Geography Division Address Canvassing Recommendation

Geography Division:
Timothy F. Trainor

Version 1.01

Nov. 15, 2014

Table of Contents

List of Tables ii
List of Figures ii
Executive Summary 1
1 Introduction 3
1.1 Purpose and Scope 3
1.2 Background 4
1.2.1 Address List Improvement for Decennial Censuses 4
1.2.2 The Geographic Support System Initiative 5
2 Assess and Improve MAF/TIGER 6
2.1 Quality Indicators 6
2.1.1 Examples of QIs Applied with Other Data 7
2.2 Partnership Program 7
2.2.1 Description 7
2.2.2 Method 8
2.2.3 Results 8
2.2.4 Discussion 12
2.3 Third-Party Address Data and Administrative Records 14
2.3.1 Third-Party Address Data 14
2.3.2 Administrative Records 15
2.3.3 Discussion 15
3 Research, Modeling, and Area Classification 15
3.1 Assessing Stability and Consistency 16
3.1.1 Delivery Sequence File Stability Index 16
3.1.2 Measuring Housing Unit and Address Consistency Across Datasets 21
3.2 Comparison of the MAF with Aerial Imagery 23
3.3 Statistical Models 24
3.4 Applying Address Canvassing Research 26
3.5 Analysis 29
4 Recommendations 31
4.1 Recommendation for a Reengineered Address Canvassing 31
4.1.1 Basis of the Recommendation 31
4.1.2 Other Information Applicable to the Recommendation 31
4.2 Further Recommendations to Support a Successful Reengineered Address Canvassing 32
5 Impacts 33
6 Benefits 33
7 Risks 34
8 Conclusion and Future Work 35
Document Logs 37
Abbreviations and Glossary 38
Acknowledgments 39

List of Tables

Table 1. Partner Type 10
Table 2. Reasons for Non Participation 10
Table 3. MAF Updates from Partner Address Data 11
Table 4. TIGER Updates from Partner Road Data 12
Table 5. Production Rate Comparison of GSS-I and Other Address Review Operations 13
Table 6. Impact on Ungeocoded Addresses After Processing Partner Address and Road Data: Areas Covered by Partners to Date 14
Table 7. Initial Match Rates of Third-Party Data and the MAF 15
Table 8. Number of Census Tracts and Addresses by DSF Stability Index Value: 2009-2012 ... 21
Table 9. Comparison of 2010 Census Housing Unit Counts with Spring 2012 DSF Address Counts, by Census Tract 23
Table 10. Address Canvassing Statistical Modeling Outcomes at Selected Housing Unit Canvassing Levels, using 2010 Census Blocks 25

List of Figures

Figure 1. GSS-I Timeline 5
Figure 2. Distribution of Entities Contacted, by Type 9
Figure 3. Partner Data Processing 10
Figure 4. Reengineered Address Canvassing Continuum 17
Figure 5. DSF Stability by Census Tract: 2010-2014 22
Figure 6. Reengineered Address Canvassing Decision Flow 27
Figure 7. Special Land Use Areas With Little Expected Development: 2014 28
Figure 8. Non-Mail Out Types of Enumeration Areas: 2010 30

Executive Summary

Address canvassing is the process by which the U.S. Census Bureau validates, corrects, or deletes existing Census Bureau addresses, adds missing addresses, and adds or corrects locations of specific addresses before a decennial census. The Census Bureau is determining if a full, nationwide address canvassing in the field can be reengineered for the 2020 Census. A reengineered address canvassing involves both in-office and in-field components to update and validate the address list in preparation for the 2020 Census. The Geography Division recommends that the Census Bureau move forward with plans for a reengineered address canvassing for the 2020 Census. The main findings from research and activities to improve the quality and coverage of the Master Address File/Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) System and to support a reengineered address canvassing are summarized below.

Address evaluation and update processes have been implemented that are intended to minimize the areas where in-field address canvassing would be necessary by ensuring a complete, comprehensive MAF/TIGER. These processes focus on acquiring address and road data from local governments and researching other sources to update and maintain MAF/TIGER. The Geographic Support System Initiative (GSS-I) Partnership Program has demonstrated that it can acquire, evaluate, and use data from local-government partners to supplement the twice-yearly United States Postal Service (USPS) Delivery Sequence File (DSF). Of the 30.5 million addresses from the GSS-I processed to date (covering 37.0 percent of the nation's housing units), 76.6 percent matched existing MAF addresses and of those, 97.7 percent were geocoded to the correct block, thereby confirming that the large majority of addresses in partner lists already are contained in the MAF and are candidates for inclusion in Census 2020. Also, local government files have utility in maintaining and updating the TIGER road network and, further, support geocoding new and existing MAF addresses that previously were not included in the Census. These results suggest that there are many addresses already in the MAF that can be independently and successfully validated using partner data, reducing the need to field validate all addresses in the MAF. Through processing the local-government partner data acquired to date, 64,183 new addresses were added and 327,799 previously ungeocoded MAF addresses were geocoded using fewer resources than in-field address canvassing and traditional in-office geocoding.

Feasibility studies are underway to determine whether third-party address lists are a viable source for updating MAF/TIGER and to determine the coverage of these in areas where partner files are not available. Initial results show that 82.5 percent of the addresses in a national third-party file match with the MAF, showing promise for validating existing addresses in the MAF. A large number of unclassifiable addresses suggest, however, that third-party files may be less efficient for identifying new addresses and addresses missing from the MAF.

Address canvassing research has demonstrated that (1) high levels of housing unit and address consistency exist throughout the United States; (2) comparison of MAF addresses and imagery for change detection can successfully highlight census blocks to be canvassed in the field; and (3) statistical modeling can contribute to the process for selecting geographic areas for review, leading to the identification of census blocks to be canvassed in the field. These activities impact and benefit not only the recommendation for a reengineered address canvassing, but also provide opportunities for updating MAF/TIGER throughout the decade, thereby improving address and feature coverage for intercensal surveys and estimates programs.

It is important to note that none of the methods employed offers a single solution, rather, each works with the others to create more certainty about areas most likely to require in-field address canvassing. While the research and results discussed in this paper are sufficient for developing a recommendation for a reengineered address canvassing, research must continue. Establishing the process for determining the percentage of housing units for in-office and for in-field canvassing will be followed by the development of an implementation plan identifying the optimal sequence of file acquisition, data analysis, and other activities for more effective decision-making and the geographical distribution of the specific areas to be canvassed in the field. This work will proceed in conjunction with the development of the 2020 Census design.

1 Introduction

Address canvassing is the process by which the U.S. Census Bureau validates, corrects, or deletes existing Census Bureau addresses, adds missing addresses, and adds or corrects locations of specific addresses before a decennial census. The Census Bureau is determining if a full, nationwide address canvassing in the field can be reengineered for the 2020 Census. A reengineered address canvassing involves both in-office and in-field components to update and validate the address list in preparation for the 2020 Census. The Geography Division recommends that the Census Bureau move forward with plans for a reengineered address canvassing for the 2020 Census.

1.1 Purpose and Scope

In a reengineered address canvassing, the Geography Division has the crucial work of assuring a complete, valid address list using sources and methods that are not directly observed in the field. This report summarizes activities that improve the quality of address data in, and coverage of, the Master Address File/Topologically Integrated Geographic Encoding and Referencing (MAF/TIGER) System in support of a reengineered address canvassing. Major activities to improve the quality and coverage of MAF/TIGER included in this report are:

- Continuous quality measurement of MAF/TIGER (e.g., Quality Indicators)
- Processes to acquire addresses that both validate and improve MAF/TIGER (e.g., Partnership Program)
- Research and analysis to identify, and even predict, the types of geographic areas that could be managed through in-office review and areas that would require address validation and updating with field canvassing (e.g., area classification and statistical modeling)

Also included is preliminary but promising work on other information sources:

- Third-party data and administrative records
- Comparison of the MAF to imagery

After a brief background section, this report covers methods and results for each of the activities listed above; provides the detail for the recommendation to move forward with plans for a reengineered address canvassing; gives representative impacts, benefits, and risks inherent to the recommendation; and furnishes next steps, which will include a process for the determination of the percentage of housing units for in-office and in-field canvassing and an implementation plan for the geographic component of a reengineered address canvassing.

Out of Scope. The report does not name specific places to be canvassed in the field, speak to specific decennial census operations in a reengineered address canvassing, or estimate costs to implement in-field address canvassing.

1.2 Background

1.2.1 Address List Improvement for Decennial Censuses

Prior to an address canvassing field operation, the Census Bureau creates and updates its address list using information from the United States Postal Service (USPS) Delivery Sequence File (DSF), other sources, and through limited fieldwork such as the Community Address Updating System (CAUS) program. In previous address canvassing operations, field representatives traversed every road and visited each residential address in the United States (except for the most remote locations, for example, the greatest portion of Alaska). In a reengineered address canvassing, the field work would be limited to specific geographic areas, where necessary, to ensure an accurate and complete address list, i.e., field staff would visit housing units only in selected areas.

Each decennial census that has relied on mailed questionnaires has used some form of canvassing to validate and update the Census Bureau address list prior to mailing questionnaires. For the 1970, 1980, and 1990 censuses, the Census Bureau began with a commercially purchased address list for available metropolitan areas then conducted canvassing operations to improve the list. For Census 2000, an objective was to build and maintain a permanent housing unit address list for future use. The starting point to build the initial MAF was the1990 Address Control File. It was supplemented with addresses from the DSF; from partners through a new decennial program, the Local Update of Census Addresses (LUCA); and with address updates from Census 2000 field activities.

For the 2010 Census, 96.7 percent of the addresses included in the final address list were city-style (house number, street name) and 3.3 percent were non-city-style (e.g., rural route, post office box, or general delivery).[1] For the 2000 and 2010 censuses, the principal sources of address data in city-style areas were the DSF and Census Bureau field activities, including address listing for the Decennial Census and the ACS. Field activities have served to verify the existence and quality of addresses already in the MAF and as sources of new address information.

The advent of the American Community Survey (ACS) following Census 2000 required that the MAF be updated throughout the decade, rather than once per decade, to provide new and changed addresses for the ACS address frame. Between 2000 and 2010, addresses in MAF/TIGER were updated multiple times using the DSF, CAUS, and prior to the 2010 Census, LUCA. A full in-field address canvassing was conducted in 2009 to validate and correct the address list used for the 2010 Census and to add information such as geographic coordinates for housing units.

[1] 2010 Census Operational Assessment for Type of Enumeration Area Delineation www.census.gov/2010census/pdf/2010_Census_TEA_Delineation_Assessment.pdf

1.2.2 The Geographic Support System Initiative

In 2009, the Census Bureau proposed the Geographic Support System Initiative (GSS-I), an integrated program of improved address coverage, road updates, and quality assessment of MAF/TIGER System. Work under the GSS-I commenced in 2011 (see Figure 1. GSS-I Timeline). GSS-I research and analysis activities fulfill a primary goal of supporting a reengineered address canvassing for the 2020 Census. Major participants in the GSS-I are federal, state, local, and tribal governments.

