CARRA Working Paper Series

Working Paper *#2014-06*

Comparison of Survey, Federal, and Commercial Address Data Quality

Quentin Brummet
U.S. Census Bureau

Center for Administrative Records Research and Applications
U.S. Census Bureau
Washington, D.C. 20233

Paper Issued: June 30, 2014

1 Introduction

This report summarizes matching of survey, commercial, and administrative records housing units to the Census Bureau Master Address File (MAF). Matching survey records to the MAF allows us to attach a unique identifier ("MAFID") to each unit and link records across files. These linkages can potentially improve data quality and decrease costs and respondent burden in surveys.

This report considers matching from three separate data sources: the 2009 American Housing Survey (AHS), year 2009 commercial data obtained from CoreLogic, and administrative records obtained from the U.S. Department of Housing and Urban Development (HUD). The HUD data come from the 2010 and 2011 Tenant Rental Assistance Certification System (TRACS) and Public and Indian Housing Information Center (PIC).

Table 1 provides a brief overview of the sizes of the data sets and our overall match rates. In general, we tend to match at higher rates for the federal data sets. The survey data from the AHS match at a higher rate than the HUD administrative records, and the commercial data from CoreLogic match at a much lower rate than the other data sets**.**

Note also that the AHS match rate falls in the range of MAF match rates from other recent surveys. For example, the 2008 Survey of Income and Program Participation had a match rate of 91.15 percent while the 2010 and 2011 Current Population Surveys had match rates of 95.43 and 97.28 percent, respectively.

Table 1: Observation Count and MAF Match Rate by Data Set

Data Set	N	MAF Match Rate
2009 AHS	62,135	91.96%
2009 CoreLogic	169,024,241	63.41%
2010 HUD TRACS	2,615,443	88.51%
2011 HUD TRACS	2,613,962	89.75%
2010 HUD PIC	7,576,347	85.54%
2011 HUD PIC	7,732,490	84.42%

The rest of the report is composed as follows. Section 2 provides a brief overview of our MAF Matching process. Then, sections 3-6 provide analysis of matching for AHS to MAF, CoreLogic to MAF, AHS to CoreLogic, and HUD administrative records to MAF, respectively. For each match, we first document the prevalence of missing data and overall match rates for the data sets, then evaluate how the match rate varies by characteristics of the housing unit and address. Section 7 concludes, and discusses potential limitations of the work.

2 Overview of MAF Match Process

The Center for Administrative Records Research and Applications (CARRA) MAF matching process uses information on house number, street prefix, directional prefix, street name, street suffix, directional suffix, apartment number/description, and five-digit zip code to match addresses to the MAF.

Prior to matching, addresses are standardized. First, the matching routine uses Pitney Bowes Code1[1] to correct misspelled street names and update zip codes. After this, addresses are parsed using a standardizer from the U.S. Census Bureau's Geography Division. This parsing is particularly important as it extracts information such as street prefix and suffix types (e.g., identifiers for rural routes or state highways).

The match software uses a probabilistic matching algorithm to compare information from the parsed fields. Each potential match is assigned a probability based on the similarity of the fields. The matching routine is composed of two passes. The first pass blocks on three digit zip code and house number, then matches based on the similarity of the other address fields.[2] If two addresses are determined to be a match, they are removed from the potential matching pool for the next pass, which is designed to match rural routes and P.O. boxes. This pass blocks on three digit zip code and box/route number and then matches on the rest of the address fields.

3 Matching 2009 American Housing Survey to MAF

3.1 Summary of Missing Data and Matching

This section contains information on match rates from the 2009 AHS to the July 2010 MAF.[3] The sample is restricted to completed interview cases in the 2009 AHS. As documented in Table 2, the AHS data set contains relatively little missing information on address. House number and zip code are missing for a small fraction of observations, and street name is never missing.

[1] In 2012 and later, DataFlux is used to preprocess addresses.
[2] Blocking refers to selecting a subset of data from which possible matches can be drawn (in this case, only potential observations with the same house number and three-digit zip code).
[3] CARRA receives updates to the MAF twice a year.

Table 2: Summary of Missing Address Information in 2009 AHS

Variable	Percent of Observations Missing
House Number	0.77%
House Number Suffix	95.37%
Street Name	0.00%
Unit Description	73.67%
Physical Description	29.59%
City	0.03%
State	0.00%
5-digit Zip Code	1.01%
Any Primary Matching Field (House Number, Street Name, Zip)	1.77%

N=62,135.

We have difficulty matching observations with missing address information, particularly when information is missing for house number. As shown below in Table 3, we can match less than a quarter of observations for housing units lacking a house number in their address.

Table 3: 2009 AHS MAF Match Rate by Address Type

	MAF Match Rate	Percent of Observations
Missing House Number	24.48%	0.77%
Missing Street Name	-	0.00%
Missing Zip Code	87.30%	1.01%
Missing None of Above	92.53%	98.23%
Total	91.96%	

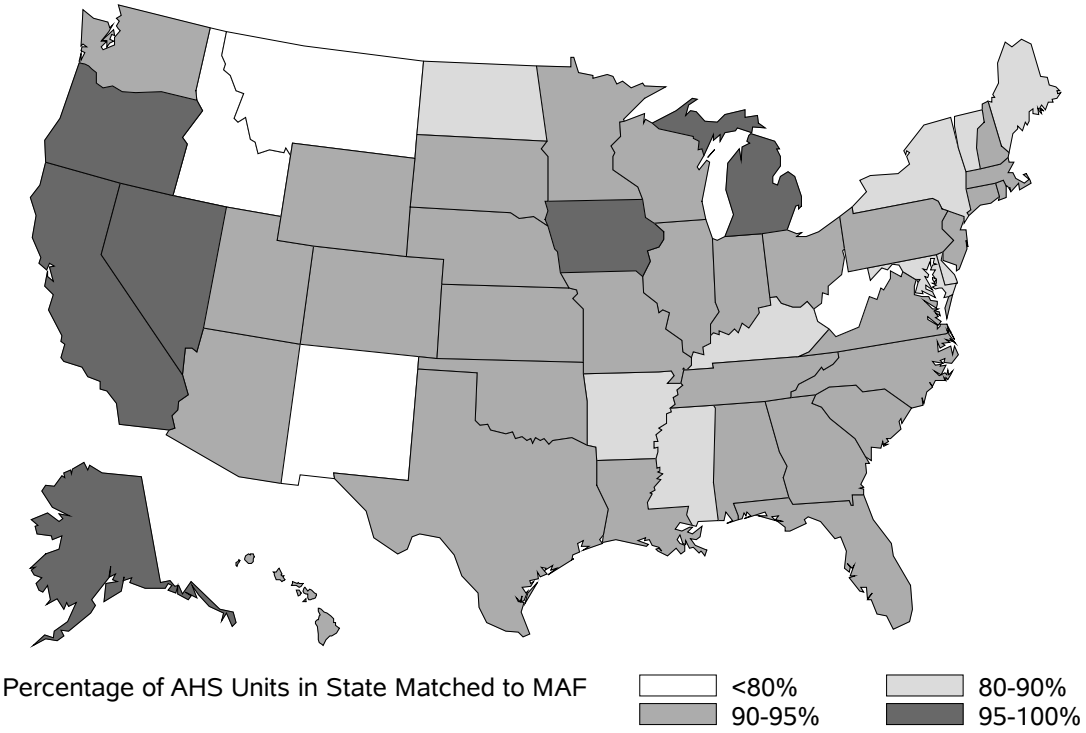
N=62,135.

