Content and Coverage Quality of a Commercial Address List as a National Sampling Frame for Household Surveys

Timothy L. Kennel[1] and Mei Li[2]
[1]US Census Bureau, 4600 Silver Hill Road, Washington, DC 20233
[2]US Census Bureau, 4600 Silver Hill Road, Washington, DC 20233

Abstract
There has been much interest in the possible use of address-based lists from information resellers (IRs) as a potential frame for household surveys. In this study, we matched the InfoUSA file to the Census Bureau's Master Address File (MAF) for all 50 States and the District of Columbia. We use this matching to compare the coverage and content qualities of the InfoUSA file to the MAF. Additionally, we compared the InfoUSA file to a nationally representative sample of housing unit field enumerations.

Key Words: Residential Mailing Address List, Gross Coverage, USPS, Listing, Sampling Frame, Area Probability Sampling

1. Introduction

With the increasing number of cell phone only households, declining response rates to telephone surveys, and the increasing cost of field enumeration of housing units, many survey centers are drawn toward purchasing residential mailing address lists from information resellers (IRs), which commonly provide addresses for direct marketing. The content and coverage qualities of such purchased address lists is of great interest to those concerned with survey quality and methods.

The US Census Bureau maintains an inventory of addresses for the nation's living quarters. This inventory is called the Master Address File (MAF). It currently supports a number of survey and census operations at the US Census Bureau. However, due to legal restrictions on the MAF, some surveys may want to use alternative frames. Thus, the Census Bureau is interested in the possibility of using IR address lists as a potential sampling frame for some surveys. After comparing the cost, content, and coverage of a number of files from several IRs, the Census Bureau decided to explore the possibility of using an address file from InfoUSA as a potential sampling frame.

We received a file with over 360 million household records on it from InfoUSA. After unduplication, the file contained about 130 million housing unit records. The unduplicated InfoUSA file was matched to the MAF. In this paper, the match and nonmatch rates are used to measure the relative coverage of the unduplicated InfoUSA file. Since the MAF contained a nationally representative sample of approximately 2,800 tabulation blocks, which were completely listed, we were able to compare the InfoUSA file to the MAF as well as a ground listing of 2,800 blocks. In addition to measuring the relative coverage of the InfoUSA file, we also assessed the locatability of addresses on the InfoUSA file, the quality of unit designations, and the quality of geocodes on the InfoUSA file.

2. Background

2.1 Introduction

In 2006, a US Census Bureau team conducted evaluations of several IR files. The goals of the 2006 evaluations were to determine whether one or more of these IR files could meet the needs of the US Census Bureau to enhance data collection, reduce disclosure risks, and be used as an alternative source for developing address frames for household surveys. The results from these 2006 evaluations suggested that IR data might be able to provide adequate housing unit coverage for survey frames and to improve the MAF in some areas.

There is great interest in the possibility of using IR files as a primary sampling frame. Researchers from NORC, RTI, and Westat have all published journal articles or JSM proceedings papers evaluating the quality of IR files as a sampling frame. Each evaluation uses a slightly different methodology. In this section, we summarize several of the articles.

2.2 Literature Review

Most papers find that IR files are missing between 15% and 23% of the housing units on the ground (Dohrmann *et al*. 2006, Iannacchione *et al* 2007, O'Muircheartaigh *et al*. 2002, O'Muircheartaigh *et al*. 2006). The estimated coverage rates depend on the methods used to measure the rates as well as the IR vendor.

All studies consistently find that IR coverage is much higher in urban areas than in rural areas (Dohrmann *et al*. 2007, Iannacchione *et al*. 2007, O'Muircheartaigh *et al*. 2007). In looking further at covariates associated with coverage, O'Muircheartaigh and others also found that population density, median tract income, the percent of city-style addresses, and the Census Type of Enumeration Area were indicators of the coverage of the ADVO frame (O'Muircheartaigh *et al*. 2007). For areas needing supplementation different enhancements including the half-open interval; the Waksberg method; using augmented addresses in areas with simplified delivery; the Check for Housing Units Missed (CHUM); and area frame listing are suggested.

Although most studies focus on housing unit coverage regardless of the occupancy status, Iannacchione and others found that coverage of occupied housing units is much higher than coverage of vacant housing units (Iannacchione *et al* 2007). For North Carolina, they found that an IR file had 82.1% of the housing units in the state; but 95.7% of the occupied housing units in the state.

Sylvia Dohrmann and others found some differences among IR vendors. In looking at the cost and coverage of files from Donnelley Marketing (InfoUSA), ADVO, CIS, and Anchor Computer, they found that vendors with a Computerized Delivery Sequence (CDS) license have better coverage of the nation at the housing unit level, but tend to be more expensive than vendors with the DSF2 license. Vendors with a CDS license must meet USPS coverage requirements, but receive some new DSF addresses that are not given to vendors with a DSF2 license (Dohrmann *et al*. 2006).

Although the US Census Bureau has limited experience with using IR files as a sampling frame, it has a long history of creating and using address lists, often supplemented with additional frames (See US Department of Labor 2006).

3. Methods

3.1 Data

In this paper, we compared three sampling frames:

- the InfoUSA frame, a sampling frame constructed using InfoUSA household data,
- the MAF-based frame, a sampling frame constructed from applying filtering rules to the MAF, and
- the Ground Frame, a sampling frame obtained from field block canvassing in Frame Assessment for Current Household Surveys - National Evaluation Sample blocks.