Figure 1. GSS-I Timeline

Sections 2 and 3 that follow cover methods and results for activities of the GSS-I that inform the recommendation to move forward with plans for a reengineered address canvassing. Section 2 includes Quality Indicators, the GSS-I Partnership Program, and third-party data and administrative records. Section 3 includes area classification research and statistical modeling to assist in determining where in-field address canvassing is needed.[2]

2 Assess and Improve MAF/TIGER

2.1 Quality Indicators

Quality Indicators (QIs) measure the quality of MAF/TIGER address and road data and their completeness by census block and census tract, and contribute to an overall assessment of MAF/TIGER data quality. QIs allow the comparison of census blocks and census tracts based on their quality evaluation. The evaluation is used to identify geographic areas on which to focus partnership activities and contribute to a process for determining where to conduct in-field address canvassing.

Address QI scores are established through analyzing selected characteristics (known as sub-QIs) such as the existence of a house number, street name, and ZIP Code; whether the address is used by the USPS for mail delivery; and whether the address can be accurately geocoded. Sub-QIs are applied to each address and are scored on a scale from zero (poor) to 100 (excellent). The scores are aggregated by creating a census block score that contributes to an overall tract score, also on a zero-to-100 scale. Scores are recalculated as updates are made to MAF/TIGER.

Address QI scores measure consistency of data as the means to demonstrate confidence in MAF/TIGER address data. Address sub-QI scores take into account factors such as DSF history, whether latitude/longitude coordinates are associated with the address, whether it is city-style, and whether it is geocoded to a census block. The scores are high or increase when addresses in the census tract indicate a combination of factors such as multiple years of consistent DSF history, existence of a geographic location for an address, and completeness of city-style addresses with associated geocodes. Scores are low or decrease when there is an inconsistent or short history of DSF updates for addresses in the census tract or when there are missing, inconsistent, or multiple locations assigned to an address.

Road QIs measure the quality of the road network within MAF/TIGER. Similar to address QIs, road QI scores result from characteristics at the individual road segment level, collected at the census block level and aggregated to a score for each census tract. The road QI score is composed of three sub-QIs related to spatial accuracy, name quality and consistency, and address

[2] While these are major activities of the GSS-I contributing to a reengineered address canvassing, they are not the whole of the GSS-I nor do they include related 2020 Census research and testing projects.

range quality. With regard to address ranges, the sub-QI takes into account whether address ranges along a road feature are mutually exclusive, overlap, or are imputed.

Address and road QI scores provide a measure of the fitness for use of data within MAF/TIGER and contribute to a greater understanding of, and confidence in, the contents of MAF/TIGER for use in decision making. As with other measures discussed later in this report, QIs provide one method for assessing information within MAF/TIGER and are most effective when used in combination with other data and measures.

2.1.1 Examples of QIs Applied with Other Data

Census tract 6011.07 in Howard County, Maryland has an address QI score of 77.8, the lowest in the county. According to the 2010 Census, there were 1,807 housing units in the tract. The 2013 MAF contained the same number of units (though only 1,721 have city-style addresses associated with them). The Spring 2012 DSF contained 1,742 addresses for the tract. The relatively low address QI score for this tract combined with the variation in counts of addresses and housing units between sources suggests the need for further review to determine the reason for variation and to identify the appropriate process for improving information in the MAF. A review of imagery for this census tract showed the inclusion of a mobile-home park, which may account for the difference between the number of housing units counted in the census and contained in the MAF and the number of addresses in the DSF; it is not uncommon for mail to be delivered to the address associated with the mobile-home park office and then distributed from there to the residents.

As another example, census tract 6023.05, also in Howard County, Maryland, has an address QI score of 90.9, the highest in the county. The numbers of housing units and addresses in this tract are relatively consistent across various sources and across time: 1,267 housing units according to the 2010 Census; 1,260 addresses in the Spring 2012 DSF; and 1,273 addresses in the 2013 MAF (1,272 of which have city-style addresses). The high address QI score combined with the consistency of address and housing unit counts suggests that address information for this census tract can be managed and updated in-office.

2.2 Partnership Program

2.2.1 Description

The GSS-I Partnership Program provides an opportunity for state, local, and tribal governments to supply data to the Census Bureau throughout the decade in a joint effort to improve the quality and coverage of Census Bureau address and road data. A foundation of the program is the recognition that local governments are authoritative sources for quality address and road data within their communities and acquisition and evaluation of these data (using aerial imagery and

comparison to the Census Bureau's existing inventory of addresses) will minimize the need to conduct an address canvassing in many areas.[3]

An accurate road network is important for ensuring that the addresses can be precisely associated with their location on the ground through geocoding. Partner-provided road data is a low cost source to help the Census Bureau ensure that MAF/TIGER is accurate and current without costly field activities. Where partner-provided road data are either not available or are low quality, third-party data will be pursued as a source.

2.2.2 Method

In 2012, a partnership program strategy was implemented that identified particular partners based on characteristics of the MAF addresses for their area, on their associated QIs, and that included varying sizes and locations across the country (see Figure 2. Distribution of Entities Contacted, by Type). For example, could files be obtained from partners in sparsely populated rural areas to confirm that the MAF had sufficient and accurate coverage in areas with low QIs? The GSS-I Partnership Program solicited state, local, and tribal governments within these parameters to submit address and road data.

Once acquired, each partner file was examined, the content inventoried, and determinations were made on compliance with data content guidelines for addresses and roads.[4] Prior to use, the partner-provided data underwent a series of automated checks and analytic reviews. This process matched partner addresses to the MAF, interactively reviewed non-matches to avoid duplication in the MAF, and validated that new addresses represented structures that actually existed on the ground (See Figure 3. Partner Data Processing).

2.2.3 Results

Unless otherwise indicated, data in the following sections result from files acquired for the GSS-I Partnership Program by April 30, 2014.

Program Partners. The GSS-I Partnership Program contacted 301 potential partners: 255 sent data while 46 governments were unable to participate because the requested data did not exist or they required a use agreement or user fee (see Table 1. Partner Type and Table 2. Reasons for Non Participation). The partner files acquired encompassed 7,864 governmental entities, 37 percent of the housing units, and 39 percent of the population in the 2010 Census.

[3] Local governments are an authority for address and road data based on both their charter- or ordinance-derived legal authority over zoning and permitting activities and their function as a source of data for other authorities, including their respective states and the USPS.
[4] www.census.gov/geo/gssi/addgdln.html