While this missing information does contribute to our inability to merge some observations, it does not explain the majority of unmatched observations. The following subsections discuss characteristics of the housing units or addresses that are associated with low match rates.

3.2 Characteristics of Matched Units

While there is some variation across states in the match rate for the AHS, Figure 1 documents that most states have match rates in the 90-95 percent range. We match at higher rates in West Coast states, and slightly lower rates in more rural states. Four states (Idaho, New Mexico, Montana, and West Virginia) stand out as having particularly low match rates of less than 80 percent.[4]

Figure 1: 2009 AHS MAF Match Rate across U.S. States



Source: 2009 AHS matched to July 2010 MAF.

Table 4 also indicates that our match performs worse in rural areas. In particular, our match rate is almost ten percentage points lower for housing units classified as rural in the AHS.[5]

[4] These patterns have been observed in other studies. For instance, Idaho, New Mexico, Montana, and West Virginia have four of the seven worst match rates from the 2010 decennial census to administrative records. See Sonya Rastogi and Amy O'Hara, 2012, *2010 Census Match Study*, Center for Administrative Records Research and Applications Report, U.S. Census Bureau.

[5] Note that there are no P.O. boxes in the AHS. In addition, addresses that contain "Rural Route," County Road," or "State Route" are more difficult to match and more likely to be classified as "Rural." This difference is not large enough to explain the difference in match rates between Urban and Rural locales, however.

Table 4: 2009 AHS MAF Match Rate by Urban/Rural (Using 1990 Urban-Rural Code)

	MAF Match Rate	Percent of Observations
Urban	93.84%	76.21%
Rural	85.92%	23.79%
Total	91.96%	100%

N=62,135.

In addition, Table 5 demonstrates that the match rates are also consistently lower for rural areas when using the more detailed "City/Balance/Urban/Rural" codes that are available in the AHS.

Table 5: 2009 AHS MAF Match Rate by City/Balance/Urban/Rural

	MAF Match Rate	Percent of Observations
City	92.58%	20.32%
Balance	94.77%	23.81%
Urban	91.84%	8.84%
Rural	84.06%	16.90%
Code Missing	93.78%	30.14%
Total	91.96%	100%

N=62,135, City/Balance/Urban/Rural code based on 1980 definition.

Note that the AHS is a panel and units entered the panel in different years. These different years of introduction into the AHS may correlate with different match rates. However, Table 6 indicates no substantial differences in match rates based on the year in which the unit was introduced to the AHS.

Table 6: 2009 AHS MAF Match Rate by Year Introduction

Year of Introduction	MAF Match Rate	Percent of Observations
1985	92.46	57.52
1987	91.56	2.82
1989	90.34	2.33
1991	88.57	2.14
1993	91.43	1.80
1995	90.24	2.05
1997	91.91	2.29
1999	92.53	2.46
2001	92.69	2.60
2003	91.46	2.64
2005	91.03	5.42
2007[6]	88.27	3.36
2009	92.25	12.58
Total	91.96%	100%

N=62,135.

3.3 Programmatic Determinants of Match Rates

Length of the street name is often associated with low match rates, as long street names may be difficult to parse or represent data errors.[7] As shown below in Table 7, the AHS match performs worse with longer street names. The match is particularly low for the small fraction of observations with street names longer than two words.

[6] The match rate for units introduced to the AHS in 2007 is slightly lower than the other years; note that this cannot be explained by prevalence of multi-unit buildings, urban/rural concentrations, or length of street names for this group of units.

[7] For example, pre-processed street names in the AHS sometimes include descriptive information about the unit that is difficult to match. For example, some contain interviewer notes such as "Call sun PM 5-10," or refer to specific buildings on college campuses.

Table 7: 2009 AHS MAF Match Rate by Length of Street Name

Number of Words in Street Name[8]	MAF Match Rate	Percent of Observations
1	93.03%	82.07%
2	90.22%	16.24%
3	69.73%	0.84%
4	36.07%	0.10%
5+	18.18%	0.02%
Missing[9]	45.08%	0.74%
Total	91.96%	100%

N=62,135.

The unit description field in the AHS contains apartment numbers or other references that help to identify the unit. As documented in Table 8, the match performs worse when this sort of information is present in the data set. Reviewing the unmatched observations, one commonly occurring issue is the case where this field contains some sort of descriptive information about the unit (e.g., "REAR" or "BASEMENT APARTMENT").

Table 8: 2009 AHS MAF Match Rate by Unit Description Field

	MAF Match Rate	Percent of Observations
Unit Description Blank	94.62%	26.33%
Unit Description Non-Blank	84.52%	73.66%
Total	91.96%	100%

N=62,135.

Looking more generally at the number of units in the building, Table 9 demonstrates that low match rates for multi-unit buildings are concentrated in buildings with two or three units. In fact, our match rate for buildings with four or more units is almost as high as our rate for single-unit buildings.

[8] Note that address lengths in all sections of this report refer to addresses standardized and parsed using Pitney Bowes Code1 and a standardizer from the U.S. Census Bureau's Geography Division. The distribution of word length for all data sets considered in this report is available in Appendix B.

[9] Note that our standardization process occasionally sets street name to missing. These units are largely comprised of rural routes and mobile home parks, but also include college dormitories and units that are identified in the AHS by mostly descriptive information.

Table 9: 2009 AHS MAF Match Rate by Number of Units

Number of Units	MAF Match Rate	Percent of Observations
1 Unit	93.85%	62.62%
2 Units	78.52%	3.59%
3 Units	84.30%	1.28%
4 Units	90.54%	2.60%
5-9 Units	89.92%	4.50%
10+ Units	92.55%	11.42%
Variable Missing[10]	88.09%	13.98%
Total	91.96%	100%

N=62,135.

4 Matching 2009 CoreLogic to MAF

4.1 Summary of Missing Data and Matching

Table 10 summarizes missing data in the 2009 CoreLogic data set. Unlike the AHS, the CoreLogic file includes substantial amounts of missing information on address. In particular, for property addresses (which were used for MAF matching), over 18 percent of observations are missing a primary match field.

Table 10: Summary of Missing Data for Key Matching Variables in 2009 CoreLogic

Variable	Percent of Observations Missing (Mailing Address)	Percent of Observations Missing (Property Address*)
House Number	12.06%	0.01%
House Number Suffix	99.29%	99.40%
Street Name	0%	14.32%
Apartment Number	0%	94.00%
City	1.25%	16.88%
State	1.14%	0.12%
Zip Code	1.31%	17.72%
Any Primary Matching Field (House Number, Street Name, Zip)	12.16%	18.18%

N=169,024,241.