This section describes these three frames in greater detail.

3.1.1 InfoUSA

According to their website (infousa.com), "InfoUSA is the leading provider of business and consumer information products, database marketing services, data processing services and sales and marketing solutions." InfoUSA's national database of housing units was developed from collecting over 4,300 telephone directories across the country. They add new households from real estate and marketing sources. As a DSF2 licensed vendor, their addresses are standardized and verified using the Delivery Sequence File (DSF), the United States Postal Service's (USPS) national file of mail delivery points. Because InfoUSA collects addresses from diverse sources, other IRs often use their addresses as a source for city-style addresses in areas with simplified mail delivery.

We received a database with over 360 million household records from InfoUSA in February 2008. Since the national InfoUSA database contained both current and historic households, each address on InfoUSA may have multiple household records. The abundance of historic households and other duplicates explains why the number of records on the InfoUSA database is much greater than the national housing unit estimate of approximately 128 million housing units. The Census Bureau unduplicated the InfoUSA household database. When multiple households were listed for the same housing unit, we kept the household with the most recent source date and removed the others. We also removed duplicate households within the same housing unit. After unduplication, the February 2008 file contained 133,058,311 housing units. Since the Census Bureau's unduplication rules may differ from InfoUSA's unduplicaton rules and the unduplication may impact match rates, the results in this study may not exactly mirror independent findings by other researchers.

The InfoUSA household database has numerous variables collected from a variety of sources. All household records contain address variables, modeled income, the household person count, and the name, age, race, and the date of birth of up to four adults in the household. The database also contains additional variables such as telephone numbers, block group geocodes, GPS coordinates, housing values, the number of bathrooms in the house, the number of square feet in the house, whether anyone in the household is interested in beauty products, and tenure status for many households.

3.1.2 Master Address File

The MAF is a Census Bureau database intended to contain addresses for all living quarters in the United States. The MAF is continually updated through a series of

different operations. The MAF contains housing units collected from Census 2000 along with semiannual updates from the US Postal Service's Delivery Sequence File (DSF) and field listing operations. The DSF is the largest source of new addresses on the MAF outside the census years. This study used the MAF extracts delivered in July 2008, which contained addresses from the March 2008 DSF delivery. In terms of timing, the July 2008 MAF delivery roughly corresponds to the February 2008 InfoUSA delivery.

Table 1: Master Address File Summary Counts

Category	Total
Unfiltered MAF	171,657,062
Filtered MAF	136,122,595
Complete City-Style	131,627,085
Match to DSF	120,570,506
Match to Census 2000	115,506,951
Post-Census DSF Adds	15,263,433

Since the MAF is a cumulative inventory of all housing units, past and present, it contains numerous records that cannot be found on the ground and sometimes multiple records for the same housing unit if different sources to the MAF provided different forms of the address. A filter is applied to the MAF to determine which records on the MAF should be eligible for sample. Users of the MAF can determine their own filter based on their needs and requirements. Table 1 shows that of the 170 million addresses on the MAF, 136,122,595 passed the filter used for this project. For this report we call the set of records that passed the filter the Filtered MAF. The filter is subject to misclassification errors.

Although use of the MAF is limited to the Census Bureau, Table 1 shows that over 88% of the Filtered MAF addresses were found on the DSF. Thus, the properties of the MAF are very similar to the properties of the DSF, the primary source of addresses for many IR files. For this reason, survey samplers and managers outside the Census Bureau have been interested in the coverage properties of the MAF for quite some time.

3.1.3 Ground Frame

The National Evaluation Sample (NES) is a nationally representative sample of blocks across the United States. The portion of the NES used for this project was most representative of the ground in February 2008, the InfoUSA and MAF reference dates. The field listing of all housing units in 2,800 tabulation blocks occurred between January and March 2008. These listings updated the July 2008 MAF. In this report, the NES listings are referred to as the Ground Frame because field representatives on the ground have verified them. For more information on the NES see Loudermilk (2009).

3.2 Matching

After standardizing and unduplicating the InfoUSA database, the Census Bureau assigned tabulation block geocodes to the InfoUSA file. Then, the unduplicated InfoUSA file was matched to the Unfiltered MAF. We used the SAS MATCH macro developed by the US Census Bureau to do the matching. We used the outcome from this matching process to produce various measures of quality in this report. The match and nonmatch rates are used to measure the relative coverage of the unduplicated InfoUSA file compared to the MAF. Since the MAF contained updates from the NES, restricting our analysis to the

2,800 NES tabulation blocks allowed us to compare the InfoUSA file to the Ground listing.

All matching was blocked within the first three digits of the ZIP code. For example, addresses in ZIP code 20010 could be matched to addresses in ZIP code 20009, but would not be matched to addresses in ZIP code 20233. Within each three digit ZIP code, we attempted to match each address in three passes. In the first pass, addresses were matched at the unit level. If addresses didn't match in the first pass, a second pass was made to match the address at the Basic Street Address (BSA) level. A BSA is an address without a unit designation. The BSA often represents a single structure with one or more housing units in it, although there are exceptions. For example mobile homes in some trailer parks share the same BSA, but are distinguished by a unit designation. Finally, a third pass was developed to match rural-style addresses that didn't match in the previous two matches. Table 2 shows how many addresses matched during each pass. Any InfoUSA record that matched to the common Unfiltered MAF during any one of the three passes was considered a match.