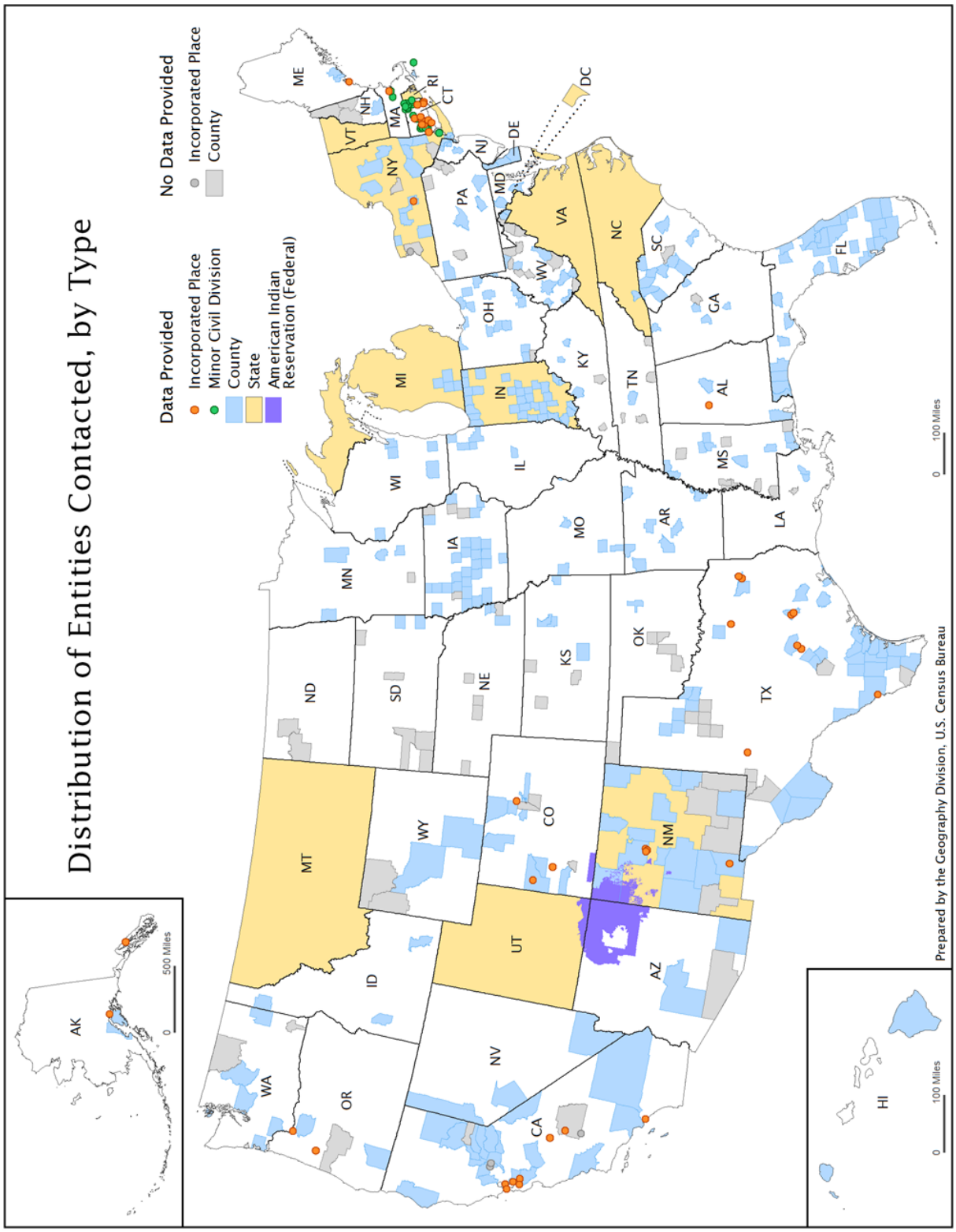

Figure 2. Distribution of Entities Contacted, by Type

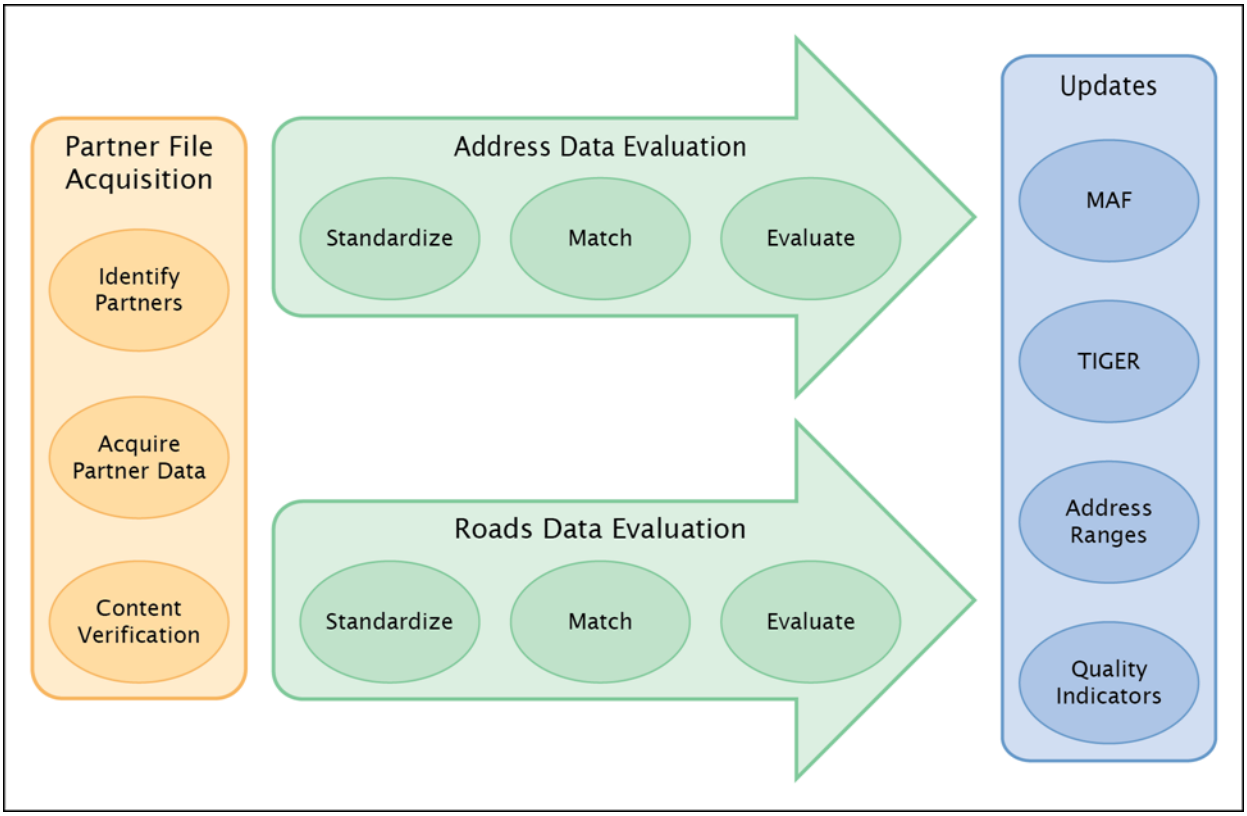

Figure 3. Partner Data Processing

Table 1. Partner Type

	Number of Participating Partners	Number of Non-Participating Partners
State	10	1
County	175	45
Local	67	0
Tribal	3	0

Table 2. Reasons for Non Participation

	Number of Non-Participating Partners
Data did not exist	28
Use agreement required	13
Fee required	4
Refused to participate	1

Address and Geocode Results. Of the address files received from 255 partners to date, 90 percent of the files, or 30,547,359 addresses, were processed. The remaining 10 percent of the address files received were not processed because the data did not meet minimum address submission guidelines, for example, data did not contain unique identifiers.[5] After processing, the MAF was updated with 76.8 percent of the addresses submitted by partners. Examples of

[5] www2.census.gov/geo/pdfs/gssi/Address_Data_Submission_Guidelines_v1.1.pdf

addresses not used include duplicates in the partner file, non-residential addresses, and addresses missing required content such as unit designations within a multi-unit structure.

Of the 30.5 million partner addresses used to update the MAF, 76.6 percent matched existing MAF addresses and did not need additional review. The remaining 23.4 percent (7.1 million) of partner-provided addresses were systematically reviewed for possible inclusion in the MAF. During detailed review, many of the non-matched partner-provided addresses were found to already exist in the MAF in a slightly different format that prevented automated matching, and many others were non-residential addresses. At the conclusion of the review, 64,183 of the non-matched addresses were identified as unique housing units that should be added to the MAF.

For 97.7 percent of the addresses that matched to the MAF, the partner-provided geocode associated with the address in the MAF matched, confirming that the Census Bureau had the address and it was correctly geocoded. For 1.0 percent, the MAF was updated with the partners' geocodes. In addition, 1.3 percent (306,835) of previously ungeocoded addresses were geocoded in the MAF for the first time using latitude/longitude information contained in the partner-provided data (see Table 3. MAF Updates from Partner Address Data).

Table 3. MAF Updates from Partner Address Data

	Number	Percent
Total partner addresses	30,547,359	
Partner addresses matched existing MAF addresses	23,410,653	76.6
Geocode unchanged	*22,876,559*	*97.7*
Geocode updated	*236,259*	*1.0*
Geocode added (previously ungeocoded)	*306,835*	*1.3*
Partner addresses did not match existing MAF addresses	7,136,706	23.4
Partner addresses added to the MAF	*64,183*	*0.2*
Partner addresses not used for MAF update[a]	*7,072,523*	*99.8*

[a] An address was not used for MAF update, e.g., it was a duplicate in the partner file or was a non-residential address

Road Results. The GSS-I Partnership Program received road data from 255 partners covering 39 percent of the 2010 Census population. Three files were not processed because the data did not conform to minimum feature data submission guidelines.[6] MAF/TIGER was updated by adding 11,922 miles and modifying 35,815 miles of road. Added roads generally consisted of newly-constructed roads. Modified roads generally were those that required alignment to imagery (see Table 4. TIGER Updates from Partner Road Data).

[6] www2.census.gov/geo/pdfs/gssi/Feature_Data_Submission_Guidelines_DRAFTv1.0.pdf

Table 4. TIGER Updates from Partner Road Data

	Number of Road Miles
Total miles of road in the nation	6,346,165
Total miles of roads in areas evaluated to date	1,367,713
Total miles of roads added to date	11,922
Total miles of roads modified to date	35,815

2.2.4 Discussion

Availability of Address Data. The nationwide availability of high-quality address data is a critical component for determining the methodological approach to a reengineered address canvassing. If high quality data are not available, the ability to improve MAF/TIGER is limited and a nationwide in-field address canvassing operation will be necessary. The GSS-I Partnership Program has been successful at acquiring, evaluating, and incorporating current, high-quality address and road data from 255 partners to date that encompass 7,864 governmental entities, 37 percent of the housing units, and 39 percent of the population in the 2010 Census.

The GSS-I Partnership Program shows that the Census Bureau is more likely to acquire partner address and road data in urban and suburban areas, many of which are likely to show growth. This supports the notion that partner data can be used to both validate and supplement the MAF in these areas. Experiences indicate partner data are less likely to be acquired and successfully processed for sparsely-populated rural areas containing non-city-style addresses. Planning is underway to use alternative methods to supplement the address list in these areas with third-party data and possibly addresses from administrative records; see Section 2.3.

Utility of Partner Data Acquired. One key finding from the initial processing of partner data is that of the 30.5 million addresses processed, 76.6 percent matched existing MAF addresses, and of those, 97.7 percent already were geocoded to the correct block. In other words, the large majority of addresses in partner data are already in the MAF and therefore are candidates for inclusion in the 2020 Census. As a result, many addresses in the MAF can be independently validated using partner data, reducing the need to field validate all addresses in the MAF. A second key finding concerns efficiency. The acquisition, review, and use of partner data to update MAF/TIGER is also one of the most resource-efficient methods of updating the MAF compared to various 2010 Census field operations and in-office geocoding resolution (see Table 5. Production Rate Comparison of GSS-I and Other Address Review Operations).

Table 5. Production Rate Comparison of GSS-I and Other Address Review Operations

Program	Addresses Per Hour
Partnership Program updates (through March 31, 2014)[a]	1,240.7
2014 in-office geocoding	14.0
2010 Address Canvassing nationwide average[b]	15.4
2010 Address Canvassing large blocks average[b]	25.2
2000 Census Block Canvassing[b]	24.1
2000 Census Address Listing[b]	4.0

[a] Assumes 40 hours per file for address evaluation and processing
[b] From *2010 Address Canvassing Operational Assessment for Type of Enumeration Area Delineation*, http://www.census.gov/2010census/pdf/2010_Census_TEA_Delineation_Assessment.pdf

Prior to conducting a census, it is important that MAF/TIGER has an accurate and complete depiction of roads to ensure that addresses and their associated housing units are geocoded to the correct location. Accurate geocoding will ensure that statistical models and other geographic analyses that use the location of addresses as an input are accurately reflecting what exists on the ground. Roads also help with fieldwork and boundary delineation. MAF/TIGER was updated with 11,922 new and 35,815 modified miles of road as a result of the GSS-I Partnership Program.

The number of roads added and modified vary by partner. In some areas, few updates were made to MAF/TIGER because roads already were current and high quality. For example, for King County, Washington, a large urban county with 14,843 miles of road, only 48 (0.3 percent) miles of road were added and 151 (1.0 percent) miles modified because TIGER was already current and complete for the county. In contrast, for Columbia County, Georgia, a smaller rural county with 1,185 miles of road, 116 (9.8 percent) miles of road were added to TIGER and 130 (11.0 percent) of miles modified.

Another significant positive result of matching partner-provided address data to MAF/TIGER is the reduction in the number of ungeocoded addresses in the MAF. Ungeocoded addresses are not included in the census frame because, without a census block location, the enumeration data associated with the address cannot be tabulated to the correct jurisdiction and census block. When ungeocoded addresses are resolved, the coverage for that area improves and there is less need to conduct in-field address canvassing. As a result of processing partner files, 327,799 ungeocoded addresses in MAF/TIGER were geocoded (see Table 6. Impact on Ungeocoded Addresses After Processing Partner Address and Road Data: Areas Covered by Partners to Date). After including the 64,183 newly added addresses, the number of addresses added at this time to the 2020 Census universe is 391,982.

Table 6. Impact on Ungeocoded Addresses After Processing Partner Address and Road Data: Areas Covered by Partners to Date

Areas Evaluated to Date	Number of Addresses	Percent of Area Total
MAF addresses in partner areas	57,303,095	
Ungeocoded MAF addresses before processing partner data	1,545,676	2.7
Geocoded based on partner data	327,799	0.6
Ungeocoded addresses remaining in MAF	1,217,877	2.1

2.3 Third-Party Address Data and Administrative Records

As a means to enhance the coverage and accuracy of MAF/TIGER in areas where partner data is not available, and to augment the available partner data, the feasibility of using third-party data and administrative records is being investigated. The Census Bureau is evaluating the use of commercial address and road data from third-party vendors to enhance the coverage and accuracy of the MAF. Current activities include a feasibility study using a national third-party data set and a Request for Information (RFI) about the current state of vendor-supplied address data.[7]

2.3.1 Third-Party Address Data

In the third-party address feasibility study, duplicates within the third-party data (0.4 percent) were removed and addresses matched to the MAF using processes similar to the GSS-I Partnership Program (see Section 2.2.2). The initial results show that 82.5 percent (139.2 million) of the addresses from the third-party data match addresses in the MAF and that 17.1 percent (28.8 million) of the addresses did not match and would require review before being added to the MAF (see Table 7. Initial Match Rates of Third-Party Data and the MAF). While the overall match rate of 82.5 percent is higher than most partner files, the high volume of non-matched addresses requiring interactive review introduces some efficiency issues in using current methods to identify addresses in the third-party data that are missing from the MAF. In addition, very few adds have been identified in the initial evaluation of these non-matched addresses, and the percentage of new addresses identified from the third-party file is much lower than the percentage of new addresses identified in the typical file processed in the GSS-I Partnership Program. If these initial efforts, however, indicate third-party data are a viable source for address validation and additions to the MAF, the Census Bureau will pursue the use of additional third-party data.