*Property address was used for matching to MAF

[10] Addresses with missing information on number of units are similar to single-unit structures in terms of prevalence of homeownership and fraction of units that are mobile homes.

As shown below in Table 11, the CoreLogic file matches to the MAF at a lower rate than the AHS (91.96 percent). While the prevalence of missing information explains roughly half of this low match rate, our match rate for observations with complete address is still substantially lower than match rates for AHS.

Table 11: 2009 CoreLogic MAF Match Rate by Address Type

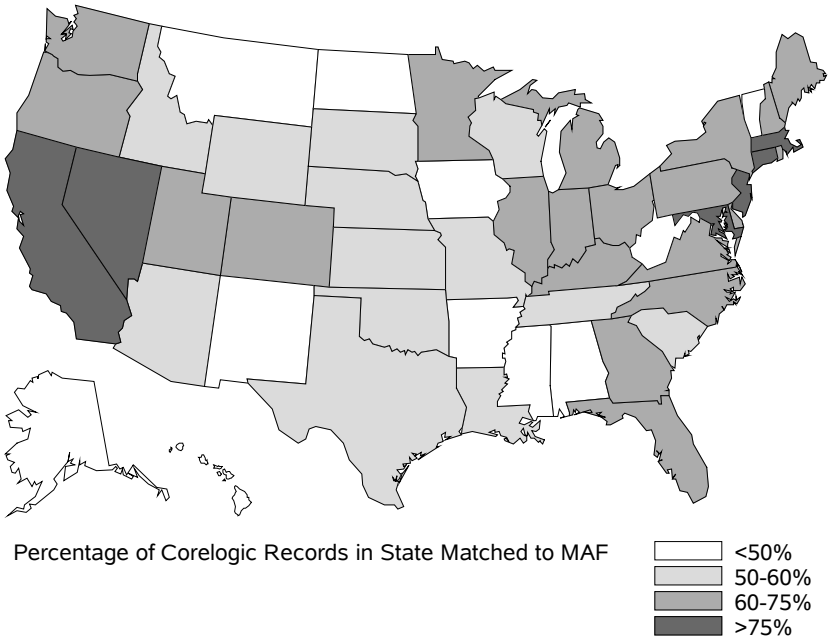
Property Address Type	MAF Match Rate	Percent of Observations
Missing House Number	0.00%	0.01%
Missing Street Name	0.00%	14.32%
Missing Zip Code	0.04%	17.72%
Missing None of Above	77.49%	81.82%
Total	63.41%	

N=169,024,241.

4.2 Characteristics of Matched Units

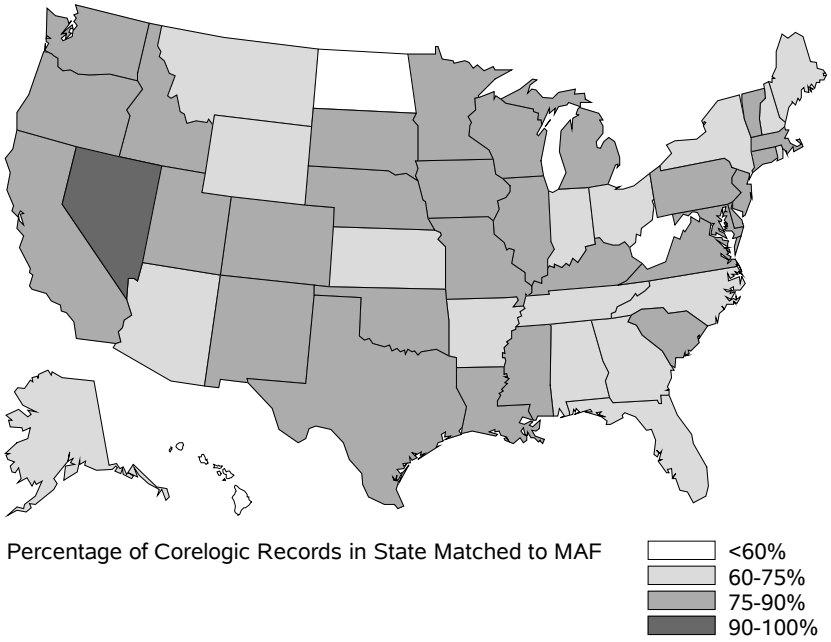
Figures 2 and 3 display match rates across U.S. states for all CoreLogic records and CoreLogic records with non-missing addresses, respectively. Again, western states match at higher rates. However, the Northeast seems to be match at higher rates when compared to the rest of the country in CoreLogic. This pattern does not appear in any of the other data sources.

Figure 2: 2009 CoreLogic MAF Match Rate across U.S. States



Source: 2009 CoreLogic data matched to January 2011 MAF.

Figure 3: 2009 CoreLogic MAF Match Rate across U.S. States, Conditional on Non-Missing Address



Source: 2009 CoreLogic data matched to January 2011 MAF.

4.3 Programmatic Determinants of Match Rates

Table 12 documents that we again match at lower rates for long street names. Units with street names longer than two words match at very low rates. Also note that CoreLogic contains a slightly higher percentage of observations with long street names than does AHS.

Table 12: 2009 CoreLogic MAF Match Rate by Length of Street Name

Number of Words in Street Name	MAF Match Rate	Percent of Observations
1	74.18%	69.96%
2	74.90%	14.69%
3	57.58%	0.77%
4	42.94%	0.06%
5+	30.57%	0.01%
Missing	0.25%	14.51%
Total	63.41%	100%

N=169,024,241.

As with the AHS, CoreLogic matches at lower rates in multi-unit buildings. In particular, Tables 14 and 15 show that our match performs worse for the relatively small fraction of observations with apartment numbers. Of the over 16 million records coded in CoreLogic as apartments, duplexes, or condominiums, 52.56 percent have a missing apartment number. Hence, many of these units are likely "header records," where one observation represents an entire condominium or apartment building. [11]

Table 13: 2009 CoreLogic MAF Match Rate by Apartment Number Field

	MAF Match Rate	Percent of Observations
Apartment Number Blank	63.84%	94.00%
Apartment Number Non- Blank	56.67%	6.00%
Total	63.41%	100%

N=169,024,241.

[11] Our match rate varies across apartment/condominium type, but unit type does not explain why we match observations with apartment numbers at lower rates. In particular, note that we match 74.3 percent of condominiums and 58.70 percent of apartments without apartment numbers. However, we only match 63.05 percent of condominiums and 54.80 percent of apartments with apartment numbers.

Table 14: 2009 CoreLogic MAF Match Rate by Apartment Number Field, Conditional on Complete Address[12]

	MAF Match Rate	Percent of Observations
Apartment Number Blank	79.05%	75.90%
Apartment Number Non- Blank	57.49%	5.92%
Total	77.49%	81.82%

N=169,024,241.