Table 2: Summary of matching InfoUSA to the common UnFiltered MAF

Universe	Total	Percent
Pass 1: Unit level match	120,826,384	91.0%
Pass 2: BSA level match	3,058,464	2.3%
Pass 3: Rural match	14,991	0.0%
Nonmatch	8,958,974	6.7%
Total	132,858,813	100%

In this report, we present unit level coverage rates under two different frame creation processes. In the first scheme, a sample of the IR units is selected and interviewed. Under the second sampling scheme, the InfoUSA files are unduplicated at the BSA level. After creating a file of basic street addresses, the BSA's are sampled. At the time of interview, the BSA's are listed and a subsample of units within the listing are interviewed. To determine if a unit on the Filtered MAF matched to a BSA on the InfoUSA file, we defined a BSA identifier on both files. Then we looked to see if any unit in the MAF BSA matched to a unit on the InfoUSA file. If at least one unit in the MAF BSA matched to the InfoUSA file, then we determined that all the units in the MAF BSA matched to InfoUSA. Assuming that listers don't make any mistakes listing the units, BSA sampling should provide better coverage of units. On the other hand, the time and resources needed to list multi-unit structures at the time of interview can be costly.

All standard error estimates were calculated using the delete a group jackknife method in SUDAAN to take the stratification, clustering, and sampling into account.

4. Results

4.1 Coverage Analysis

In this section we report and discuss estimates of net coverage, gross undercoverage, and gross overcoverage. Using match rates to compare the InfoUSA files to the MAF and the Ground Frame at the national, regional, and state level, we find that InfoUSA undercoverage is greater than MAF undercoverage and little coverage improvement can be made to the MAF by adding InfoUSA units to a MAF-based frame. On the other hand, the InfoUSA file is comparable to other national sampling frames, provides a wealth of

auxiliary data about households, and covers BSAs for all regions within Office of Management and Budget coverage guidelines.

In September 2006, the US Office of Management and Budget released standards and guidelines for statistical surveys. Standard 2.1 and the following guidelines deal with the coverage of sampling frames. Of particular interest, Guideline 2.1.3 claims, "Coverage rates in excess of 95 percent overall and for each major stratum are desirable. If coverage rates fall below 85 percent, conduct an evaluation of the potential bias." It is important to recognize that the target population for many household surveys is the non-institutionalized US population, whereas this report is primarily focused on housing unit coverage. The coverage and match rates in this report may differ from person-level coverage rates.

4.1.1 Undercoverage

Comparing the MAF to the InfoUSA file shows that 16.1% of the complete city-style addresses on the Filtered MAF did not match to a unit on the InfoUSA file, and 7.9% did not match to a BSA on InfoUSA as seen in Table 12. Comparing the ground frame to InfoUSA shows similar results: an estimated 15.5% (s.e. 0.7) of the complete city-style addresses on the ground could not be found on InfoUSA and 8.4% (s.e 0.5) of the complete city-style addresses on the ground did not match to a BSA on InfoUSA as seen in Table 3. As expected, these relative coverage rates vary by region and state. These coverage rates tend to agree with findings highlighted in the literature review.

Table 3: Nonmatch rates by characteristics

Characteristic	Estimated Total	Estimated addresses on the ground that do not match to InfoUSA unit	Estimated addresses on the ground that do not match to InfoUSA BSA	Estimated addresses on the ground that do not match to MAF
Complete Address	122,468,317	15.5% (0.7)	8.4% (0.5)	5.4% (0.5)
Permit Issuing	115,778,383	15.9% (0.7)	10.5% (0.5)	5.7% (0.5)
Urban	94,094,930	14.0% (0.8)	5.6% (0.5)	4.5% (0.6)
100% of addresses in block are city-style	71,364,583	9.4% (0.7)	4.7% (0.5)	4.1% (0.7)
Mobile Homes	7,136,171	40.0% (2.9)	31.2% (2.8)	20.2% (2.8)
Nation	125,444,526	17.5% (0.7)	10.5% (0.5)	6.4% (0.5)

Table 3 shows estimated relative coverage rates comparing units on the ground frame to InfoUSA units, units on the ground to InfoUSA BSAs, and units on the ground to the Filtered MAF for several domains. As we see, InfoUSA coverage is much better as a BSA sampling frame used with unit listing at the time of interview. We also see that InfoUSA strengths and weaknesses are similar to the strengths and weaknesses of the MAF, but the magnitude of InfoUSA undercoverage is greater than MAF undercoverage. Rural areas, tabulation blocks with less than 100% complete addresses, and tabulation blocks not covered by a building permit office should benefit the most from coverage improvement. Coverage of mobile homes is a challenge for InfoUSA and the MAF.

Every six months, all city-style addresses from the DSF are merged into the MAF. Table 4 shows that 87.0% (s.e. 0.8) of the estimated 122 million city-style addresses found on

the Ground frame matched to the DSF, while 84.5% (0.7) matched to the InfoUSA file. IRs that heavily rely on the DSF are likely not to have much better coverage than InfoUSA, which is a DSF2 license vendor.