[7] FedBizOpps.gov, Solicitation Number DEC-14-147, Street Centerline Data and Address Data

Table 7. Initial Match Rates of Third-Party Data and the MAF

	Number	Percent
Total third-party addresses	168,623,188	
Duplicates within third-party data	626,675	0.4
Total third-party addresses to match to the MAF	167,996,513	
Matched	139,192,124	82.5
Unmatched	28,804,389	17.1

2.3.2 Administrative Records

The Geography Division is also developing a method of evaluating the feasibility of using addresses from selected administrative records to enhance the coverage and accuracy of the MAF, especially in areas that are likely to change, but where partner files are not available. This evaluation will use many of the same methods used in the GSS-I Partnership Program, where addresses are matched to the MAF, and then unmatched records are reviewed for possible inclusion in the MAF (see Section 2.2.2). Some of the administrative records data under consideration are the Indian Health Service Registration File, the U.S. Department of Housing and Urban Development Public and Indian Housing Information Center File, and the Selective Service Registration File.

2.3.3 Discussion

To date, the feasibility of using third-party address data and administrative records is unknown. Research into the use of third-party files suggest that match rates for addresses in sources investigated so far are relatively consistent with match rates obtained using GSS-I partner files, however, the volume of addresses unmatched from third-party sources will require additional research. Future efforts should include evaluation of additional third-party sources to learn whether the high number of unmatched addresses is anomalous to the third-party data reviewed and the development of alternative strategies to more efficiently identify addresses that are not in the MAF. Responses to the RFI about the current state of third-party address data will also inform the Census Bureau about the potential use and benefits of third-party address data. Finally, planned research into the feasibility of using address data from selected administrative records will also provide insight into the feasibility of using these data sources for updating the MAF.

3 Research, Modeling, and Area Classification

A component of the recommendation for a reengineered address canvassing is the feasibility of using rigorous models and methods in determining to what extent, and where, in-field address canvassing is required.

For a substantial proportion of census blocks, the number of actions (addresses added, deleted, moved, and with other changes) identified within each census block during the 2010 address

canvassing operation was small: two-thirds of the 10 million addresses added were located in only four percent of census blocks.[8] The concentration of a large proportion of actions in a small proportion of census blocks suggests that not all census blocks need to be canvassed in the field, supporting the recommendation for limiting the in-field component of a reengineered address canvassing. Building on previous research, and knowledge of address canvassing results, as well as patterns of population and housing distribution and change, current research is focused on identifying factors that provide reliable predictors of where canvassing in the field might be necessary, for example, indicators of stability and consistency in land use patterns, numbers and distribution of housing units, and existence of address lists. Specifically, research, analysis, and modeling efforts are focused on assessing census block stability and consistency, comparison of the MAF to imagery, and statistical models. The Reengineered Address Canvassing Continuum (see Figure 4) provided a means to conceptualize and guide research, data gathering, analysis, and decision-making.

3.1 Assessing Stability and Consistency

A key aspect of research is focused on measuring stability and consistency of housing unit and address counts as well as classifying census blocks based on land uses and types of housing units. In general, where stability and consistency can be identified between housing unit and address counts, there can be greater confidence in the MAF and in the ability to rely upon the DSF, local files obtained through the GSS-I Partnership Program, and other sources to update and maintain the MAF without having to canvass in the field. Identifying highly stable areas and areas with unique or non-residential land uses that can be maintained through in-office review and update helps narrow the types of geographic areas and number of housing units requiring canvassing in the field.

3.1.1 Delivery Sequence File Stability Index

The DSF Stability Index provides an indicator of stability of addresses in the DSF over a specified duration and, by extension, the MAF in those areas where the DSF is the primary source for address updates (the MAF is updated using the DSF twice yearly throughout the decade). The DSF Stability Index is calculated by tracing the presence of addresses in the DSF at the end of the period through each preceding DSF for the time period. Index values range from zero to one; an index value of one indicates that each address in the final DSF for the time period appears in each preceding DSF. The higher the index value the greater the stability of addresses

[8] Boies, John L., Kevin M. Shaw, and Jonathan P. Holland, 2012. 2010 Census Address Canvassing Targeting and Cost Reduction Evaluation Report, 2010 Census Program for Evaluations and Experiments at www.census.gov/2010census/pdf/2010_Census_Address_Canvassing_Targeting_and_Cost_Reduction_Evaluation_Report.pdf

Figure 4. Reengineered Address Canvassing Continuum

To provide a structure to aid in determining the accuracy of the address frame for the FY2015 Targeted Address Canvassing decision

	Likely not canvass	**Likely Canvass**
Current State	QI trend positive (Address, Features, Geocoding, Geographic Areas, Completeness and aggregate QI values	QI trend negative (Address, Features, Geocoding, Geographic Areas, Completeness and aggregate QI values
	100% city style addresses	0% city style addresses
	Area not known to have hidden units	Area known to have hidden units
	Area not known to have hard to count populations	Area known to have hard to count populations
	Partner file available for the area	Partner file not available for the area
	Area not known to have informal housing, unique housing situations/communities	Area known to have informal housing, unique housing situations/communities
	Area not classified as “needs to be canvassed” or containing an address list error	Area classified as “needs to be canvassed” or containing an address list error
	Area not known to have subsidized housing	Area known to have subsidized housing
	Area not known to have seasonal housing	Area known to have seasonal housing
	Area not known to have single-to-multi-unit, or multi-to-single unit conversion	Area known to have single-to-multi-unit, or multi-to-single unit conversion
	100% MAF and TIGER agreement on geocodes	0% MAF and TIGER agreement on geocodes
	Area contains a special land use (includes federally owned and managed lands)	Area does not contain a special land use (includes federally owned and managed lands)
	100% MAF address confirmation rate (matching rate) with administrative records	0% MAF address confirmation rate (matching rate) with administrative records
	100% of MAF units have corresponding MSPs	0% of MAF units have corresponding MSPs
	Successful questionnaire delivery and response, or successful interview	Unsuccessful questionnaire delivery and response, or unsuccessful interview
	“High” DSF stability index (semi-annual and trend assessment)	“Low” DSF stability index (semi-annual and trend assessment)
	Land use/land cover data indicates no change can occur	Land use/land cover data indicates no change can occur

	Likely not canvass	Likely Canvass
	0% small multi-unit structures	100% small multi-unit structures
	100% residential or 100% non-residential addresses	A mix of residential and non-residential addresses
	100% DSF address coverage	0% DSF address coverage
	0% trailer/mobile home parks	100% trailer/mobile home parks
	Inside or over 5 km from an urban area boundary	Outside and within 5 km from an urban area boundary
	Inside or over 5 km from a college or university	Outside and within 5 km from a college or university
	Housing units with average persons per household	Housing units with higher than average persons per household
	Demographic variables indicate the unlikely presence of informal or hidden housing units	Demographic variables indicate the likely presence of informal or hidden housing units
Change Detection	Area government(s) have provided "high quality," "trusted" data	Area government(s) have not provided "high quality," "trusted" data
	Area government(s) provided multiple vintages of data	Area government(s) have not provided multiple vintages of data
	PEP indicates no/little housing unit change	PEP indicates signficant housing unit change
	iSIMPLE indicates no changes have occurred	iSIMPLE indicates changes have occurred
	Imagery change detection indicates no changes have occurred	Imagery change detection indicates changes have occurred
	Area government-provided data indicate no changes have occurred	Area government-provided data indicate changes have occurred
	No significant event: No address conversion, natural disaster, environmental disaster, and/or redevelopment	Significant event: Address conversion, natural disaster, environmental disaster, and/or redevelopment
	Static housing values	Changing housing values
	No recent or current economic change	Recent or current economic change
	DSF trends, as compared to the MAF, indicate the area is stable	DSF trends, as compared to the MAF, indicate the area is not stable
	DSF refresh does not add addresses into the MAF	DSF refresh does add addresses into the MAF

	Likely not canvass	**Likely Canvass**
	Utility customer data matches the MAF and is stable over time	Utility customer data does not match the MAF and changes over time
	No DSF excluded from delivery statistics (EDS) flags	DSF excluded from delivery statistics (EDS) flags
Predictive Change	Zoning data indicates no Hus can be built	Zoning data indicates Hus can be built
	Percentage of housing at different ages indicates current and future stability	Percentage of housing at different ages indicates current and future change
	2020 MAF error model predicts no errors	2020 MAF error model predicts significant errors
	No planned significant event: No address conversion, environmental remediation, and/or redevelopment	Planned significant event: Planned address conversion, environmental remediation, and/or redevelopment
	Urban growth model indicates no/low future growth in the area	Urban growth model indicates future growth in the area
	Urban growth boundaries (OR and WA) indicate no/low future growth in the area	Urban growth boundaries (OR and WA) indicate future growth in the area
	No annexations in the Southwest indicate no/low future growth in the area	Annexations in the Southwest indicate no/low future growth in the area
	Adjacent to a block with no/low number of MAF adds (threshold TBD)	Adjacent to a block with a certain number of MAF adds (threshold TBD)
	No GSS-I partner file provisional adds (addresses), new construction, and "shelf" features (linear)	GSS-I partner file provisional adds (addresses), new construction, and "shelf" features (linear)
2010 Baseline	No address changes (adds, deletes, changes) during 2010 operations: AdCan, LUCA, new construction, non-ID, PBO	Significant ddress changes (adds, deletes, changes) during 2010 operations: AdCan, LUCA, new construction, non-ID, PBO
	2010 HU counts equal to 2010 MAF units	Significant number or percentage difference between 2010 HU counts and 2010 MAF units
	No Census Coverage Measurement identified issues	Significant Census Coverage Measurement identified issues
	No type A non-ID adds	Significant number of type A non-ID adds
	No successful 2010 CQR cases	100% successful 2010 CQR cases
	No (0%) undeliverable addresses for Census	100% undeliverable addresses for Census

	Likely not canvass	**Likely Canvass**
	2010	2010
	MAF units with a positive enumeration	MAF units without a positive enumeration
	MAF units in 2010 Census and either Fall ’09 DSF or Spring ’10 DSF	MAF units in 2010 Census and not within either Fall ’09 DSF or Spring ’10 DSF

in the DSF. Census tracts with the highest levels of stability tend to be located in urban and suburban areas (see Figure 5. DSF Stability by Census Tract: 2010). The number of census tracts within the United States with the highest rates of DSF stability and numbers of DSF addresses within those census tracts are shown in Table 8. Number of Census Tracts and Addresses by DSF Stability Index Value: 2009-2012. The DSF Stability Index provides one measure of stability and should be used with other measures of stability, consistency, and quality in decision-making.