Below, Table 15 displays the MAF match rate for CoreLogic by type of property. We clearly match at much higher rates for single-family residences, but our match rate for condominiums, duplexes, and apartments is still roughly 65 percent. The very low MAF Match rates are concentrated in housing units classified by CoreLogic as vacant or other properties.[13] This reflects that the MAF extracts used for matching are not intended to cover non-residential units.

Table 15: 2009 CoreLogic MAF Match Rate by Property Type

	MAF Match Rate	Percent of Observations
Single Family Residence/Townhouse	80.99%	62.94%
Condominium/Duplex/Apartment	65.69%	9.58%
Vacant	12.90%	12.89%
Other[14]	30.69%	14.59%
Total	63.41%	100%

N=169,024,241.

5 Matching 2009 AHS to 2009 CoreLogic

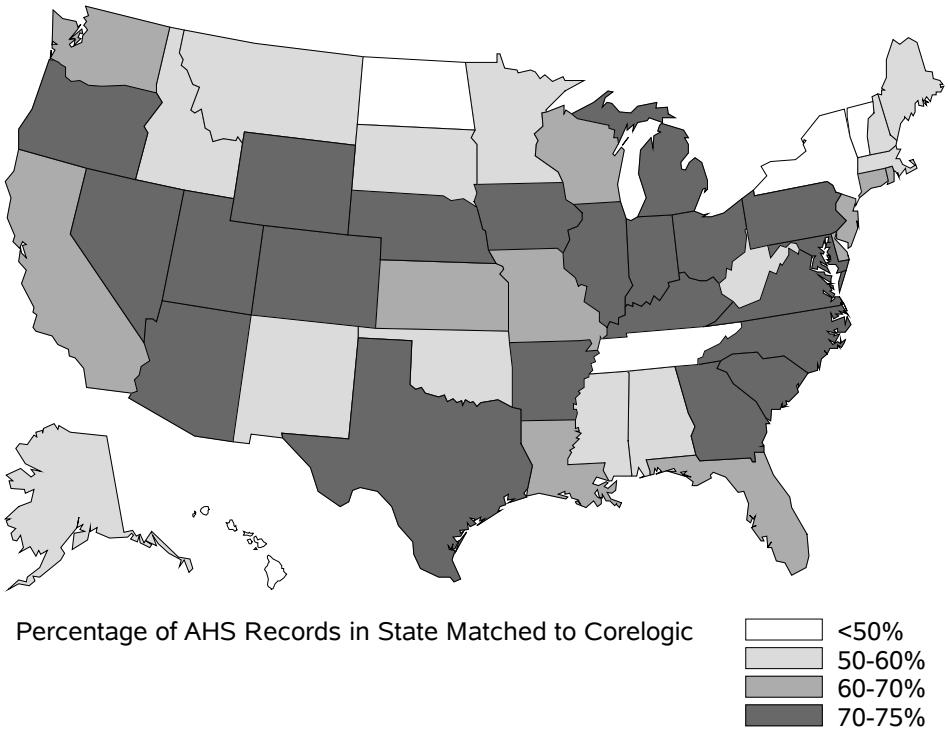
We are able to attach 56,782 unique MAFIDs to the 2009 AHS file. Of these units, we match 62.28 percent to a corresponding unit in CoreLogic by linking on MAFID. The tables and figures below provide detail on factors associated with this match rate. Note that the "AHS-CoreLogic match rate" is calculated as the fraction of AHS units with MAFIDs that link to CoreLogic. This definition is preferred because future production of the AHS will use the MAF as a sampling frame and hence future AHS samples will contain MAFIDs for all observations. Match rates calculated based on the entire AHS sample are lower than those presented below in sections 5.1-5.3, but all patterns remain qualitatively similar regardless of which definition of match rate is used.

[12] "Complete Address" is defined as an address containing non-missing house number, street name, and zip code.
[13] Note that vacant units in CoreLogic tend to be evenly distributed across states.
[14] The "other properties" category is primarily comprised of commercial and industrial locations.

5.1 Characteristics of Matched Units

As shown in Figure 4, there is relatively little variation across states in terms of match rates from the AHS to CoreLogic. A few states have particularly low match rates, but a large number of states have match rates in the 70-75 percent window and many more are in the 60-75 percent window. Tennessee stands out as having a particularly poor match rate – only 1.3 percent of AHS addresses in the state match to CoreLogic.[15]

Figure 4: 2009 AHS-CoreLogic Match Rate across U.S. States



Source: 2009 AHS linked to 2009 CoreLogic using MAFID.

Tables 16 and 17 display match rates by AHS Urban/Rural and City/Balance/Urban/Rural classifications, respectively. Unlike what we observed in Section 3.2, match rates between AHS and CoreLogic are actually higher in rural locations. This pattern is explained by the fact that multi-unit buildings match at extremely low rates between CoreLogic and AHS. Section 5.2 further explores this result.

[15] The majority of AHS addresses in Tennessee are from the original AHS sample, but over 40 percent come from more recent sample years.

Table 16: 2009 AHS-CoreLogic Match Rate by Urban/Rural

	AHS-CoreLogic Match Rate	Percent of AHS MAFID Observations	Percent of All AHS Observations
Urban	61.00%	77.80%	76.21%
Rural	66.80%	22.20%	23.79%
Total	62.28%	100%	100%

N=56,782. Source: 2009 AHS linked to 2009 CoreLogic using MAFID.
Urban/Rural code based on 1990 definition. AHS-CoreLogic match rate refers to the percent of AHS observations with valid MAFIDs that match to CoreLogic.

Table 17: 2009 AHS-CoreLogic Match Rate by City/Balance/Urban/Rural

	AHS-CoreLogic Match Rate	Percent of AHS MAFID Observations	Percent of All AHS Observations
City	51.15%	20.36%	20.32%
Balance	65.96%	24.59%	23.81%
Urban	64.15%	8.81%	8.84%
Rural	65.57%	15.38%	16.90%
Code Missing	64.43%	30.86%	30.14%
Total	62.28%	100%	100%

N=56,782. Source: 2009 AHS linked to 2009 CoreLogic using MAFID.
City/Balance/Urban/Rural code based on 1980 definition. AHS-CoreLogic match rate refers to the percent of AHS observations with valid MAFIDs that match to CoreLogic.

5.2 Programmatic Determinants of Match Rates

Table 18 shows that very long street names again match at lower rates. Interestingly, however, two-word street names match at higher rates than one-word street names. This pattern was not observed in the AHS-MAF or CoreLogic-MAF matches.

Table 18: 2009 AHS-CoreLogic Match Rate by Length of Street Name in AHS

Number of Words in Street Name	AHS-CoreLogic Match Rate	Percent of AHS MAFID Observations	Percent of All AHS Observations
1	61.86%	83.02%	82.07%
2	65.39%	15.95%	16.24%
3	62.29%	0.63%	0.84%
4	40.91%	0.04%	0.10%
5+	0.00%	0.00%	0.02%
Missing[16]	26.11%	0.36%	0.74%
Total	62.28%	100%	100%

N=56,782. Source: 2009 AHS linked to 2009 CoreLogic using MAFID. AHS-CoreLogic match rate refers to the percent of AHS observations with valid MAFIDs that match to CoreLogic.