Table 4: Estimated percent of complete addresses on the ground that were found on the DSF and InfoUSA frames

On DSF	On InfoUSA		Total
	Yes	No	
Yes	80.0% (0.8)	6.9% (0.3)	87.0% (0.8)
No	4.5% (0.4)	8.6% (0.6)	13.0% (0.8)
Total	84.5% (0.7)	15.5% (0.7)	122,468,317

Table 5 shows counts for complete city-style addresses on the Filtered MAF by the DSF status on the Filtered MAF. As we see, the Filtered MAF contains 11 million complete city-style addresses that could not be found on the DSF. Many of these units could not be found on InfoUSA. We also see that InfoUSA shares many of the complete city-style addresses on the Filtered MAF that are classified as residential delivery with the Filtered MAF. However, compared to the MAF, coverage of the DSF records that the USPS has classified as "exclude from delivery statistics" is low. Fortunately, many of these units are either under construction, unoccupied or otherwise not yet eligible for sample [see Loudermilk and Kennel (2005) and Martin and Loudermilk (2008)].

Table 5: Count of complete addresses on Filtered MAF by DSF status and BSA match status

DSF status on Filtered MAF	Matches to BSA on InfoUSA		Total
	Yes	No	
Not on DSF	6,773,863	4,282,713	11,056,576
DSF Residential	110,566,673	3,295,716	113,862,389
DSF Commercial	241,406	33,545	274,951
DSF Exclude from Delivery Statistics	3,644,745	2,788,421	6,433,166
Total	121,226,687	10,400,395	131,627,082

4.1.2 Overcoverage

Comparing the unduplicated InfoUSA files to the MAF and Ground Frame shows that 5.7% of the complete city-style addresses on the InfoUSA database could not be found on the Unfiltered MAF. Table 13 also shows that an additional 7.6% of the complete city-style addresses on the InfoUSA database matched to MAF addresses that failed the filter. These rates also varied by region and state.

Areas with overcoverage tend to overlap with areas having undercoverage. Rural blocks, blocks with few complete addresses, and blocks not covered by a building permit office tend to have higher nonmatch rates than urban blocks, blocks with many complete addresses, and blocks covered by a building permit office. The same trend also holds for InfoUSA records that match to invalid records on the MAF. We also found that the less recently an InfoUSA record has been updated, the more likely it is to be overcovered. Moreover, records with reported tenure status (owner or renter) on the InfoUSA files are more likely to match to the Filtered MAF than records with uncertain tenure status.

4.2 Content Analysis

We investigated the locatability of addresses in the unduplicated InfoUSA files as well as the presence and accuracy of unit designations in the InfoUSA files. Based on the percent

of complete addresses on InfoUSA, we conclude that most addresses on InfoUSA should be locatable, however, InfoUSA codes suggest that up to 8.5% of the InfoUSA records may be unlocatable. In terms of unit designations, it appears that the presence and accuracy of unit designations is high; however, InfoUSA is disproportionably missing units within BSAs.

4.2.1 Address Locatability

A complete address is an address with both a house number and a street name. Post office box addresses, rural route addresses, and addresses with a missing house number or street name are considered incomplete addresses. Without additional information, incomplete addresses tend to be more difficult to locate than complete addresses. The challenges in locating incomplete addresses can inflate field costs and result in interviewing the wrong unit. If an address can not be found, it can lead to decreases in sample size and coverage.

InfoUSA and the Filtered MAF have a similar number of complete addresses, but the Filtered MAF has more incomplete addresses than InfoUSA. InfoUSA classified 131,432,494 of their addresses as complete; whereas there are 131,627,082 complete addresses on the Filtered MAF. On the other hand InfoUSA classified 1,426,319 of their addresses as incomplete; whereas the Filtered MAF has 4,495,513 incomplete addresses.

Table 6: Count of InfoUSA housing units by quality of address for complete and incomplete addresses

Quality of address[1]	Complete	Incomplete	Total
Accurate	116,274,296	136,056	116,410,352
Probably Deliverable	3,238,815	1,499	3,240,314
Deliverability Questionable	581,596	0	581,596
Probably Undeliverable	11,337,787	1,288,764	12,626,551
Undeliverable	0	0	0
Total	131,432,494	1,426,319	132,858,813

Some of the complete addresses may not be locatable. According to InfoUSA classifications, up to 11,337,787 complete addresses may be unlocatable (see Table 6). Auxiliary data such as geocodes and phone numbers can be used to help locate households. Table 7 shows that a majority of the 1,426,319 incomplete addresses have a phone number on the InfoUSA file.

Table 7: Count of InfoUSA housing units by presence of phone number for complete and incomplete addresses

Presence of phone number	Complete	Incomplete	Total
Present	84,625,352	816,575	85,441,927
Absent	46,807,142	609,744	47,416,886
Total	131,432,494	1,426,319	132,858,813

Because of their small geographic size, ZIP+4 codes can also be used to locate housing units. Table 8 shows that 120,232,262 addresses have a ZIP+4 code, but only 137,555 are incomplete addresses, adding very little to the locatability of incomplete addresses. Although larger than ZIP+4 areas, InfoUSA assigned a block group geocode to 1,375,075 of the 1,426,319 incomplete addresses

[1] This variable was defined and provided by InfoUSA.

Table 8: Count of InfoUSA housing units by presence of ZIP+4 for complete and incomplete addresses

Presence of ZIP+4	Complete	Incomplete	Total
ZIP+4 Present	120,094,707	137,555	120,232,262
ZIP+4 Not Present	11,337,787	1,288,764	12,626,551
Total	131,432,494	1,426,319	132,858,813

As seen in Table 12, the percent of incomplete addresses on the InfoUSA file varies by state. For example, 85.3% of the 730,319 addresses on the InfoUSA files are complete in West Virginia, but 98.6% of the 9,944,989 InfoUSA addresses in Texas are complete.

