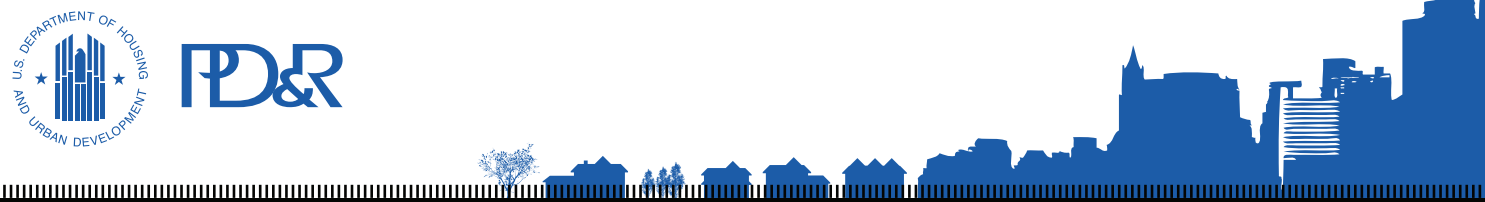
Imputing Lot Size with Property Tax Data



U.S. Department of Housing and Urban Development | Office of Policy Development and Research

Imputing Lot Size with Property Tax Data

Emily Molfino
U.S. Department of Housing and Urban Development

Contents

1. Overview[1]

The American Housing Survey (AHS) and the American Community Survey (ACS) ask respondents in single-family housing structures about the size of their lots (hereinafter referred to as lot size). Lot size is also captured in property records and tax assessment records (hereinafter referred to as property tax records). When property tax records are matched with the surveys, about 82 percent of eligible AHS and 78 percent of eligible ACS housing units have a matched property tax record with a valid lot size value. As such, lot size could be potentially imputed from property tax records. Such imputation has the potential to both reduce respondent burden and increase data quality.

This whitepaper presents a discussion and analysis that was performed to assess the use of property tax data to impute lot size for respondents who have matched property tax records. Section 2 provides a general discussion of lot size as a survey construct and as a field in property tax assessment records. Section 3 presents a statistical analysis of the potential to match AHS and ACS respondents to their property tax records with a lot size value, concluding that the availability of lot size from matched tax assessment records is sufficient to consider it a viable source for imputation.

Sections 4 and 5 lay the evidentiary foundation for using property tax records to fully replace lot size values when there is a matched record—a method known as cold decking. Section 4 compares the aggregate distribution of lot size from respondent-reported values to the aggregate distribution of lot size from property tax records, demonstrating that the aggregate distributions are similar. Section 5 evaluates and demonstrates individual-level correspondence between respondent-reported lot size values and values from property tax records.

Section 6 focuses on survey respondents who do not have matched property tax reports. For these respondents, the cold-deck method described and explored in prior sections is not an option. An alternative imputation method, the local cumulative distribution function (CDF), is described and evaluated, concluding that the local CDF method outperforms the existing hot deck method.

Section 7 concludes the whitepaper by discussing the findings from the prior sections and the decision to use lot sizes from property tax records for imputing AHS responses.

2. Lot size as a survey construct and property tax construct

This section presents a discussion of lot size as a construct. First, we lay out the validity of lot size for both respondents and property tax records. Second, we discuss the reliability of the measurement of lot size in both data sources. This whitepaper focuses on the Census Bureau's

[1] *Disclaimer*: This report is released to inform interested parties of research. Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau or the U.S. Department of Housing and Urban Development. The Census Bureau Disclosure Review Board have reviewed this data product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. CBDRB
Approval: CBDRB-FY20-344.

two largest demographic surveys that collect lot-size information from respondents: the AHS and ACS. Results should be generalizable to other demographic surveys depending on the ability to link property tax records and sample design.

2.1 Validity of the construct, from the respondent's perspective

In both the AHS and ACS, lot size is only asked of housing units that are single-family detached, single-family attached, or mobile homes. Moreover, lot size is not asked of AHS households who report that their unit is in condominium or cooperative.

The most likely instance for when a respondent's lot size is subject to interpretation is when a home sits on more than one parcel of land. While the respondent may consider a lot to be the continuous piece of land on which a house sits, the property tax jurisdiction may consider it to be split into multiple parcels.