Table 19 documents that our match rate between the MAF and CoreLogic is extremely low for multi-unit buildings. Neither the CoreLogic-MAF nor the AHS-MAF matches display as strong a pattern when comparing single family and multi-unit buildings, but this likely reflects CoreLogic containing information on apartment buildings, but not individual apartment units. In particular, recall that 52.56 percent of the over 16 million records coded in CoreLogic as apartments, duplexes, or condominiums have a missing apartment number.

Table 19: 2009 AHS-CoreLogic Match Rate by Number of Units

Number of Units	AHS-CoreLogic Match Rate	Percent of AHS MAFID Observations	Percent of All AHS Observations
1 Unit	78.97%	64.11%	62.62%
2 Units	19.68%	3.04%	3.59%
3 Units	12.31%	1.17%	1.28%
4 Units	12.98%	2.56%	2.60%
5-9 Units	13.31%	4.41%	4.50%
10+ Units	14.81%	11.52%	11.42%
Variable Missing	62.88%	13.18%	13.98%
Total	62.28%	100%	100%

N=56,782. Source: 2009 AHS linked to 2009 CoreLogic using MAFID. AHS-CoreLogic match rate refers to the percent of AHS observations with valid MAFIDs that match to CoreLogic.

[16] Addresses that match from AHS to CoreLogic but have missing street names are primarily rural routes and P.O boxes.

5.3 AHS-CoreLogic Match Rates by Characteristics of CoreLogic-MAF Match

This section assesses how the quality of CoreLogic's coverage for a given county influences AHS-CoreLogic match rates in that county. In particular, if certain counties have tax record systems that were difficult for CoreLogic to aggregate, CoreLogic would have poor coverage for units in that county. To examine this possibility, the tables below examine whether certain counties are responsible for poor match rates between AHS and CoreLogic and analyzes counties in two ways. First, counties are sorted according to the percent difference in unit counts between CoreLogic and the MAF. [17] The second analysis displays AHS-CoreLogic match rates sorted by how well the CoreLogic data could be matched to the MAF.

Table 20 shows match rates by the percent difference in county-level unit counts between CoreLogic and the MAF.[18] The first column displays AHS-CoreLogic match rates for a given group of counties, while the second column indicates the percent of observations in the AHS with MAFIDs from those counties. The final column shows the percent of all AHS observations from those counties.

In general, there is little variation in the CoreLogic-MAF match across counties. However, the final two rows of the table shows that a small percentage of AHS observations belong to one of two groups. First, one group of counties have much smaller unit counts in CoreLogic than in the MAF. In addition, a smaller group of counties contains no CoreLogic information for any unit in that county. These records have particularly poor AHS-CoreLogic match rates, but the two groups combined account for only 11.89 percent of all AHS observations.

Table 20: 2009 AHS-CoreLogic Match Rate by Difference in Unit Counts between CoreLogic and MAF

Percent Difference in County-level Aggregate Unit Count between CoreLogic and MAF	AHS-CoreLogic Match Rate	Percent of AHS Observations with MAFID	Percent of All AHS Observations
Greater than 50%	70.69%	8.82%	9.13%
25.001% to 50%	66.75%	12.53%	12.33%
0.001% to 25%	63.87%	10.83%	11.14%
-24.999% to 0%	68.07%	25.65%	25.70%
-49.999% to -25%	62.24%	33.46%	32.96%
Less than -50%	36.12%	8.69%	5.97%
CoreLogic Missing County[19]	12.25%	2.79%	2.76%
Total	62.28%	100%	100%

N=56,782. Source: 2009 AHS linked to 2009 CoreLogic using MAFID. AHS-CoreLogic match rate refers to the percent of AHS observations with valid MAFIDs that match to CoreLogic.

[17] This percent difference is defined as $\frac{Corelogic\ Count - MAF\ Count}{MAF\ Count}$.

[18] Note that CoreLogic tends to have fewer units per county than the MAF extracts to which we are matching. This results from the fact that the MAF contains old addresses as well as records, which assist in Census counts but are unlikely to have a counterpart in CoreLogic (e.g., the underside of an overpass that may contain a sizeable homeless population).

[19] The CoreLogic data set contains at least one housing unit in 2,955 of the 3,146 United States counties. Because our match procedure does not require valid county identifiers, we still match some units in these counties even if the CoreLogic data set contains no information on which county a unit is located.

Table 21 shows AHS-CoreLogic match rates by the aggregate county-level CoreLogic-MAF match rate. Overall, AHS-CoreLogic match rates are higher for counties with higher CoreLogic coverage, but match rates are relatively similar across the majority of counties.

However, a small group of counties (8.07 percent of all AHS observations) stands out as having both low CoreLogic coverage and low CoreLogic-MAF match rates. In addition, 2.76 percent of AHS observations fall into a county where CoreLogic contains no housing units. This pattern suggests that a small group of counties have poor coverage in CoreLogic which leads our match rate to be low in these counties. Nonetheless, AHS-CoreLogic match rates are relatively similar across the remaining 89.17 percent of AHS observations.

Table 21: 2009 AHS-CoreLogic Match Rate by CoreLogic MAF Coverage

County-level Match Rate between CoreLogic and MAF	AHS-CoreLogic Match Rate	Percent of AHS Observations with MAFID	Percent of All AHS Observations
<40%	49.37%	6.90%	8.07%
40%-50%	63.28%	4.62%	4.85%
50%-60%	63.87%	9.28%	9.56%
60%-70%	60.72%	16.53%	16.53%
70%-80%	64.71%	32.03%	31.37%
80%-90%	67.93%	25.56%	24.63%
>90%	68.10%	2.29%	2.22%
CoreLogic Missing County	12.25%	2.79%	2.76%
Total	62.28%	100%	100%

N=56,782. Source: 2009 AHS linked to 2009 CoreLogic using MAFID. AHS-CoreLogic match rate refers to the percent of AHS observations with valid MAFIDs that match to CoreLogic.

6 Matching HUD Administrative Records to MAF

6.1 Summary of Missing Data and Matching

The HUD administrative records in this analysis come from two sets of files.[20] The TRACS files cover individuals in privately owned, subsidized housing. Owners of subsidized multifamily projects are required to submit data for housing assistance payments through TRACS. The PIC files contain information on individuals who receive housing assistance. The sources include individuals in public housing, housing choice vouchers, and Section 8 moderate rehabilitation programs. Both of these files are person-level, as opposed to the MAF, AHS, and CoreLogic files, where each observation represented a housing unit.

[20] Note that the U.S. Census Bureau's Geography Division matched the 2010 HUD administrative records, while CARRA matched the 2011 HUD administrative records. Matching methodology is similar between the Geography Division and CARRA, so all patterns should remain qualitatively similar.

As shown in Table 22, HUD administrative records have relatively little missing information on address in input files. Note that Table 22 refers to data prior to pre-processing, where HUD administrative records include all address elements together in one data field. Missing information for addresses after processing is still a concern, and section 6.3 documents that roughly 4.49 and 7.02 percent of street names are missing for the 2011 TRACS and PIC files, respectively.

