the address level and by 2010 tabulation block ID at the block level, allowing us to carry potentially useful variables over from older files to newer ones. For each set of models, the independent variables are created from earlier data than the dependent variables.

To create our database, many data sources, each with a large number of records and variables, needed to be merged together. Working with such large data sets requires large amounts of disk space and significant CPU time. A couple of methods were used to deal with this in our SAS© programming. We used a hash object merge, which efficiently combines data sets by loading one of the datasets into memory; this is much faster than disk-based operations. Parallel processing was also incorporated into our programs by running independent parts of the programs concurrently.

An indication of the size of our database is given in Table 2, which shows the counts of addresses and blocks currently in our database. Not only are there millions of records in the files, there are over 1,000 variables in the address-level files and over 2,000 in the block-level files.

Table 2: Addresses and Blocks in the Current Database

	2009 Model Files	2013 Model Files
Addresses	188.0 mil	195.5 mil
Blocks	11.2 mil	11.2 mil
Blocks with Living Quarters	6.6 mil	6.5 mil
Empty Blocks	4.6 mil	4.7 mil

Another challenge in creating our database was that the files used two different geographic identifiers. Some data sources used collection geography and some used tabulation geography. Collection geography supports the management of field enumeration in the census (including AC), and consists of addresses, boundaries, and geographic features. Collection blocks are bounded by visible features and the boundaries do not change. Tabulation geography is used for legal, statistical, and administrative

areas and can be bounded by non-visible boundaries. The 2010 AC results file uses collection geography and the 2010 Census files use tabulation geography. We decided to use tabulation geography in our database because collection blocks are not used in other operations and tabulation blocks will continue to be used throughout the decade. Because of this, some data that was only available in collection blocks had to be mapped to the appropriate tabulation block by using a crosswalk file at the address level. We will keep this mapping until the 2020 Census geography is available.

5. Variables

To assess the coverage of the MAF and predict which areas require AC, we used different types of models with various dependent and independent variables. Our dependent variables consisted of various counts and indicators of AC action codes, depending on the model. Our independent variables consisted of various address and block characteristics, as well as demographic information.

5.1 Dependent Variables

The dependent variables are based on the action codes from AC. In AC, census field workers collect address data and compare it against the address list to make changes as needed. Every record on the address list (the Pre-AC MAF extract) was assigned an action code from AC. The following action codes were possible.

- Add – The living quarter was observed on the ground, but it was not on the address list, so it was added.
- Change – The address was corrected.
- Verify – The address had no changes.
- Duplicate – The address was a duplicate of another address in the AC universe.
- Move – The address was deleted by the lister and the same address was added in a different block (note: move actions are identified later in the update process; they were not a valid field action).
- Nonresidential – The address is for a commercial establishment or other nonresidential use, and there is no living quarters at the address.
- Uninhabitable – The living quarters must be vacant, open to the elements, condemned, or burnt out, and as a result, unfit for habitation.
- Single Delete – The address does not exist and the delete record was not verified by a second lister.
- Double Delete – The address does not exist and the delete record was verified by a second lister in quality control or in the Final Delete Verification.

Using these action codes, we have run models on predicted counts of adds in a block, using the two different types of adds (true and reinstated), as well as predicted counts of deletes (primarily double deletes), and predicted counts of all negative actions (which includes both types of deletes, duplicates, nonresidential, and uninhabitable). Other models have aimed at predicting the probability of a block containing one or more adds, or predicting the probability of it containing one or more adds or deletes. All the possible definitions of adds and deletes have been used in these models.

Other exploratory model analysis has also been done using the change and move action codes in our dependent variables; however, these models have not shown as much promise in providing answers to our research questions. The verify action code has not

been used in any models.

5.2 Independent Variables

The independent variables in the statistical models are selected from the remaining variables in our database. Our block-level files contain over 2,000 binary dummy variables and other categorical variables that are eligible for selection as independent variables in the statistical models. The variables can be classified into five main categories:

1. MAF Flags and Indicators

These can include information on the address source, e.g., if the address came from the Local Update of Census Addresses (LUCA), or if it was present in the 2000 Census, and so on. The address status (e.g., valid address, duplicate, nonresidential) as of various prior operations is also included. Finally, this category also includes descriptions of the address itself, such as whether the address is from a single unit or if it is a unit within a multi-unit structure, or a group quarters.

2. USPS Delivery Sequence File (DSF) Measures

These variables include information related to the DSF, such as if the address was a valid DSF address, or if it was excluded from delivery statistics (EDS). EDS addresses are flagged by the USPS because they are not current mail delivery points, though they could be in the future and therefore are left in their system. These variables also include measures of DSF change over time, e.g., the first time an address appeared on the DSF, and whether it remained in the DSF from then to the present.

3. Demographic Measures

Demographic measures come primarily from the decennial files, and include information on age, race, sex, and Hispanic origin. These variables were originally in the decennial files at the person level, so variables were created at the address level that give the total counts in an address for each demographic characteristic, e.g., the number of people 65 years old or older in the unit.

4. Geographic Characteristics

These variables include geographic information for the block. This includes whether the block is a land or water block, if it contains city-style addresses, or the percentage of the block that is covered by a national park landmark.

5. Administrative Records Coverage

These variables are still being developed. We hope to incorporate information from the AR files soon, and are planning to create variables such as the ratio of AR addresses to MAF addresses.

6. Conclusion

The statistical modeling efforts related to researching address coverage error and selective field canvassing requires a large database utilizing datasets from numerous

sources. These datasets cover the U.S. at the person, housing unit, and census block levels and contain longitudinal data. During design and construction, challenges in incorporating data sources at different levels, with different geographic identifiers, and at various time periods, were encountered and had to be resolved. Our approach creates a set of data files that is flexible enough to be used for different types of statistical models and evaluation methods.

References

Bainter, Steven (2008), "MAF Basics," http://www.geo.census.gov/apmb/mafbasics.html; April 1, 2014.

Boies, John (2012), "Final Report for the 2010 Census Evaluation of Address Canvassing Targeting and Cost Reduction," DSSD 2010 CPEX Memorandum Series, No. A-04, July 12, 2012.

Boies, John and Christine Tomaszewski (2014), "Fielding a Targeted Address Canvassing Operation: Alternative Approaches to Moving from Predictive Statistical Modeling to a Cost Effective Address Canvassing Field Operation for the 2020 Census," JSM Proceedings, 2014.

Mazur and Wilson (2011), 2010 Census Briefs, Housing Characteristics: 2010, Issued October, 2011.

Pritts, Mary and Sally Snodgrass (2014), "Customer Requirements Document for the Adjudication of DAAL Updates for the 2020 MAF Model Validation Test (MMVT)," DSSD 2020 Decennial Census Memorandum Series, No. ##, U.S. Census Bureau, Forthcoming.

Tomaszewski, Christine and John Boies (2014), "Recent Advancements in Statistical Modeling to Identify Address Updating Areas for the 2020 Census," JSM Proceedings, 2014.

Tomaszewski, Christine and Kevin Shaw (2013), "Examining Census 2000 Tabulation Block and Tract Homogeneity of Census 2010 Address Canvassing Action Codes, for 2020 Census Targeted Address Canvassing Modeling," DSSD 2020 Decennial Census R&T Memorandum Series, No. R-03, May 21, 2013.

U.S. Census Bureau (2012), 2010 Census Address Canvassing Operational Assessment Report No. 168, January 17, 2012.

Young, Derek and Nancy Johnson (2014), "Zero-Inflated Regression Modeling for Coverage Errors of the Master Address File," JSM Proceedings, 2014.