Table 8. Number of Census Tracts and Addresses by DSF Stability Index Value: 2009-2012

DSF Stability Index	2009-2010		2010-2011		2011-2012	
	Census Tracts	Addresses Spring 2010 DSF	Census Tracts	Addresses Spring 2011 DSF	Census Tracts	Addresses Spring 2012 DSF
1.0	12,380	19,058,903	13,834	21,656,324	19,167	30,558,455
0.980 – 0.999	47,771	81,370,407	50,405	86,173,997	46,642	80,834,772
0 – 0.979	12,906	19,405,534	8,818	12,793,266	7,248	10,009,467
Total	73,057	119,834,844	73,057	120,623,587	73,057	121,402,694

3.1.2 Measuring Housing Unit and Address Consistency Across Datasets

The number of addresses in the Spring 2012 DSF for each census tract in the United States was compared to the 2010 Census housing unit count (the latter providing a baseline against which to measure the DSF counts). Counts were considered consistent if the 2010 Census housing unit count was within 0.5 percent of the number of addresses in the Spring 2012 DSF. Research also used a broader definition of consistency and identified census tracts where the 2010 Census housing unit count was within two percent of the number of addresses in the Spring 2012 DSF (see Table 9. Comparison of 2010 Census Housing Unit Counts with Spring 2012 DSF Address Counts, by Census Tract). Because this research was occurring at the same time as processing of data in the GSS-I Partnership Program, the analysis could not include counts of addresses from partner files in the calculation of housing unit and address count consistency. Counts of addresses from partner files can be included in future calculations, providing a stronger indicator of consistency or inconsistency across various address sources. As with the DSF Stability Index, this measure of housing unit consistency should be used with other measures of stability, consistency, and quality in decision-making. For example, inconsistency between the 2010 Census count and the Spring 2012 DSF may be a result of housing unit change. The consistency measure could be compared to other indicators of change to determine the extent to which each address source reflects expected levels of change.

About the Term Consistency. Consistency is important because when multiple independent sources validate the same address it provides confidence that the address is accurate. Measuring housing unit and address count consistency across data sources as a range around a baseline value can help identify where to focus efforts to achieve greater benefit in updating the MAF.

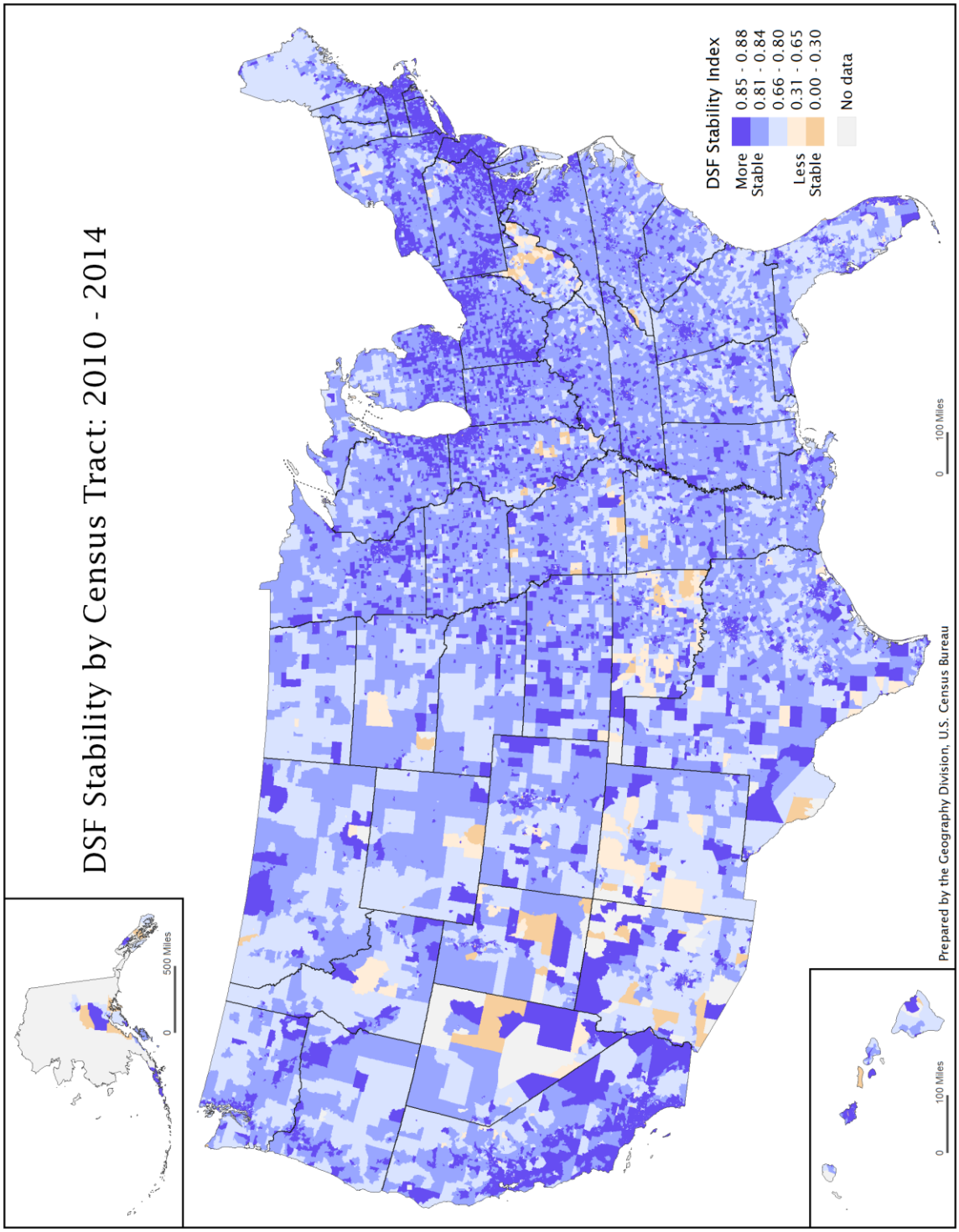

Figure 5. DSF Stability by Census Tract: 2010-2014

Table 9. Comparison of 2010 Census Housing Unit Counts with Spring 2012 DSF Address Counts, by Census Tract

	Census Tracts		2010 Census Housing Units		Addresses in Spring 2012 DSF	
	Number	Percent	Number	Percent	Number	Percent
Within 0.5% of DSF count	11,648	15.9	20,675,277	15.7	20,670,633	17.0
Greater than 0.5% and less than or equal to 2.0% of DSF count	16,987	23.3	30,929,810	23.5	30,817,378	25.4
Greater than 2.0% of DSF count	44,422	60.8	80,099,643	60.8	69,914,683	57.6
Totals	73,057	100.0	131,704,730	100.0	121,402,694	100.0

3.2 Comparison of the MAF with Aerial Imagery

Comparing the MAF with imagery consists of using imagery for a visual review of housing units to identify and classify individual census blocks. This is a three-step process:

- Automated census block classification based on MAF housing unit type (collected as part of the 2010 address canvassing) and whether the number of MAF addresses has remained stable, increased, or decreased compared to 2010 Census housing unit counts;
- Manual review of imagery to identify stability or change including whether a census block appears to be built out or whether there is evidence of possible future growth; and
- Manual comparison of census block classification and number of MAF addresses with housing units visible on imagery to affirm consistency between the number of addresses in the MAF and housing units on imagery or to identify discrepancies.

Automated classification based on comparison of 2010 Census housing unit counts and 2013 MAF address counts suggest a substantial amount of stability in the nation. Of the 11.3 million census blocks (as of 2013), nearly 4.3 million (38.0 percent) had zero population and housing in 2010 and no addresses within the MAF in 2013. An additional 5.1 million blocks (45.3 percent) encompassing nearly 80.7 million addresses (58.2 percent) contained the same number of addresses in 2013 as housing units counted in the 2010 Census. This classification process, based on the comparison of the 2013 MAF to the 2010 Census, may not reflect all changes that have occurred since. Change detection through comparison of multiple vintages of imagery and other sources is critical to fully assessing the extent of housing unit change and the degree to which the MAF is kept current. This process also identified the amount of change within blocks by housing unit type and highlighted categories of areas that may warrant higher priority in review based on imagery and other sources. For example, the number of addresses in census blocks containing five or more mobile homes increased by 4.4 percent between 2010 and 2013. The number of addresses in blocks containing small multi-unit structures (two to nine units) increased by 1.8 percent. Both of these types of housing units present challenges when updating the address list based solely on the DSF and other sources.

In this research project, interactive review focused on census blocks in 29 counties. The counties chosen for review reflected different rates of estimated housing unit change, urban/rural status, contained Address Validation Test (AVT) sample blocks, or whether the government was a participant in the GSS-I Partnership Program.[9] Within each county, reviewers selected census blocks representing a variety of development patterns, such as densely and sparsely developed areas, older suburban neighborhoods, newer development, areas of single family homes, areas containing concentrations of apartment buildings, and areas with mixes of residential and commercial uses. Reviewers visually compared multiple vintages of imagery and classified census blocks as stable (no visible change), growth, or decline. Of the 11,086 census blocks reviewed in the 29 counties, 82.0 percent showed no change according to both the imagery-based review and comparison of 2013 MAF address counts to 2010 Census housing unit counts. Reviewers also compared results of automated classification to imagery-based analysis results for the census block. The automated classification results matched imagery-based results in 95.4 percent of the census blocks. This research project demonstrated the value in using imagery to assess whether the MAF is complete and accurately reflects the number of housing units visible on imagery. The project also confirmed expectations that, for many census blocks, especially those that have no additional land area available for residential development, the number of addresses and housing units has remained stable and consistent since the 2010 Census.

3.3 Statistical Models

As part of the 2010 Census Program for Evaluations and Experiments (CPEX), research began on how regression models could be used in a reengineered address canvassing. One of the primary goals of the 2010 Census Address Canvassing operation was to identify additions and deletions to the census address frame prior to the self-response phase of the census, as these frame errors have a direct impact on census coverage. In the CPEX research, using only data available prior to the 2010 Census Address Canvassing operation, each census block was assigned a probability value from several binary logistic regression models denoting the likelihood of each block containing an erroneously excluded or included housing unit (i.e., an add or delete address action code from canvassing).[10] With those modeled probabilities and the 2010 address canvassing results (specifically the listing action codes for each address in every census block), the efficacy of statistical models was evaluated. At various housing unit (HU)

[9] The Address Validation Test, underway at the time of this report, has two components: the MAF Model Validation Test (MMVT) and the Partial Block Canvassing (PBC) Test. The MMVT will collect address information for use in validating and calibrating statistical models. The PBC component will test the effectiveness of canvassing only the portion of a census block in which change is concentrated. Results from both will be used to further refine statistical models, validate accuracy of local government address lists, validate results from in-office comparisons of the MAF to imagery and other in-office review processes, and to inform planning for implementing reengineered address canvassing.