Even though InfoUSA has a ZIP code, phone number, and rough geocodes for many incomplete addresses, the InfoUSA database lacks specific identifying data needed to locate incomplete addresses. Costly clerical and field operations could be developed to locate some of the addresses with incomplete addresses.

4.2.2 Unit Designations

The unit designation is a part of an address that differentiates one housing unit from another housing unit within a basic street address. For example, "Apt 1C," "Unit 3," "Lot 2," and "Trailer 9" are examples of unit designations. Although BSAs tend to correspond to multiunit structures, this is not always the case. For example, mobile homes in the same trailer park may be distinguished by their unit designation. Unit designations are used to differentiate housing units during sampling, to prepare sample materials, and to find sample units.

Table 9 shows the percentage of Filtered MAF units that match to the InfoUSA files by presence of unit designation on both files. Since most housing units in the nation are detached single units, 74.6% of units on the MAF and 64.4% of units on InfoUSA do not have unit designations.

Many of the Filtered MAF records that could not be matched to InfoUSA end up being units within multi-units, indicating that InfoUSA tends to miss units within BSAs. Among the matches, we see that there aren't many records that have a unit designation on the Filtered MAF, but are missing a unit designation on InfoUSA. Only 0.1% of units on the Filtered MAF do not have a unit designation, but do have a unit designation on InfoUSA. Vice versa, only 0.8% of the units on the Filtered MAF have a unit designation on the Filtered MAF, but lack a unit designation on InfoUSA.

Table 9: Percent of Filtered MAF units that match to InfoUSA by presence of unit designation

Presence of Filtered MAF unit designation	Presence of InfoUSA unit designation		Address not found on InfoUSA	Total
	Present	Absent		
Present	15.8%	0.8%	8.8%	25.4%
Absent	0.1%	64.4%	10.0%	74.6%
Total	15.9%	65.2%	18.8%	136,122,595

Focusing on the units on the Filtered MAF that were not covered by the InfoUSA files, we see that many of them have a unit designation on the Filtered MAF. Thus, we conclude that much of the InfoUSA undercoverage is concentrated in units within multiunit structures.

Table 10 confirms that InfoUSA tends to miss units within multi-unit structures. For example, we found 8,704,561 units on the Ground frame that were in BSAs on InfoUSA that had two or few units than the Ground Frame count. Indeed, there are more units in BSAs with an undercount on InfoUSA than there are units in BSAs with an overcount on InfoUSA.

Table 10: Estimated units on ground difference in BSA counts between InfoUSA and ground

Size of structure on ground	InfoUSA had					Nonmatch
	2+ Fewer Units	One fewer Unit	Equal	One More Unit	2+ More Units	
Single Unit			83,741,547	425,665	106,942	13,227,607
Small Multi (2-4)	358,292	2,056,093	1,456,507	2,058,093	21,330	2,929,167
Large Multi (5+)	8,346,270	1,426,618	2,086,651	3,568,032	910,147	5,782,052
Total	8,704,561	3,482,678	87,284,705	6,051,789	1,039,419	21,938,827

Overall, the InfoUSA files have unit designations for units that should have unit designations. Many of the InfoUSA records that could not be found on the Filtered MAF have a unit designation on the MAF. Although the quality of unit designation data appears to be favorable, the InfoUSA file tends to undercover some units within large multi-units.

4.2.3 Geocoding

Geographic codes, or geocodes, describe the geographic location of buildings. They are often used in sampling to define clusters and primary sampling units. They are also used in the field to find sample units. Thompson and Turmelle (2004) and Turmelle, Rodrigue, and Thompson (2005) discuss how the Canadian Labor Force Survey uses the outcome of the geocoding process to determine which geographic areas need coverage improvements. Geocodes also play an integral role in data collection. During estimation, geographic codes can be used to define cells for nonresponse weighting adjustments. Thus, the presence and accuracy of geographic codes is advantageous for surveys that use census geography for sampling and weighting.

We investigated the success of our internal TIGER® geocoding system. We assigned state, county, tract, and tabulation block geocodes to 113,997,670 (85.8%) of the 132,858,813 records on the unduplicated InfoUSA file. Of those TIGER® geocoded records on InfoUSA, 90.3% had the exact same state, county, tract, and tabulation block as found on the MAF. The remaining 9.7% did not match to the MAF, matched to ungeocoded records on the MAF, or disagreed with the geocodes on the MAF.

The unduplicated InfoUSA files contain block group geocodes for nearly all units. Most of these geocodes agree with the MAF geocodes. Table 11 shows that only 9,508 records on InfoUSA lack geocodes at the block group level. InfoUSA assigned block group

geocodes to 123,667,822 units based on the address of the unit. Almost 110 million of the InfoUSA records that were geocoded to a block group at the site level had the same state, county, tract, and block group geocodes on the MAF. A little over eight million additional InfoUSA records were geocoded based on their ZIP code, but these geocodes less frequently agree to the MAF.