2.2 Reliability of the measurement, from the AHS respondent's perspective

In the AHS, lot size may be determined using multiple items. The respondent may reply in either square feet, feet, or acres. If the respondents are uncertain about the size of their lots, they are asked to estimate the dimensions of the lots. Responses greater than 25 acres are grouped together. While detailed lot size information is collected by the AHS, the AHS public use files only include seven categories/ranges of lot size. In the ACS, respondents are provided a smaller set of categories (ranges) for lot size. In fact, they are provided only three groupings. Exhibit 2.1 shows the lot size aggregation used in this whitepaper.

Exhibit 2.1. Lot Size Groupings

Lot Size (AHS)	Lot Size (ACS)
Less than 1/8 acre	Less than 1 acre
1/8 up to 1/4 acre	
1/4 up to 1/2 acre	
1/2 up to 1 acre	
1 up to 5 acres	1 up to 10 acres
5 up to 10 acres	
10 acres or more	10 acres or more

For either survey, the respondent can provide an answer from memory or use other sources of information to provide an answer. This information can be private (mortgage or home inspection documentation) or public (online real estate databases or county property databases). Nonetheless, the larger ranges in the ACS help respondents who might not know the exact size of their lots. This is compared to the AHS, where either the exact lot size, dimensions of the lot, or a range of lot sizes are asked for. Unless the respondent knows the exact lot size for the unit, respondents tend to round their responses to whole or common values.[2]

[2] Manski, Charles F., and Francesca Molinari. 2010. "Rounding Probabilistic Expectations in surveys," *Journal of Business & Economic Statistics* 28.2: 219–231.

The universe of both surveys includes owners, renters, and vacant units. Respondents who own their housing units are likely well informed about the size of their lots, especially those who are recent buyers of their housing units. However, owners may be more likely to forget the exact lot size of their units as time passes following the initial purchase.[3]

The same may not be true of respondents who are renters or individuals responding for vacant units.[4] They may be able to estimate lot size or use a second-hand source. Nevertheless, renters or individuals responding for vacant units would need to determine or confirm lot size of the unit by searching through documents they may have or by researching the neighborhood. The amount of effort to find the answer is not uniform across respondents.[5]

2.3 Validity and reliability of lot size from the taxing jurisdiction's perspective

In property tax records, lot size is one of the recorded characteristics for a parcel. Pertinent for the AHS and ACS, single family housing structures are typically on their own parcels and thus have their own lot sizes in the property tax records. For a taxing jurisdiction, lot size is a straightforward concept. It is most often measured precisely using digital boundaries or coordinates from a Meets and Bounds system. When lot size appears in a property tax record, there is low risk of reported inaccuracies, especially when converted into broader categories.

2.4 Conclusion

The ACS, AHS, and property tax records intend to capture the lot size upon which a housing unit is built on. However, this is not always what is captured. Jurisdictions have both the incentive and the means to accurately capture the lot size of a parcel. Thus, we assume that there is greater accuracy within the property tax records than in respondent reported values, especially when not limited to three broad categories as in the ACS.

Moreover, the universe of eligible respondents for lot size and unit of analysis in the property tax records are closely aligned. This increases the likelihood of matching an eligible AHS and ACS record to a property tax record.

3. Analysis of the availability of lot size in property tax records

Section 2 illustrated how lot size in property tax records is a more reliable measure as compared to respondent-reported values. A potential imputation technique is to simply replace a value, whether it be provided by the respondent or missing, with the lot size value from the matched property tax record. This is called cold decking. Cold decking requires that data be available from an auxiliary data source for that sampled housing unit. This section explores how often lot size is available from property tax records for AHS and ACS sample housing units.

---

[3] Gaskell, G.D., D.B., Wright, and C.A. O'Muircheartaigh. 2000. "Telescoping of Landmark Events: Implications for Survey Research." *Public Opinion Quarterly*, 64, 77–89.

[4] For vacant housing units, a landlord, owner, real estate agent, or knowledgeable neighbor can provide data on the unit.

[5] Gummer, Tobias, and Tanja Kunz. 2019 "Relying on External Information Sources When Answering Knowledge Questions in Web Surveys." *Sociological Methods & Research* .

All tables in this whitepaper were calculated using the AHS/ACS household weights. This allows us to compare results between the two surveys, which have different sample sizes and procedures.[6] While weights are used, tables are approximations of rates of agreement. Statistical testing is only conducted in Section 4 when comparing estimated distributions.

3.1 Initial review of availability of a lot size value from a matched property tax record

The *availability* of a lot size value for an AHS or ACS respondent is contingent upon two things: the ability to match the respondent to a property tax record and the presence of a lot size value for the matched property tax record.