[10] For detailed methodology, see 2010 Census Address Canvassing Targeting and Cost Reduction Evaluation Report, 2010 Census Program for Evaluations and Experiments, at www.census.gov/2010census/pdf/2010_Census_Address_Canvassing_Targeting_and_Cost_Reduction_Evaluation_Report.pdf

coverage thresholds, several outcomes of a simulated 2010 reduced canvas (e.g., HU and block listing workloads, quantities of other forfeited address action codes, and cost avoidance estimates) were measured.

Between 2010 and 2014, as part of the Census Bureau's reengineered address canvassing research, various Generalized Linear Models were studied: Logistic (logit), and Negative Binomial and Poisson regression models. Logistic regression was chosen to model binary response data; negative binomial and Poisson to model count response data. Given the high frequency of census blocks with no add or delete actions, zero-inflated versions of the Negative Binomial (ZINB) and Poisson (ZIP) models were developed to achieve better model fits. In this report, we present two of the top-performing logistic regression models available as of July 2014 (see Table 10. Address Canvassing Statistical Modeling Outcomes at Selected Housing Unit Canvassing Levels, using 2010 Census Blocks).

Table 10. Address Canvassing Statistical Modeling Outcomes at Selected Housing Unit Canvassing Levels, using 2010 Census Blocks

		Block-Level Statistical Models[a]					
Outcomes and Housing Unit (HU) Canvassing Level		Logistic Regression Models					
		Adds Model			Adds and Deletes Model		
In 2009, at a HU canvassing level of …		*Canvassing only the selected blocks yields …*					
Number of Housing Units[b] (millions)		Add Capture Rate[c]	Delete Capture Rate[d]	Percent Blocks Canvassed[e]	Add Capture Rate	Delete Capture Rate	Percent Blocks Canvassed
5 percent	7.2	9.0	11.0	0.3	6.0	14.0	0.2
10 percent	14.5	17.0	20.0	1.1	13.0	25.0	0.6
20 percent	29.0	30.0	34.0	3.2	25.0	41.0	2.2

[a] The numbers in each column are specific to each model and are not mutually exclusive

[b] In total, there were approximately 145 million HUs eligible for canvassing in the 2010 Census Address Canvassing operation conducted in 2009. This is referred to as the dependent list count (U.S. and Puerto Rico).

[c] The Add Capture Rate refers to the percent of all add actions captured by canvassing only the selected blocks in 2009. The total number of add actions was about 10.8 million HUs.

[d] The Delete Capture Rate refers to the percent of all Type D (double deletes) actions captured by canvassing only the selected blocks in 2009. The total number of Type D actions was about 15.8 million HUs

[e] The total number of 2010 Census blocks is about 11.2 million (U.S. and Puerto Rico)

The first logit model, an adds-only model, predicts the presence of two or more add actions from the 2009 address canvassing operation. This model performs better than the second model, an adds and deletes model, for capturing add actions. For example, in 2009, at the 20 percent HU canvassing level (approximately 29.0 million HUs), the Adds Model captures about 30 percent of all add actions and 34 percent of delete actions; while the Adds and Deletes Model captures 25 percent of all add actions and 41 percent of delete actions. As just quantified, the Adds and Deletes Model captures approximately an additional seven percent of delete actions. Both

models identify a very small percentage of blocks for canvassing, less than four percent of all blocks for both models, at each of the five percent, 10 percent, and 20 percent HU canvassing levels.

While reducing workloads by 80 to 95 percent will achieve improvements for a reengineered address canvassing, it is inescapable that some quality degradation will follow. Moreover, it will be critical to measure the effects of a reduced canvas on the downstream census operations (e.g., nonresponse follow-up). Results from the AVT will provide the first opportunity for assessment and additional model development post-2010 Census.

While much work remains to further develop and evaluate the current models and quantify downstream effects, in terms of the original goal of the research—to substantially reduce the workload of the census listing operation, while maintaining an acceptable level of accuracy—the statistical models have yielded favorable results thus far. And, more so, the methodology provides an exceptionally low cost solution. Further, the current models have used available data as independent variables (predictors). Additional variables are now being processed that reflect geographic aspects of addresses or blocks, such as proximity to recent housing change, stability of the block over time, and indicators of quality (e.g., locatability, mailability) which are anticipated to be stronger predictors of change.

3.4 Applying Address Canvassing Research

MAF comparisons to imagery and outputs from the statistical models each produce sets of census blocks that can be organized into categories for further review. Blocks that exhibit high levels of stability and consistency can be categorized as requiring low levels of monitoring for potential change. Blocks in which change is detected or for which models predict change can be selected for additional in-office review and update of MAF/TIGER or eventually in-field address canvassing. The Reengineered Address Canvassing Decision Flow (Figure 6) outlines the current vision of the process or path that will be applied to a census block that results in a decision to canvass or not canvass in the field. The address canvassing decision flow concept assumes that in-office work improving MAF/TIGER continues in the years leading up to the 2020 Census with the goal that the in-field address canvassing universe can be reduced further and the number of updates reported from fieldwork minimized.

An additional approach focuses on removing specific areas from the address canvassing universe because the address list for such areas can be maintained and updated through methods or operations other than in-field address canvassing. These areas are characterized by sparse settlement patterns and low levels of expected population and housing unit change. The assumption is that the Census Bureau will work with federal and state agencies to assure an accurate list of addresses and housing unit locations on public lands (for example, National Park Service areas and Bureau of Land Management grazing lands) and military bases, which account for 17 percent of the nation’s land area, but less than one percent of addresses (see Figure 7. Special Land Use Areas With Little Expected Development: 2014). For areas where the Census