Table 22: Summary of Missing Data for Key Matching Variables in HUD Administrative Records

	TRACS		PIC	
Variable	2010	2011	2010	2011
Address	0.00%	0.00%	0.00%	0.00%
City	1.59%	2.00%	4.37%	5.47%
State	0.59%	0.72%	4.37%	5.47%
Zip Code	1.89%	2.00%	4.37%	5.47%
Any Primary Matching Field[21] (House Number, Street Name, City, Zip)	1.89%	2.00%	4.37%	5.47%
N	2,615,443	2,613,962	7,576,347	7,732,490

Table 23 documents match rates across the HUD data sets. The match rates are similar between data sets, with slightly higher rates in the TRACS data sets.

Table 23: MAF Match Rates in HUD Administrative Records

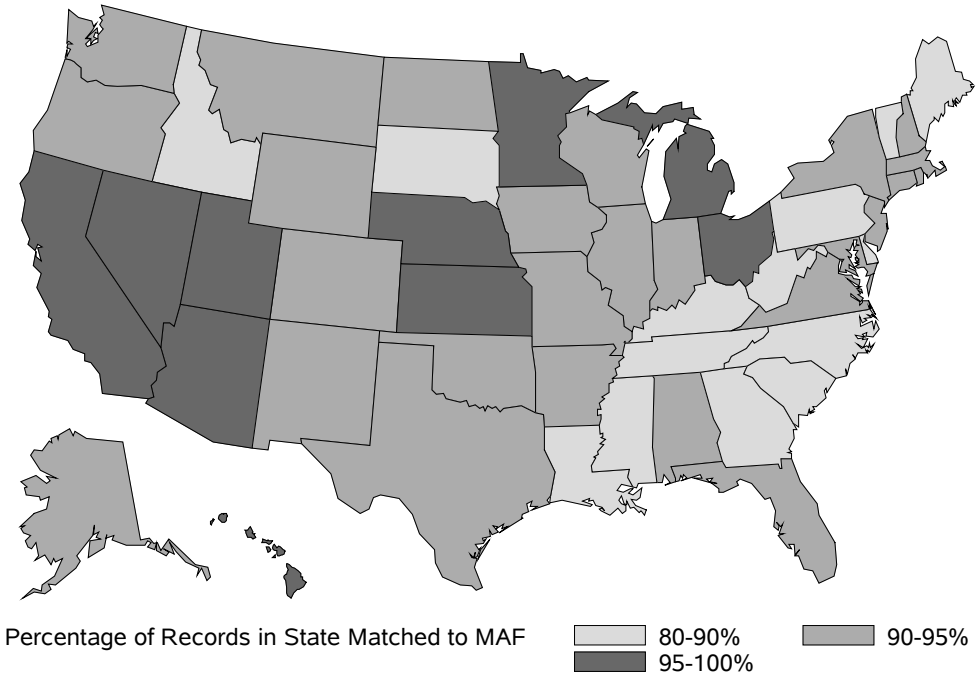
	TRACS		PIC	
	2010	2011	2010	2011
MAF Match Rate[22]	88.51%	89.75%	85.54%	84.42%
MAF Match Rate Conditional on Non-Missing Address and Zip	90.21%	91.58%	89.45%	89.30%
N	2,615,443	2,613,962	7,576,347	7,732,490

[21] With the exception of a handful of observations, there is complete overlap in missing data prevalence between address fields in the PIC files and 90 percent of the records are in public housing.
[22] Note that the 2011 files used DataFlux rather than Pitney Bowes Code1 to preprocess addresses.

6.2 Characteristics of Matched Units

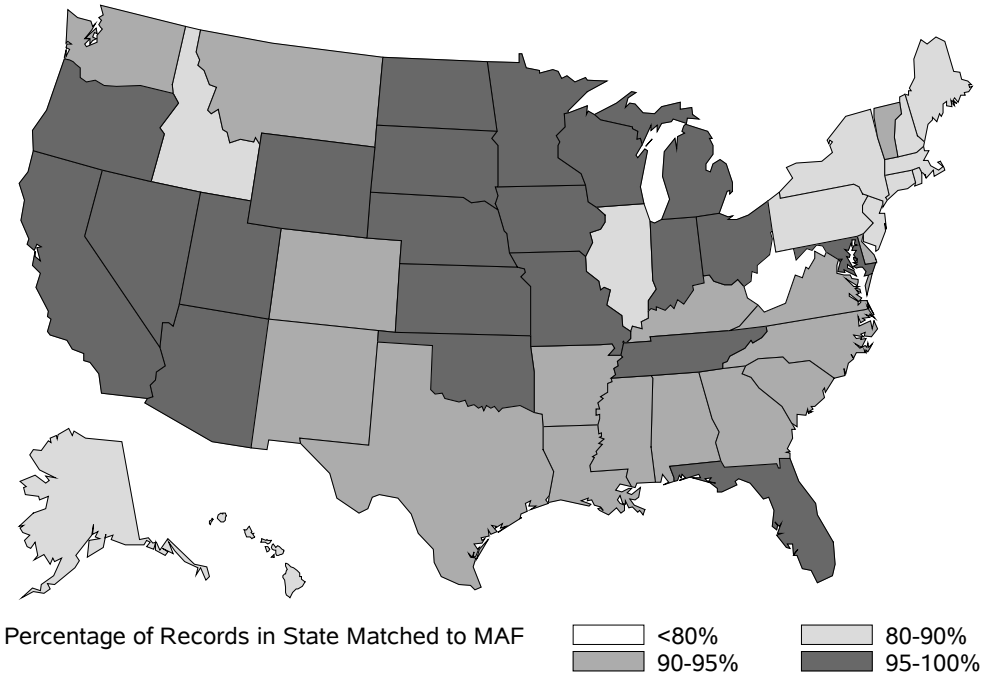
Figures 5 and 6 show the 2011 match rates across states in the U.S. for the two HUD sets. Again, there is some variation across states in match rates with states, on the West Coast and in the Midwest matching at relatively high rates and states in the South matching at relatively low rates.

Figure 5: 2011 HUD TRACS MAF Match Rate across U.S. States



Source: 2011 TRACS data matched to January 2011 MAF.

Figure 6: 2011 HUD PIC MAF Match Rate across U.S. States



Source: 2011 PIC data matched to January 2011 MAF.

Table 24 summarizes MAF match rates by program type in the HUD TRACS files. As shown below, there is very little variation in match rates among different certification types in the TRACS files.

Table 24: MAF Match Rate by Certification Type in TRACS Files

	2010		2011	
	MAF Match Rate	Percent of Observations	MAF Match Rate	Percent of Observations
Annual Recertification	88.78%	69.35%	90.10%	70.14%
Initial Certification	90.46%	2.02%	90.07%	1.70%
Interim Recertification	88.19%	15.69%	89.16%	15.65%
Move-In	87.13%	12.94%	88.46%	12.51%
Total	88.51%	100%	89.75%	100%

N=2,615,443 in 2010.
N=2,613,962 in 2011.