Matching was performed by address. The U.S. Census Bureau performs address matching using the Census Bureau's Master Address File (MAF). First, the property tax data is matched to the MAF by address using a blocking strategy: the potential matches in the property tax data are limited to records in the same ZIP Code or Census Tract. This results in each property tax record being assigned an MAF identification code (MAFID) or no MAFID if a match is not found. While this address matching process provides computational efficiency gains, an inherent assumption of the process is that both data sources have correct ZIP Codes and similar unit designations for each address. Analysts working on the AHS at HUD and the Census Bureau were able to confirm that this technique was resulting in failures to find matches.

HUD and the Census Bureau developed a process to improve AHS matches to property tax assessment data. The first stage in the matching process is a direct MAFID match. The second stage is applied to any AHS record that did *not* have a MAFID match from the first stage matching process. Matches are made using the Census tract, house number, and street name of the remaining AHS records and remaining property tax records.[7] For the 2019 AHS, this resulted in 10 percent more AHS records being matched to the property tax data. A similar process was not replicated with the ACS because HUD did not have access to the sampled addresses.

The AHS and ACS samples are selected from the MAF, so sampled housing units will have a MAFID. The property tax data and survey data can then be matched to each other using MAFID.

Exhibit 3.1 shows the availability rate of lot size from property tax records for AHS and ACS records. Roughly 82 percent of AHS records and 78 percent of ACS records have matched property tax records with a lot size values. Owner-occupied units have a higher availability rate, primarily due to their concentration in single-family detached housing units, which have a higher property tax record matching rate. The differences between the AHS and ACS lot size

[6] For more information on confidentiality protection, sampling error, non-sampling error, and definitions, see https://www.census.gov/programs-surveys/acs/technical-documentation/code-lists.html for the ACS and https://www.census.gov/programs-surveys/ahs/tech-documentation/def-errors-changes.html for the AHS.

[7] When matching on street name, an algorithm is used that calculates the likelihood two text strings matching.

availability rates reflect the improved algorithm to match AHS respondents to their property tax records.

Exhibit 3.1 Lot Size Availability Rates

	Percent of *eligible* respondents with matched record containing a valid lot size value (%)	
	2015 AHS	2014 ACS
All	82.0	78.2
By Tenure		
Owner	87.8	84.3
Renter	67.7	64.8
Vacant	69.0	60.7
By Structure Type		
Single-family detached	89.0	84.3
Single-family attached	55.9	60.9
Mobile home or RV	36.2	34.6

Note: Rates are calculated using weights to allow for comparison only and are thus approximations.
Sources: U.S. Census Bureau, 2015 American Housing Survey, U.S. Census Bureau, 2014 1-year American Community Survey, 2015 CoreLogic Property Tax Database

3.2 What share of AHS and ACS records can be cold decked from their matched property tax records?

Section 3.1 showed that roughly 82 percent of AHS and 78 percent of ACS respondents have matched property tax records with a lot size values. This rate indicated that we were capturing a significant enough portion of survey respondents to consider using tax records to impute lot size using total replacement cold decking.

Exhibit 3.2 shows the potential cold-deck rates from the AHS and ACS. For the AHS, roughly 14 percent do not have a matched property tax record with a lot size value, but the respondents did provide lot size values. For the ACS, about 21 percent do not have a matched property tax record with a lot size value, but the respondents did provide lot size values.

Exhibit 3.2 Potential Cold-deck Fates for Lot Size

	Owner (%)	**Renter (%)**	**Vacant (%)**	**All Tenures (%)**
2015 AHS				
Could be cold decked using matched value	87.8	67.7	69.0	82.0
No matched value, but respondent-reported available	10.5	20.8	23.4	13.8
No matched value or respondent reported value	1.7	11.6	7.6	4.2
Total	100.0	100.0	100.0	100.0
2014 ACS				

Could be cold decked using matched value	84.3	64.8	60.7	78.2
No matched value, but respondent-reported available	15.0	32.8	35.6	20.5
No matched value or respondent reported value	0.6	2.4	3.7	1.3
Total	100.0	100.0	100.0	100.0

Note: Rates are calculated using weights to allow for comparison only and are thus approximations.
Sources: U.S. Census Bureau, 2015 American Housing Survey, U.S. Census Bureau, 2014 1-year American Community Survey, 2015 CoreLogic Property Tax Database

3.3 Conclusion

The results in Section 3 show that AHS and ACS records can be matched to their property tax records that also have lot size values. While there is variation by structure type and tenure, this analysis is evidence that property tax records can be used for complete replacement cold deck imputation.