Table 11: Count of InfoUSA units that matched to the Unfiltered MAF by geocoding agreement and InfoUSA geocoding source

InfoUSA geocoding source	Geocodes Agree	Geocodes disagree or blank on MAF	Address not found on Unfiltered MAF	Total
Site level	109,914,103	8,649,880	5,103,839	123,667,822
ZIP +4 centroid	151,061	54,257	19,795	225,113
ZIP +2 centroid	790,127	2,781,629	360,152	3,931,908
ZIP centroid	265,364	1,287,600	3,471,498	5,024,462
Ungeocoded	0	5,818	3,690	9,508
Total	111,120,655	12,779,184	8,958,974	132,858,813

5. Limitations

Although we used probabilistic matching software tailored to match large files by address, there are some known limitations to the matching. First, address with complex house number structures made matching in some areas difficult. For this reason, nonmatch rates in Hawaii and Queens, New York may be inflated. Second, match rates for incomplete addresses were low due to the difficulty with matching such addresses. The low match rate for incomplete addresses likely reflects the difficulty in matching such addresses rather than true undercoverage on the InfoUSA database. Third, although the InfoUSA file was directly matched to the unfiltered MAF, the unfiltered MAF was not explicitly matched to the InfoUSA file. Presumably, directly matching the unfiltered MAF to the InfoUSA file would result in slightly lower nonmatch rates at the unit level, but negligible differences at the BSA level. Lastly, the focus of this paper is on housing unit coverage rather than occupied housing unit coverage. In 2007, Iannacchione and others found that coverage of occupied housing units was much better than coverage of all housing units for their frame. All of these caveats lead us to conclude that our undercoverage and overcoverages rates may be slightly inflated.

6. Conclusion

In this paper, we investigated the content quality and gross coverage of the unduplicated InfoUSA files. We estimated InfoUSA gross undercoverage of complete city-style addresses to be 15.4%. However, if we collapsed the InfoUSA file to the BSA level and listed units at the time of interview, we estimate an undercoverage rate to be 8.4%. We estimate the gross overcoverage of the InfoUSA file to be about 14%. Our results tend to agree with the findings in papers by NORC, RTI, and Westat. We found some strengths and weaknesses of the InfoUSA database as a sampling frame. Coverage of BSAs and of complete city-style addresses on the DSF are two of InfoUSA's strongest areas of coverage

Whether the InfoUSA files can be used as a sampling frame should depend on acceptable coverage requirements, the budget for frame creation, and appropriate frame

enhancements. The coverage of the InfoUSA database is better for BSAs than for units; suggesting that the InfoUSA database is a much more attractive sampling frame when used to select a BSA sample paired with BSA listing at the time of interview. Further improvements can be made through listing or coverage improvements in areas where coverage does not meet coverage requirements.

In nearly every area of address quality and coverage considered in this paper, the MAF is superior to the InfoUSA database. Moreover, since MAF coverage strengths and weaknesses tend to be parallel with InfoUSA strengths and weaknesses, the InfoUSA database does not fill in the differential coverage gaps on the MAF. This study focused on coverage rates of housing units. The effect of undercoverage on survey estimates, coverage rates for occupied housing units, and the relationship between undercoverage and response are needed areas for future research.

Table 12: Summary of Undercoverage by State

State	Filtered MAF Count	Relative Undercoverage					
		Complete City-Style Addresses			DSF records on MAF		
		Percent of Filtered MAF	Does not match to InfoUSA	Does not match to InfoUSA BSA	Percent of Filtered MAF	Does not match to InfoUSA	Does not match to InfoUSA BSA
AK	300,314	79.5%	22.1%	15.0%	62.2%	8.0%	3.2%
AL	2,344,561	94.2%	18.6%	11.8%	84.5%	12.6%	7.1%
AR	1,400,607	93.3%	16.5%	11.5%	79.8%	9.3%	5.5%
AZ	2,892,840	96.4%	20.2%	9.5%	87.3%	14.9%	6.4%
CA	13,771,954	99.4%	14.3%	5.2%	94.2%	11.7%	3.8%
CO	2,247,606	98.5%	13.6%	7.8%	88.7%	8.5%	4.7%
CT	1,505,128	99.7%	11.6%	4.2%	94.1%	8.4%	3.4%
DC	302,757	100.0%	18.3%	1.9%	93.6%	13.6%	1.2%
DE	452,625	95.4%	20.3%	14.6%	86.9%	15.9%	11.1%
FL	9,269,230	99.1%	13.6%	6.5%	94.0%	10.5%	5.2%
GA	4,311,602	96.6%	18.5%	11.5%	89.1%	14.2%	8.4%
HI	519,458	95.6%	52.7%	34.2%	81.2%	48.4%	32.2%
IA	1,415,776	99.3%	12.9%	7.8%	91.0%	9.1%	5.1%
ID	670,923	94.6%	16.9%	11.8%	82.7%	10.3%	6.5%
IL	5,581,433	98.5%	18.8%	5.7%	90.2%	13.4%	4.0%
IN	2,944,143	98.5%	12.1%	7.1%	92.4%	8.8%	4.8%
KS	1,289,908	97.3%	13.2%	8.2%	89.4%	8.9%	4.8%
KY	2,040,286	92.5%	17.8%	11.6%	81.7%	11.9%	6.9%
LA	2,131,166	96.8%	14.6%	8.6%	85.2%	9.0%	4.3%
MA	2,851,798	99.4%	14.8%	5.3%	91.5%	10.3%	3.9%
MD	2,427,142	99.6%	12.3%	4.5%	94.6%	9.4%	3.3%
ME	748,741	81.2%	26.1%	16.7%	65.3%	15.4%	7.9%
MI	4,736,058	98.4%	10.2%	6.3%	91.2%	5.9%	3.7%
MN	2,418,292	95.8%	11.9%	6.6%	89.2%	8.4%	4.0%
MO	2,843,415	93.0%	14.7%	9.3%	85.4%	10.2%	5.9%
MS	1,409,465	94.3%	22.0%	15.0%	82.3%	15.1%	9.0%
MT	480,194	89.3%	23.8%	18.1%	72.4%	13.6%	9.0%
NC	4,451,577	96.0%	16.5%	12.0%	85.2%	10.5%	7.2%
ND	325,839	92.6%	17.5%	11.2%	76.1%	9.8%	4.6%
NE	833,671	96.0%	12.2%	7.7%	87.9%	8.4%	4.9%
NH	633,604	92.8%	24.0%	15.6%	78.2%	15.5%	9.2%
NJ	3,669,998	99.7%	19.2%	5.1%	92.8%	14.8%	4.1%
NM	923,362	89.1%	24.3%	15.0%	75.2%	16.4%	8.0%