Unlike certification type studied above in Table 25, there is substantial variation in match rates across program type, voucher type, and type of action in the PIC files. Units in public housing have particularly poor match rates. Other program types in the PIC files match at rates comparable to those in the TRACS files.

Table 25: MAF Match Rate by Program Type in PIC Files

	2010		2011	
	MAF Match Rate	Percent of Observations	MAF Match Rate	Percent of Observations
Section 8 Certificates	91.09%	0.08%	92.04%	0.07%
Section 8 Moderate Rehabilitation	89.09%	0.74%	85.62%	0.77%
Public Housing[23]	74.06%	30.77%	71.19%	31.06%
Section 8 Vouchers	90.66%	68.40%	90.44%	68.10%
Total	85.54%	100%	84.42%	100%

N=7,576,347 in 2010.
N=7,732,490 in 2011.

Table 26 below shows match rates by voucher type; homeownership vouchers match at extremely high rates, but are a very small fraction of total vouchers.[24]

Table 26: MAF Match Rate by Voucher Type in PIC Files

	2010		2011	
	MAF Match Rate	Percent of Observations	MAF Match Rate	Percent of Observations
Homeownership Voucher	97.10%	0.35%	96.26%	0.39%
Project Based Voucher	89.12%	0.84%	88.06%	1.15%
Tenant Based Voucher	90.64%	67.21%	90.44%	66.57%
Voucher Type Missing[25]	74.46%	31.60%	71.58%	31.90%
Total	85.54%	100%	84.42%	100%

N=7,576,347 in 2010.
N=7,732,490 in 2011.

[23] Note that while these are person-level files, addresses refer to the buildings and often do not include apartment numbers.
[24] HUD Computerized Home Underwriting Management System (CHUMS) data will be analyzed in a future study.
[25] Note that the voucher type only applies to Section 8 Vouchers. Hence, the missing category includes all program types listed in Table 25 with the exception of Section 8 Vouchers.

6.3 Programmatic Determinants of Match Rates

Tables 27 and 28 document match rates by length of street name in the HUD administrative records files.[26] Similar to before, we match at lower rates for very long street names.[27] However, there appears to be a larger drop off in match rates between one- and two-word street names when compared to AHS.

Table 27: MAF Match Rate by Length of Street Name in 2011 TRACS File

Number of Words in Address	MAF Match Rate	Percent of Observations
1	94.30%	83.04%
2	89.55%	11.54%
3	70.68%	0.74%
4	88.89%	0.14%
5+	74.64%	0.05%
Missing	9.31%	4.49%
Total	89.75%	100%

N=2,613,962.

Table 28: MAF Match Rate by Length of Street Name in 2011 PIC File

Number of Words in Address	MAF Match Rate	Percent of Observations
1	91.85%	81.18%
2	86.91%	10.55%
3	49.88%	0.82%
4	45.61%	0.30%
5+	39.82%	0.13%
Missing	1.31%	7.02%
Total	89.75%	100%

N=7,732,490.

[26] Note that these addresses are the unparsed addresses in the HUD administrative records files, and this address contains all information about a unit's location. Our matching procedure parses this address prior to matching to the MAF.

[27] As with the AHS, there are also instances of more descriptive information such as "3rd Trailer" in a specific mobile home park.

7 Conclusion

These results show the challenges of matching addresses from multiple housing file sources, with a focus on the AHS. We match over 90 percent of AHS records to the MAF. Moreover, while match rates for AHS and CoreLogic are lower, we match at high rates for single-family residences. Because we match 78.97 percent of single-unit structures between AHS and CoreLogic, data from CoreLogic may be particularly useful for the 62.62 percent of AHS records that are single-family residences. In addition, MAF match rates for HUD administrative records are also high, and open the possibility of using this information in surveys such as the AHS.

A few limitations with the current analysis are worth noting. First, our match rates between AHS and CoreLogic are substantially worse for multi-unit structures. Improving this match rate and understanding to what extent specific buildings drive these low match rates is an important subject for future research. In addition, there appears to be a subset of counties where CoreLogic has particularly poor coverage. Understanding the characteristics of these counties is important for understanding the implications of using CoreLogic to augment surveys.

Appendix A: Match Rates across U.S. States

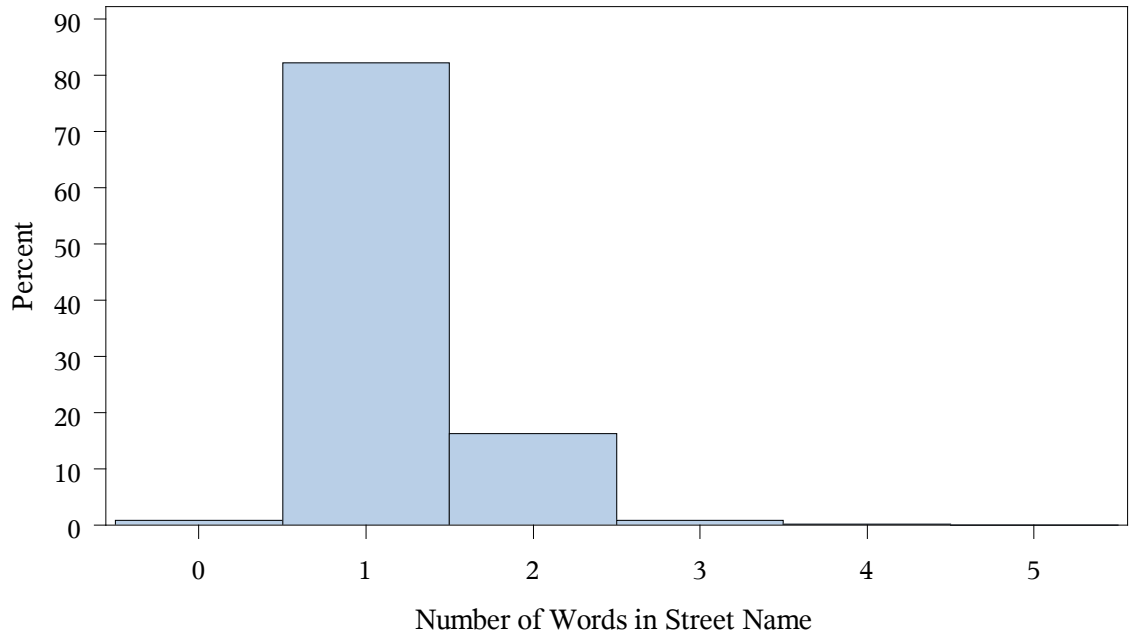
Table A.1: Summary of AHS-MAF-CoreLogic Match by State