4. Aggregate distribution correspondence

Section 3 showed that approximately 82 percent of AHS responses and 78 percent of ACS responses could imputed with cold decking because they have matched property tax records with a lot size values. About 14 percent of AHS responses and 21 percent of ACS responses do not have a matched record, but there is a respondent-reported value.

For cold decking using lot values from matched property tax records to be acceptable, the property tax records must be an unbiased source of information for lot size. If the tax records were systematically different from respondent-reported values, imputation using tax records could result in bias depending on the sources of these differences.

There are two ways to measure systematic disagreement between respondent-reported values and tax records. The first is to measure the aggregate distributional correspondence, which is the similarity of accumulated respondent-reported lot size values and accumulated lot sizes in property records. This is covered in Sections 4.1 and 4.2. The second way is to measure individual-level correspondence, which is the similarity between a respondent's reported lot size value and the lot size value in the property tax record. Section 5 includes a discussion of individual-level correspondence. A finding of high correspondence rates would strengthen the case that property tax records are a good source of data for imputing lot size, while low correspondence rates would suggest some possible systematic disagreement.

4.1 *Aggregate distribution of lot size*

Exhibit 3.1 showed that roughly 82 percent of AHS responses can be matched to property tax records with valid lot size values, while the rate is 78 percent for ACS responses. Exhibits 4.1 and 4.2 compare the aggregate distribution of lot size values from two sources: the respondent-reported lot size values from the housing surveys (AHS and ACS) and lot size values from the respondent's matched property tax records. They include owners, renters, and vacant households.

Exhibit 4.1 does show some aggregate disagreement. This is largely driven by differences in the share of records of lots less than 1 acre. This is true for each of the smaller lot-size ranges as well as the cumulative distribution. This makes sense because respondents in these smaller lots might round their lot sizes to the next half an acre—a quarter of an acre difference might not seem apparent for these respondents.

A comparison of the aggregate distribution of AHS and property tax records shows that AHS respondents report lot sizes that are larger than what is shown in their property tax records. In the AHS, about 66 percent of respondents report a lot sizes of one-half acre or less, while their tax records show that about 76 percent of AHS respondents have lot sizes of one-half acre or less.

Exhibit 4.1 AHS Aggregate Distribution of Lot Size for All Tenures

Lot Size	AHS Share of Respondents (90% Margin of Error) (%)	AHS Cumulative Share (%)		Property Tax Share (%)	Property Tax Cumulative Share (%)	Difference in Cumulative Share (%)
Less than 1/8 acre	26.7 (0.5)	26.7		22.4	22.4	4.3
1/8 up to 1/4 acre	20.8 (0.5)	47.5		33.8	56.2	– 8.7
1/4 up to 1/2 acre	18.0 (0.6)	65.5		19.3	75.6	– 10.0
1/2 up to 1 acre	11.7 (0.4)	77.2		8.3	83.8	– 6.5
1 up to 5 acres	17.7 (0.6)	94.9		11.6	95.4	– 0.5
5 up to 10 acres	2.7 (0.2)	97.7		2.5*	97.9	– 0.2
10 acres or more	2.3 (0.3)	100.0		2.1*	100.0	0.0

Notes: Rates are calculated using weights to allow for comparison only and are thus approximations.
* signifies property tax share is within a 90% confidence interval of AHS respondent share.
The difference in cumulative shares was calculated using unrounded percentages and may appear different from the differences calculated using the rounded percentages.
Sources: U.S. Census Bureau, 2015 American Housing Survey, 2015 CoreLogic Property Tax Database

Exhibit 4.2 shows that the larger ranges in the ACS do absorb some of these differences at the aggregate level. The larger differences seen in records of less than 1 acre might be a result of respondents rounding up to 1 acre.