State	Relative Undercoverage						
	Filtered MAF Count	Complete City-Style Addresses			DSF records on MAF		
		Percent of Filtered MAF	Does not match to InfoUSA	Does not match to InfoUSA BSA	Percent of Filtered MAF	Does not match to InfoUSA	Does not match to InfoUSA BSA
NV	1,176,160	98.5%	15.9%	7.4%	91.7%	12.0%	5.7%
NY	8,323,338	97.6%	24.6%	6.6%	83.1%	14.5%	4.0%
OH	5,345,829	99.3%	9.4%	4.7%	93.9%	6.5%	3.1%
OK	1,739,215	86.4%	15.0%	10.0%	77.6%	9.5%	5.3%
OR	1,701,664	98.8%	12.6%	6.6%	91.0%	8.7%	4.0%
PA	5,800,193	94.1%	15.9%	7.6%	86.6%	11.4%	4.9%
RI	477,107	99.5%	17.6%	4.3%	92.3%	13.0%	3.4%
SC	2,230,942	96.1%	18.4%	11.6%	86.7%	13.1%	7.8%
SD	372,423	91.0%	17.4%	12.0%	77.4%	9.7%	5.4%
TN	2,955,386	97.5%	15.2%	9.4%	90.7%	11.1%	6.6%
TX	10,404,972	95.0%	17.4%	8.4%	88.5%	13.9%	5.8%
UT	982,481	97.1%	13.4%	8.2%	89.2%	8.9%	5.2%
VA	3,400,549	95.8%	12.6%	6.9%	89.1%	8.9%	4.1%
VT	328,880	85.9%	31.5%	22.3%	63.5%	16.6%	9.4%
WA	2,904,631	98.8%	14.7%	6.7%	92.1%	11.1%	4.2%
WI	2,653,674	99.2%	13.3%	8.7%	91.3%	9.0%	5.5%
WV	922,577	63.4%	23.5%	17.3%	53.2%	14.4%	9.3%
WY	257,101	93.8%	23.0%	15.4%	74.1%	13.2%	7.2%
Total	136,122,595	96.7%	16.1%	7.9%	88.6%	11.4%	5.1%

Table 13: Summary of Overcoverage and Geocoding by State

State	Relative Overcoverage				InfoUSA Un-geocoded Rate (block group - site level)	InfoUSA units not geocoded by TIGER® (tabulation block)
	InfoUSA Count	Complete City-Style Addresses				
		Percent of InfoUSA	Does not match to Unfiltered MAF	Does not match to Filtered MAF		
AK	257,225	97.3%	12.4%	21.4%	12.8%	23.6%
AL	2,262,601	98.4%	7.2%	17.7%	10.8%	20.9%
AR	1,427,021	98.2%	9.3%	19.6%	13.5%	22.5%
AZ	2,566,969	99.6%	5.8%	11.4%	5.6%	17.8%
CA	13,480,818	99.8%	3.1%	10.6%	3.0%	9.2%
CO	2,287,175	99.7%	6.2%	13.5%	4.5%	12.9%
CT	1,601,939	99.3%	5.0%	10.8%	2.5%	3.5%
DC	287,183	99.7%	2.0%	10.9%	1.3%	1.8%
DE	426,519	97.8%	7.0%	15.0%	10.8%	23.2%
FL	9,288,507	99.6%	4.1%	11.3%	4.8%	13.1%
GA	4,147,236	98.6%	6.8%	15.5%	10.5%	24.3%
HI	440,829	99.2%	35.0%	42.0%	11.2%	35.0%
IA	1,407,327	99.2%	4.3%	10.3%	6.5%	10.9%
ID	650,977	98.3%	8.3%	16.0%	12.4%	25.1%
IL	5,202,998	99.6%	3.6%	10.8%	3.7%	12.7%
IN	2,943,909	99.2%	4.7%	10.8%	7.2%	15.9%
KS	1,267,026	99.3%	4.2%	11.9%	6.4%	18.3%
KY	2,014,241	98.0%	8.9%	19.4%	13.6%	21.5%
LA	2,147,192	99.4%	6.6%	15.0%	6.0%	11.7%
MA	2,992,957	99.1%	4.6%	11.1%	3.3%	4.0%
MD	2,402,996	98.9%	3.3%	9.1%	4.5%	7.9%