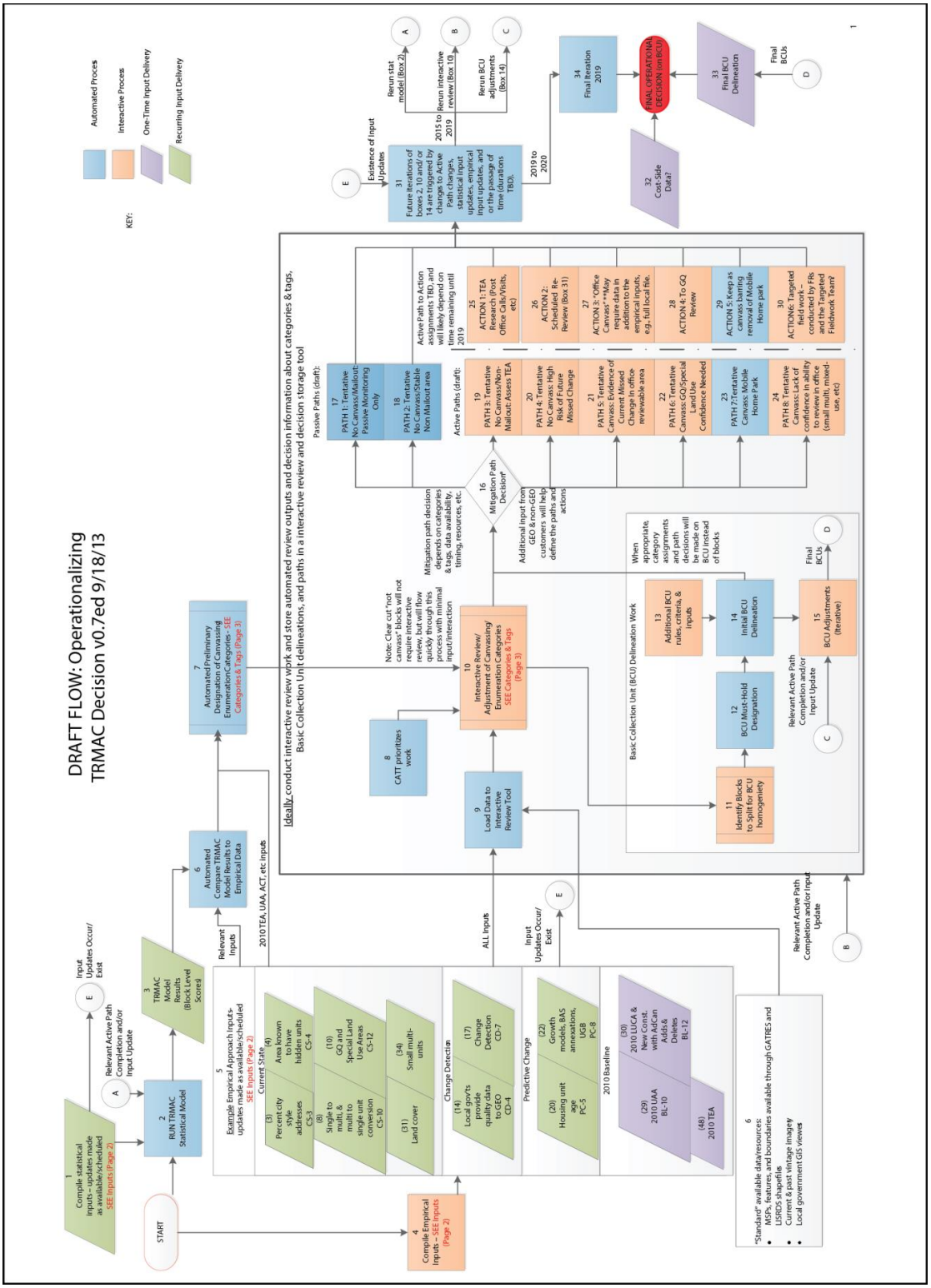

Figure 6. Reengineered Address Canvassing Decision Flow

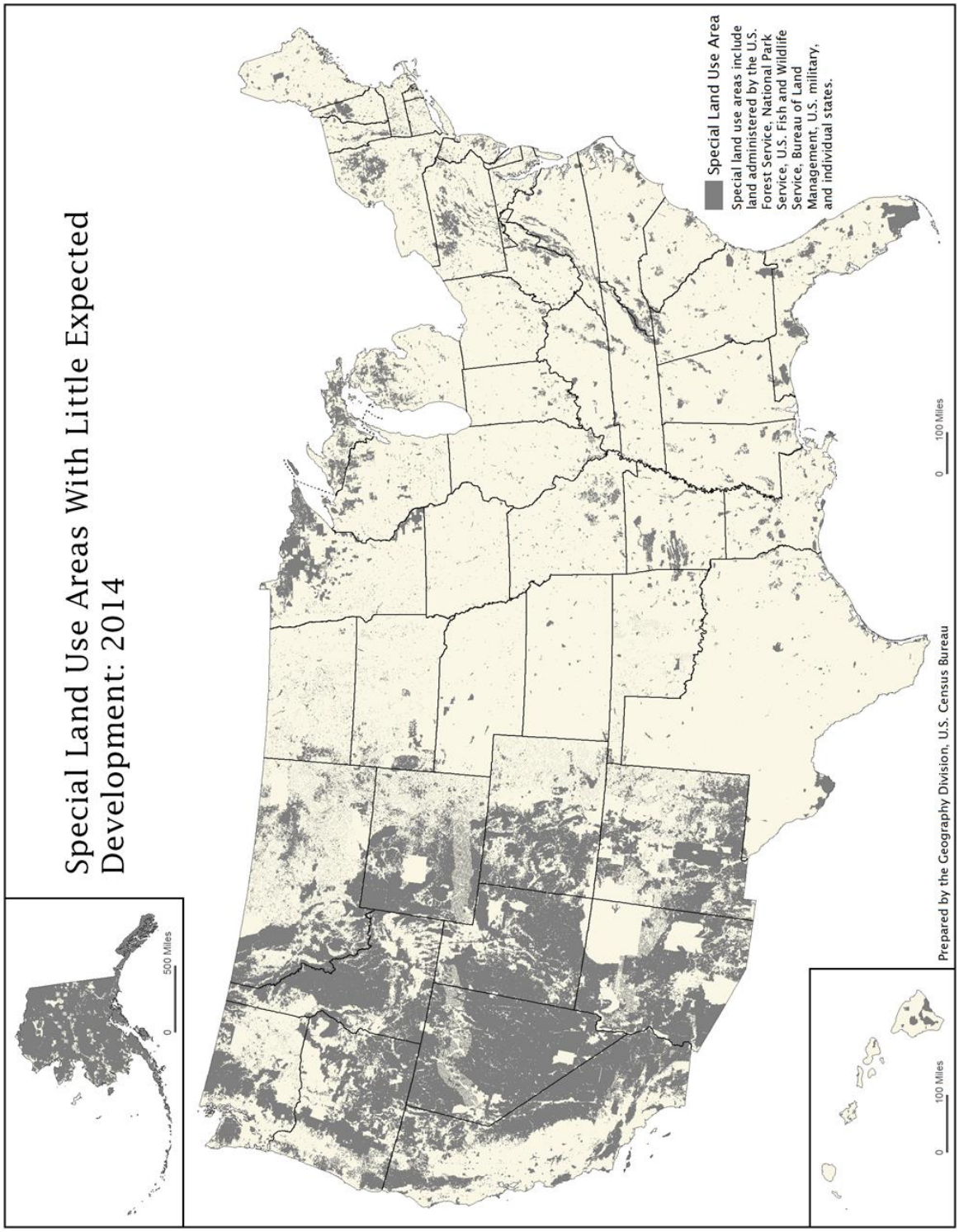

Figure 7. Special Land Use Areas With Little Expected Development: 2014

Bureau will conduct in-field enumeration for the 2020 Census, the assumption is that addresses could be updated and verified at the same time as questionnaire delivery and enumeration activities, rather than requiring two separate field operations (see Figure 8. Non-Mail Out Types of Enumeration Areas: 2010). Non-mail out types of enumeration areas accounted for approximately 50 percent of the nation's land area in 2010, but less than 10 percent of the addresses in the United States. This approach also would be applied to the 205 special land use census tracts (for example, college campuses, prisons, national, state, and local parks, airports and other nonresidential land uses) containing approximately 85,000 housing units or other types of living quarters as well as 360 census tracts consisting entirely of water area.

3.5 Analysis

Address canvassing research has demonstrated that (1) high levels of housing unit and address consistency exist throughout the United States; (2) comparison of MAF addresses and imagery for change detection can successfully highlight census blocks to be canvassed in the field; and (3) statistical modeling can contribute to the process for selecting geographic areas for review, leading to the identification of census blocks to be canvassed in the field. None of the methods employed offers a single solution, rather, each works with the others to create more certainty about areas most likely to require in-field address canvassing.

While the research and results discussed in this section are sufficient for developing a recommendation for a reengineered address canvassing, research must continue. The AVT in 2014 will provide information to validate, enhance, and improve models. Results from the AVT also will be used in comparison with data generated through the MAF to imagery comparison.

Statistical modeling efforts will continue in modeling potential change in zero population and housing blocks, incorporating address QIs, improving the ability to predict deletions, and determining the potential for integrating the various models to produce a single reengineered address canvassing model. Review of imagery and MAF housing unit counts likewise will continue in additional areas. Researchers also will integrate MAF-to-imagery comparison and statistical modeling to develop a rules- and model-based methodology for identifying which census blocks can be reviewed and managed through in-office address canvassing and the smaller number of census blocks that require canvassing in the field. In addition, output from the AVT will be used to inform the final design and implementation plan for a reengineered address canvassing.

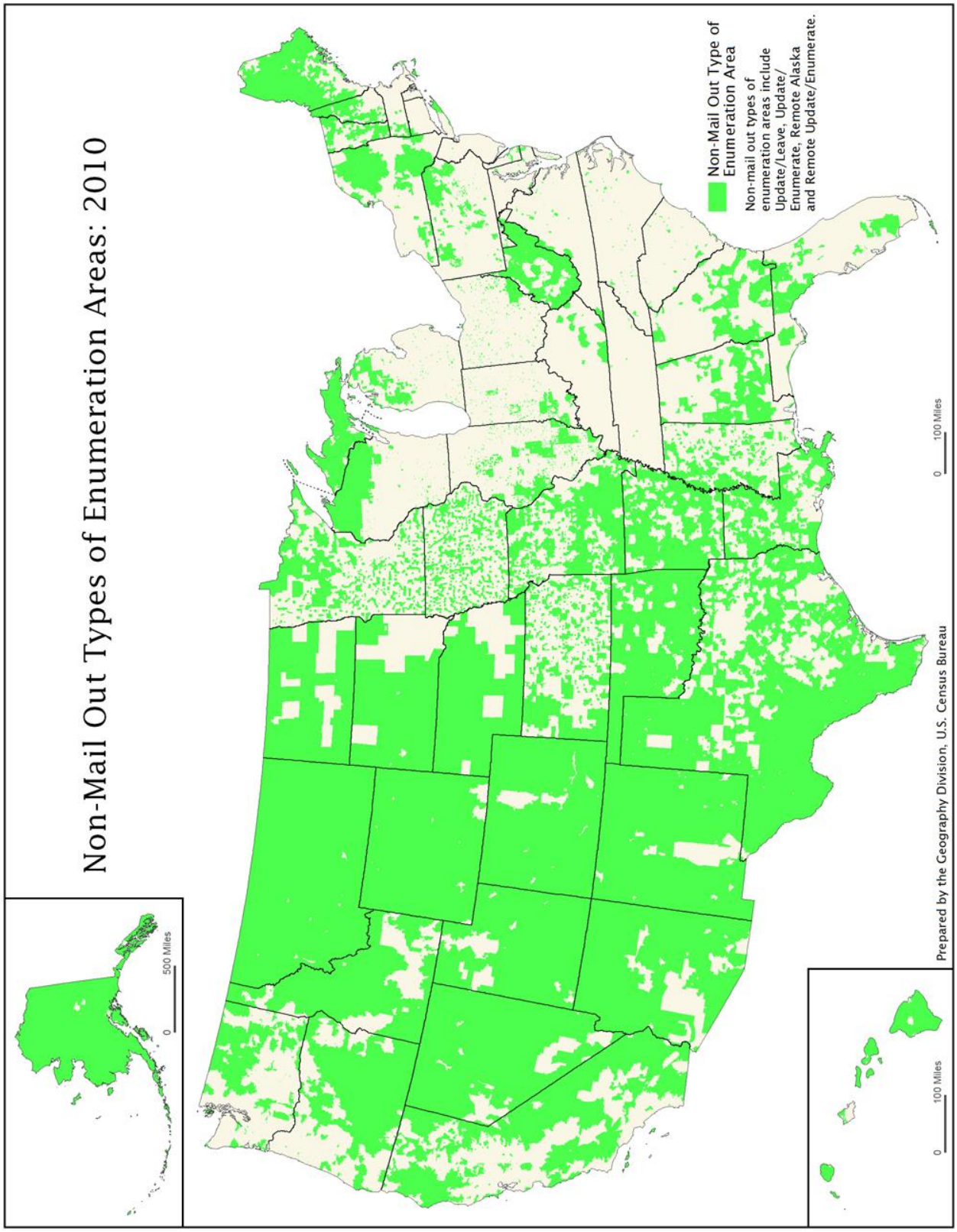

Figure 8. Non-Mail Out Types of Enumeration Areas: 2010

4 Recommendations

4.1 Recommendation for a Reengineered Address Canvassing

The Geography Division recommends that the Census Bureau move forward with plans for a reengineered address canvassing for the 2020 Census.

4.1.1 Basis of the Recommendation

Multiple address update processes have been implemented that are intended to minimize the areas where in-field address canvassing will be necessary by ensuring a complete, comprehensive MAF. These processes focus on partnering with local governments, maximizing use of the DSF, researching other sources of addresses to update and maintain the MAF, and developing new methodological approaches, including the use of aerial imagery and change detection.

The ongoing GSS-I Partnership Program has demonstrated that it can acquire data from local governments to supplement the twice-yearly DSF. Use of local government files, and other sources to update and maintain the TIGER road network, support geocoding new and existing MAF addresses. Partner files have a 76.6 percent match rate to the MAF and added 64,183 new addresses to date, therefore the program is effective at continual updates to the MAF.

Comparing the 2013 MAF to the 2010 Census address list demonstrated that address and housing unit counts in a majority of census blocks in the nation have not changed within these two sources. The census blocks identified cover 80.7 million housing units, 58.2 percent of the nation's addresses. Taken together with the results from DSF matching and MAF updates using partnership files suggest that many census blocks would not require canvassing in the field. Comparing the MAF to imagery also is an effective means of identifying change.

Statistical models can contribute to the identification of geographic areas for further in-office review and possible in-field verifications. Model results can provide an input to the review and decision-making process, leading to identification of census blocks for canvassing in the field.

The development of QIs has proven to be an effective way to measure and assess the quality of data in MAF/TIGER. More work, however, is needed to fully realize their potential for use in the Partnership Program, geographical analysis, and statistical modeling.

Third-party data were successfully matched to addresses in the MAF. Additional research and review of these sources are needed to further assess content and determine their most effective uses.

4.1.2 Other Information Applicable to the Recommendation

Results of each process show consistency but all aspects of the work have not yet been integrated. The GSS-I Partnership Program has not completed the acquisition and processing of

data nationwide (complete data acquisition is expected in fiscal year 2017). Researchers have not followed each census block through the decision flow (Figure 6) and have not compared block by block the output of each process. Some geographic areas currently identified by the models as candidates for in-field canvassing could be updated by applying data from local government files to the MAF during the years before the 2020 Census. Ultimately, the models could be used to identify areas in which the GSS-I Partnership Program should acquire partner files as source data for address updates. Likewise, QIs and imagery-based analysis and change detection can identify areas where updates to the MAF are needed, also informing the GSS-I Partnership Program. Identifying the appropriate sequence of analysis, actions, and file acquisition is needed for more effective analysis and decision-making. Over the course of the next year, further analysis of QIs, imagery-based analysis and change detection, statistical modeling, and results from the AVT will determine the percentage of HUs for in-office and in-field canvassing and result in more accurate and precise identification of the areas where in-field address canvassing will be necessary.