State	N AHS Records	Fraction matched to MAF	Fraction Matched to CoreLogic
AK	116	0.9655	0.5086
AL	938	0.9136	0.5458
AR	443	0.8036	0.5350
AZ	1201	0.9384	0.6220
CA	5715	0.9622	0.6224
CO	1060	0.9406	0.6443
CT	627	0.9426	0.6108
DC	130	0.9077	0.2692
DE	228	0.8947	0.5746
FL	3728	0.9466	0.5896
GA	1559	0.9352	0.6485
HI	145	0.9448	0.3310
IA	601	0.9700	0.6473
ID	184	0.7935	0.4076
IL	3386	0.9061	0.5936
IN	1008	0.9325	0.6617
KS	420	0.9476	0.5952
KY	957	0.8788	0.5778
LA	835	0.9198	0.5952
MA	1221	0.9083	0.5094
MD	1221	0.8993	0.6216
ME	351	0.8433	0.4900
MI	3852	0.9546	0.6469
MN	959	0.9395	0.5203
MO	1131	0.9275	0.5632
MS	556	0.8471	0.4892
MT	169	0.6923	0.3491
NC	1613	0.9144	0.6138
ND	124	0.8629	0.2984
NE	387	0.9457	0.6279
NH	194	0.9485	0.5258
NJ	3333	0.9175	0.5740
NM	422	0.7275	0.4076
NV	333	0.9580	0.7027

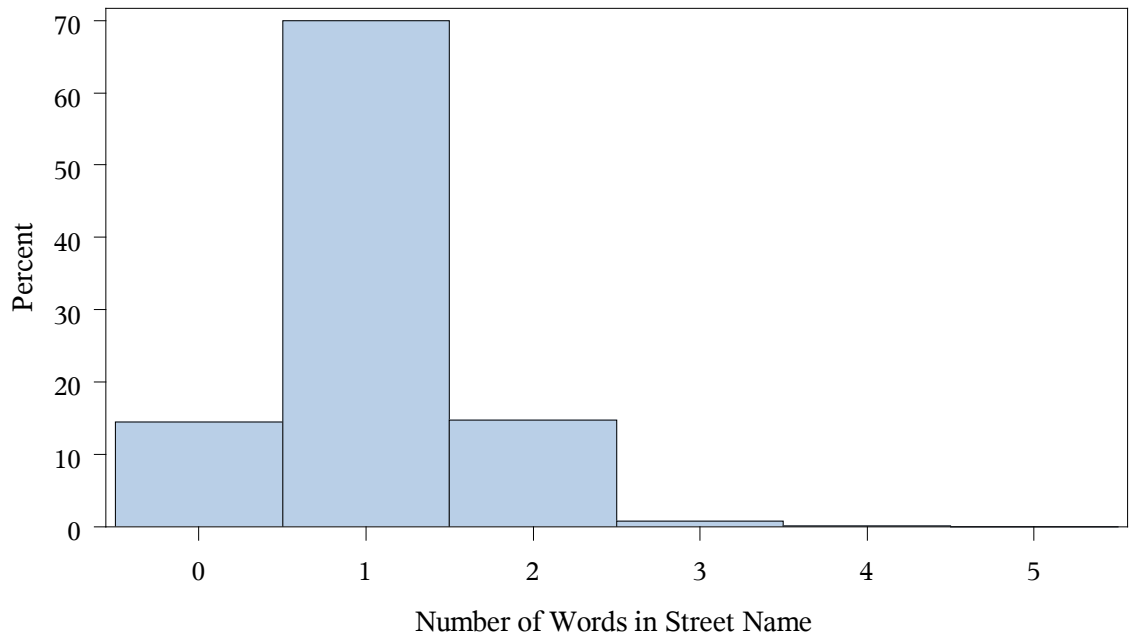
State	N AHS Records	Fraction matched to MAF	Fraction Matched to CoreLogic
NY	3911	0.8594	0.3656
OH	2423	0.9265	0.6579
OK	712	0.9059	0.5154
OR	767	0.9674	0.6871
PA	3937	0.9185	0.6203
RI	231	0.9048	0.5714
SC	916	0.9039	0.5961
SD	230	0.9130	0.5087
TN	1127	0.9335	0.0133
TX	4021	0.9140	0.5961
UT	305	0.9377	0.6328
VA	1320	0.9485	0.6530
VT	147	0.8163	0.3537
WA	1209	0.9371	0.5790
WI	1129	0.9017	0.5793
WV	484	0.7789	0.4256
WY	118	0.9068	0.6525
Total	**62,135**	**0.9196**	**0.5722**

Appendix B: Distribution of Edited Street Name Length across Data Sets

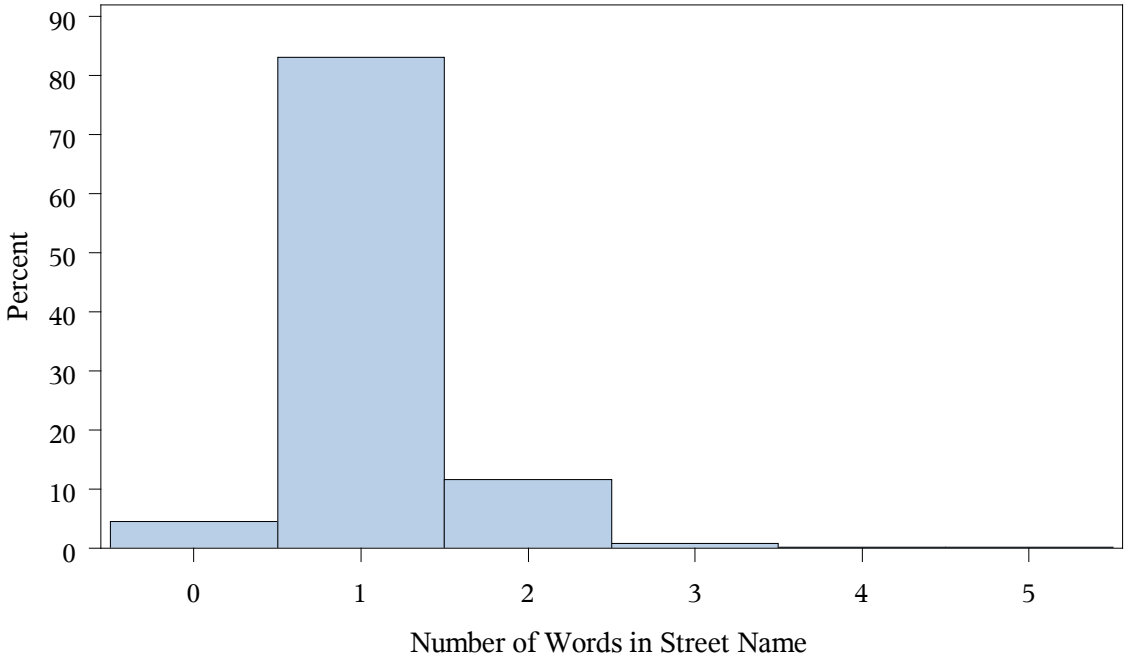
Frequency of Edited Street Name Length in AHS



Frequency of Edited Street Name Length in Corelogic



Frequency of Edited Street Name Length in 2011 HUD TRACS



Frequency of Edited Street Name Length in 2011 HUD PIC