State	Relative Overcoverage				InfoUSA Un-geocoded Rate (block group - site level)	InfoUSA units not geocoded by TIGER® (tabulation block)
	InfoUSA Count	Complete City-Style Addresses				
		Percent of InfoUSA	Does not match to Unfiltered MAF	Does not match to Filtered MAF		
ME	663,223	93.9%	9.9%	24.5%	17.4%	25.4%
MI	4,832,868	99.4%	5.1%	11.2%	4.8%	14.0%
MN	2,383,579	99.3%	4.7%	12.1%	8.4%	16.9%
MO	2,713,668	98.8%	5.1%	13.8%	10.2%	18.5%
MS	1,351,853	97.8%	8.7%	18.5%	13.8%	18.5%
MT	432,813	98.4%	10.7%	21.0%	13.5%	20.5%
NC	4,511,850	98.6%	8.6%	17.8%	9.6%	15.0%
ND	307,495	98.8%	6.9%	15.7%	10.1%	20.1%
NE	818,200	99.3%	4.6%	11.8%	8.5%	13.9%
NH	609,134	96.3%	9.2%	18.3%	11.6%	14.6%
NJ	3,554,307	99.4%	4.9%	10.8%	5.6%	7.0%
NM	795,494	98.7%	9.9%	18.6%	11.4%	19.1%
NV	1,086,039	99.7%	2.7%	8.5%	4.3%	17.4%
NY	7,815,096	98.6%	8.1%	16.0%	4.5%	8.0%
OH	5,536,822	99.4%	4.4%	10.2%	4.2%	10.2%
OK	1,636,114	98.5%	8.5%	18.7%	14.6%	23.4%
OR	1,683,398	99.6%	4.4%	11.1%	3.8%	12.8%
PA	5,585,942	98.1%	5.9%	13.2%	8.8%	12.5%
RI	478,945	98.5%	4.7%	11.3%	3.6%	4.1%
SC	2,129,406	98.7%	6.8%	14.6%	10.5%	16.9%
SD	344,856	98.2%	7.3%	15.6%	13.4%	22.5%
TN	2,935,434	98.2%	5.8%	13.1%	6.3%	15.3%
TX	9,944,989	98.6%	6.9%	15.0%	9.9%	20.3%
UT	999,487	99.5%	8.4%	15.1%	9.1%	27.8%
VA	3,335,709	98.8%	4.5%	12.0%	7.1%	12.6%
VT	325,779	93.2%	15.8%	31.9%	15.5%	18.1%
WA	2,843,150	99.6%	4.6%	11.2%	5.0%	9.9%
WI	2,725,873	99.1%	5.5%	11.4%	6.6%	14.2%
WV	730,319	85.3%	11.4%	25.3%	29.7%	39.7%
WY	243,326	98.8%	12.2%	21.0%	11.9%	20.8%
Total	132,858,813	98.9%	5.7%	13.3%	6.9%	14.2%

References

Dohrmann, S., Han, D. and Mohadjer, L, "Residential Address Lists vs. Traditional Listing: Enumerating Households and Group Quarters," Proceedings of the Annual Meeting of the American Statistical Association, 2006.

Dohrmann, S., Han, D. and Mohadjer, L, "Improving Coverage of Residential Address Lists in Multistage Area Samples," Proceedings of the Annual Meeting of the American Statistical Association, 2007.

Iannacchione, V., Morton, K., McMichael, J., Cunningham, D., Cajka, J. and Chromy, J, "Comparing the Coverage of a Household Sampling Frame Based on Mailing Addresses to a Frame Based on Field Enumeration," Proceedings of the Annual Meeting of the American Statistical Association, 2007.

Loudermilk, C, "A National Evaluation of Coverage for a Sampling Frame Based on the Master Address File," Proceedings of the Annual Meeting of the American Statistical Association, 2009.

Loudermilk, C. and Kennel. T., "Deciphering the DSF: Which Addresses from the Delivery Sequence File Should Be Included in the Sampling Frames for Demographic Surveys?," Proceedings of the Annual Meeting of the American Statistical Association, 2005.

Martin, J. and Loudermilk, C., "Assessing the Filter Rules for Extracting Addresses from the Master Address File To Construct a Housing Unit Frame for Current Demographic Surveys," Proceedings of the Annual Meeting of the American Statistical Association, 2008.

O'Muircheartaigh C., Eckman, S., and Weiss, C., "Traditional and Enhanced Field Listing for Probability Sampling, " Proceedings of the Annual Meeting of the American Statistical Association, 2002.

O'Muircheartaigh, C., English, E., and Eckman, S., "Predicting the Relative Quality of Alternative Sampling Frames," Proceedings of the Annual Meeting of the American Statistical Association, 2007.

O'Muircheartaigh, C., English, N. and Eckman, S. and Upchurch, H. and Garcia, E. and Lepkowski, J., "Validating a Sampling Revolution: Benchmarking Address Lists against Traditional Listing," Proceedings of the Annual Meeting of the American Statistical Association, 2006.

Staab, J. and Iannacchione, V., "Evaluating the Use of Residential Mailing Addresses in a National Household Survey," Proceedings of the Annual Meeting of the American Statistical Association, 2003.

Thompson, G. and Turmelle, C., "Classification of Address Register Coverage Rates," Proceedings of the Annual Meeting of the American Statistical Association, 2004.

Turmelle, C., Rodrigue, J., and Thompson, G., "Using the Canadian Address Register in the Labour Force Survey Implementation, Results and Lessons Learned," Federal Committee on Statistical Methodology, 2005.

US Department of Labor, Bureau of Labor Statistics. *Current Population Survey Technical Paper 66 - Design and Methodology*. Washington, D.C. 2006.

US Office of Management and Budget, *Standards and Guidelines for Statistical Surveys*, http://www.whitehouse.gov/omb/inforeg_statpolicy, September 2006.