4.2 Further Recommendations to Support a Successful Reengineered Address Canvassing

To support a successful reengineered address canvassing, the Census Bureau should annually:

- Commit resources for continuing partnership activities to cover the whole nation where local address lists exist and to solicit and process partnership files for areas with the most change;
- Continue review of imagery to detect change and compare results with HU counts in the MAF. In addition, this work should be integrated with address and feature data-acquisition processes;
- Identify other potential sources of information for areas where a local government cannot provide acceptable data, such as third-party data and administrative records;
- Continue to develop overall measures of certainty about the quality and completeness of the MAF at multiple levels of geography, from nationwide MAF coverage to the census block level;
- Evaluate the optimal combination of statistical modelling, auxiliary data sources, and acceptable data metrics to determine areas of interest for imagery review; and
- Continue the development of web services to improve the quality of data that local governments provide and to ease the burden on local governments with consistent approaches for address data management.

5 Impacts

The recommendation to move forward with plans for a reengineered address canvassing has the following potential impacts:

- Continuous quality assessment and enhanced measurement of MAF/TIGER enables efforts and resources to be focused on specific areas most in need of improvement;
- Address list updates are distributed throughout the decade rather than relying solely on a nationwide field operation in the year before the census;
- The MAF is enhanced by processes using additional sources for validation and update, resulting in a higher quality address list;
- A reengineered address canvassing impacts design of other field operations. For example, where address list updating is planned as part of in-field questionnaire delivery and enumeration, canvassing that area could be conducted concurrently;
- More accurate and current data about addresses and land uses will support 2020 Census design decisions;
- Technology exists to conduct in-office verification that reduces the need for fieldwork;
- Process improvements result in a better understanding of partner-provided data and formats, leading to greater efficiency in handling partner data; and
- Improved quality, coverage, and currency of address and road data results in improved products for both Census Bureau internal customers and external customers such as government agencies, partners, and data users.

6 Benefits

The GSS-I Program comprises the activities covered in this report together with others that will have benefits beyond the decision for a reengineered address canvassing, including:

- Updates to the MAF/TIGER System throughout the decade improve address and feature coverage for intercensal surveys and estimates programs, leading to improved data collection and data dissemination;
- Stronger partnerships and sustained relationships with governments because of an ongoing partnership presence throughout the decade;
- Published guidance for partners' address and road data;
- Technology innovations that improve the ability to acquire, evaluate, process, and merge partner data with existing data;
- Improved business processes for address and road data management between data providers and the Census Bureau;
- Operational infrastructure and data acquisition processes and systems for acquiring data from state, local, and tribal governments that can be used for other programs;

- MAF/TIGER quality measures can serve as an influence for internal stakeholders' use of the data in survey samples;
- Enriched TIGER products from a continuous and broader source of road data and high quality address ranges; and
- The most current address list possible for 2020 LUCA.

7 Risks

The lack of a comprehensive, national address file that can be used to independently evaluate, validate, and supplement the MAF poses a risk to the certainty that the address list is complete. The DSF has been the primary source for address updates; however, it is dependent to some extent on updates from local governments, the same partners who are providing address information to augment current update processes. Third-party data and administrative records can mitigate this risk since some of these sources are developed from program registrants, mailing lists, and customer bases, but each is limited in its scope and none are as comprehensive as the MAF. Thus, while the MAF can be measured against various other sources to match, validate, and update a large proportion of addresses, there is not absolute certainty that the DSF, local partner files, third-party data sources, and administrative records can individually ensure appropriate coverage for the 2020 Census. Fieldwork to validate and verify addresses also contains the potential for error as individual canvassers can fail to find addresses to add to the list or may introduce error in the update process. Therefore, a mitigation strategy would be to build an independent, comprehensive national address file that local government partners and other stakeholders could access to keep up to date and to validate the accuracy of address information that the Census Bureau would use in its efforts to improve and maintain MAF/TIGER. Another mitigation strategy to ensure complete address coverage would be to continue conducting an in-field sampling of the nation for address data validation.

Change detection processes such as comparing multiple vintages of imagery can provide independent sources of information with which to measure the completeness of the MAF. Technology currently exists to detect changes to structures and roads from multiple vintages of imagery. Successful use of imagery, however, requires acquisition of high-resolution imagery and the technology to extract, process and analyze data. Parcel data and zoning information also can be used to detect change, but current, nationwide datasets are lacking at this time. If a methodology for measuring the completeness of data in MAF/TIGER cannot be developed, then the confidence in and certainty of the address list for any given geographic area is diminished.

The GSS-I Partnership Program has shown that for some local governments, address lists do not exist or are not available for use, or if available, do not meet published requirements and guidelines. Thus, there is a risk that for some geographic areas, the GSS-I Partnership Program will not be able to update the MAF through acquisition of local government address files. A similar risk exists if sufficient numbers of partners do not participate in the GSS-I Partnership

Program. If change detection processes suggest substantial change in the number of housing units in these geographic areas, and the DSF and other sources of addresses are not sufficient for updates, then an accurate and complete MAF may not be assured for some geographic areas. Canvassing in the field and update activities may also be required in such instances to validate and update the address list. The impact, if these risks are realized, could be reduced coverage by MAF/TIGER and a potential increase in the number of housing units and amount of area needed to be canvassed in the field.

8 Conclusion and Future Work

Between 2012 and 2014, the Geography Division researched and acquired information from the activities described in this report, resulting in the recommendation to move forward with plans for a reengineered address canvassing for the 2020 Census.

Quality Indicators provide measures of the fitness for use of data within MAF/TIGER and contribute to a greater understanding of, and confidence in, MAF/TIGER for use in decision-making. QIs will be further refined and regularly applied to assess the quality of areas in MAF/TIGER.

The GSS-I Partnership Program successfully acquired data representing 37 percent of housing units in the United States and ascertained that these data have utility for improving MAF by validating existing addresses, adding new addresses, and geocoding addresses that previously were not included in the Census. The GSS-I Partnership Program will continue to rely on high-quality partner data to update MAF/TIGER and will promote partner confidence in the reengineered address canvassing.

Third-party data and administrative records will be used to validate addresses in the MAF. These data sources may also be used as sources of address updates in the MAF and in conjunction with other data sources to determine where in-field canvassing may be necessary. Work will continue to identify third-party data and administrative records that could be used for MAF/TIGER update and to determine how to efficiently evaluate both types of data sets.

Research demonstrates high levels of housing unit and address consistency throughout the United States. Indications are that change detection through comparing the MAF to imagery and statistical modeling each can highlight areas to be canvassed in the field. Future activity will seek to automate MAF to imagery comparison for change detection. Statistical models will be refined to better identify areas for in-field canvassing.

While the research and results discussed in this report are sufficient for developing a recommendation for a reengineered address canvassing, work must continue throughout the decade, refining methods and outcomes. The next deliverable is a defined process that will determine the percentage of HUs for in-office and in-field canvassing. Inputs will include address QIs or specific sub-QIs at the census-block level; GSS-I Partnership Program data

through the first months of fiscal year 2015; identification of third-party data and administrative records that could be used for MAF/TIGER update; statistical models (once validated and improved with results of the AVT); and further MAF-to-imagery comparison (also using AVT results for comparison). Establishing the process for determining the percentage of HUs for in-office and for in-field canvassing will be followed by the development of an implementation plan identifying the optimal sequence of file acquisition, data analysis, and other activities for more effective decision-making and the geographical distribution of the specific areas to be canvassed in the field (implementation plan due September 2015). This work will proceed in conjunction with the development of the 2020 Census design.

Document Logs

Sensitivity Assessment

Verification of Document Content	
This document does not contain any: • Title 5, Title 13, Title 26, or Title 42 protected information; • Procurement information; • Budgetary information; and/or • Personally identifiable information.	
Document Author: Timothy F. Trainor, Chief, Geography Div.	Date:

Review/Approval

Document Review and Approval Tier: Program Document		
Name	**Area Represented**	**Date**
Timothy F. Trainor	Geography Division	
Lisa Blumerman	Decennial Census Programs	
Nancy A. Potok	Deputy Director	
John H. Thompson	Director	

Version History

Version	Date	Description
1.0	09-30-2014	Initial report
1.01	11-15-2014	Add alt text to graphics, update web links

Abbreviations and Glossary

Add..... A housing unit whose address was not on the Master Address File
Address canvassing..... The process by which the Census Bureau validates, corrects, or deletes existing Census Bureau addresses, adds missing addresses, and adds or corrects locations of specific addresses before a decennial census
Address list..... Collection of addresses from the MAF for a defined geographic area
Address listing ..... A field operation to develop the address list in areas with predominantly non-city-style mailing addresses
ACS..... American Community Survey
AVT ..... Address Validation Test
Built out ..... Refers to an area in which no additional housing or other development can occur due to lack of available land or zoning restrictions or both
CAUS..... Community Address Updating System
Census block ..... The smallest geographic unit for which the Census Bureau tabulates decennial census data
Census tract..... The small, relatively permanent statistical subdivisions of a county or equivalent entity with the primary purpose of providing a stable set of geographic units for the presentation of statistical data
City-style address..... An address that consists of a house number and street or road name; for example, 201 Main Street
DSF ..... United States Postal Service Delivery Sequence File
Feature ..... Any part of the landscape, whether natural (such as a stream or ridge) or human-made (such as a road or power line) that can be shown on a map. In this report, unless otherwise indicated, feature means road
Frame ..... The set of units from which a sample is drawn. The sampling frame used for a census is an address list
Geocode ..... A code used to identify a specific geographic entity
Geocoding..... The assignment of an address, structure, key geographic location, or business name to a location that is identified by one or more geographic codes
GSS-I ..... Geographic Support System Initiative
HU..... Housing unit
LUCA..... Local Update of Census Addresses
MAF..... Master Address File
MMVT ..... MAF Model Validation Test
Non-city-style address ..... A mailing address that does not use a house number and street or road name. This includes rural routes and highway contract routes, which may include a box number; post office boxes and drawers; and general delivery
Multi-unit structure ..... A building that contains more than one housing unit (for example, an apartment building)
PBC..... Partial Block Canvassing
QI ..... Quality Indicator
RFI ..... Request for Information
Special land use areas ..... Areas of unique land uses or residential characteristics such as large municipal parks, commercial airports, federal government installations and facilities, prisons, and college campuses
TIGER..... Topologically Integrated Geographic Encoding and Referencing
USPS..... United States Postal Service
ZINB ..... Zero-inflated Negative Binomial Regression model
ZIP ..... Zero-inflated Poisson regression model

Acknowledgments

Kaile H. Bower, Curtis A. Dunson, Scott P. Fifield, Gregory F. Hanks, Andrea G. Johnson, Douglas K. Lockhart, Michael R. Ratcliffe, Trudy A. Suchan, and Timothy F. Trainor contributed to the report.