Exhibit 4.2 ACS Aggregate Distribution of Lot Size for All Tenures

Lot Size	ACS Share of Respondents (90% Margin of Error) (%)	ACS Cumulative Share (%)		Property Tax Share (%)	Property Tax Cumulative Share (%)	Difference in Cumulative share (%)
Less than 1 acre	80.7 (0.1)	80.7		81.9	81.9	– 1.3
1 up to 10 acres	16.5 (0.1)	97.1		15.3	97.2	– 0.1

10 acres or more	2.9 (0.1)	100.0	2.8*	100.0	0.0

Notes: Rates are calculated using weights to allow for comparison only and are thus approximations.
* signifies property tax share is within a 90% confidence interval of AHS respondent share.
The difference in cumulative shares was calculated using unrounded percentages and may appear different from the differences calculated using the rounded percentages.
Sources: U.S. Census Bureau, 2014 1-year American Community Survey, 2015 CoreLogic Property Tax Database

4.2 Aggregate distribution of lot size by tenure

The tables below break down the aggregate distribution of lot size values by owners (exhibit 4.3) and renters (exhibit 4.4). Owners are more likely to report their lot sizes as larger than what is in the property tax reports. In the AHS, about 61 percent of owners report a lot size of one-half acre or less as compared to 79 percent of renters, while their tax records show that 73 percent and 85 percent respectively of AHS respondents have lot sizes of one-half acre or less.

Exhibit 4.3 AHS Aggregate Distribution of Lot Size for Owner-occupied Units

Lot Size	AHS Share of Respondents (90% Margin of Error) (%)	AHS Cumulative Share (%)		Property Tax Share (%)	Property Tax Cumulative Share (%)	Difference in Cumulative Share (%)
Less than 1/8 acre	19.7 (0.5%)	19.7%		18.4	18.4	1.3
1/8 up to 1/4 acre	21.4 (0.6)	41.1		33.6	52.0	-10.9
1/4 up to 1/2 acre	20.3 (0.7)	61.4		21.2	73.3	-11.8
1/2 up to 1 acre	13.1 (0.5)	74.6		9.1	82.4	-7.8
1 up to 5 acres	19.7 (0.7)	94.3		12.9	95.3	-1.0
5 up to 10 acres	3.2 (0.3)	97.5		2.8	98.0	-0.5
10 acres or more	2.5 (0.3)	100.0		2.0	100.0	0.0

Notes: The difference in cumulative shares was calculated using unrounded percentages and may appear different from the differences calculated using the rounded percentages.

* signifies property tax share is within a 90% confidence interval of AHS respondent share.

Sources: Source: U.S. Census Bureau, 2015 American Housing Survey, 2015 CoreLogic Property Tax Database

Exhbit 4.4 AHS Aggregate Distribution of Lot Size for Renter-occupied Units

Lot Size	AHS Share of Respondents (90% Margin of Error) (%)	AHS Cumulative Share (%)		Property Tax Share (%)	Property Tax Cumulative Share (%)	Difference in Cumulative Share (%)
Less than 1/8 acre	48.5 (1.1%)	48.5		33.9	33.9	14.6
1/8 up to 1/4 acre	19.2 (1.2)	67.8		37.5	71.4	– 3.6
1/4 up to 1/2 acre	11.3 (0.9)	79.1		13.4	84.8	– 5.7
1/2 up to 1 acre	7.3 (0.8)	86.3		5.1	89.9	– 3.6
1 up to 5 acres	11.3 (1.0)	97.7		6.8	96.7	0.9
5 up to 10 acres	1.1 (0.3)	98.8		1.3*	98.0	0.8
10 acres or more	1.2 (0.4)	100.0		2.0	100.0	0.0

Notes: The difference in cumulative shares was calculated using unrounded percentages and may appear different from the differences calculated using the rounded percentages.
* signifies property tax share is within a 90% confidence interval of AHS respondent share.
Sources: U.S. Census Bureau, 2015 American Housing Survey, 2015 CoreLogic Property Tax Database

4.3 Conclusion

Respondents tend to report their lot sizes as larger than shown in the property tax record and are prone to round to one acre when given broad categories. However, this not necessarily problematic, because property tax records are an accurate source of lot sizes (see Section 2).

5. Individual-level correspondence

The aggregate distribution results in Section 4 could be masking a significant amount of disagreement (non-correspondence) between a respondent-reported value and a property tax record at the housing-unit level. Individual-level correspondence is important, because most analyses using lot size data do not focus solely on the aggregate distribution of lot size. For instance, a researcher may be interested in how lot size impacts housing values or rents. A low level of individual-level correspondence, while not impacting the aggregate distribution, could bias joint distributions between lot size and other variables of interest.

This section evaluates the individual-level correspondence between respondent-reported lot size values and the lot size values from matched property tax records. Results are presented across lot sizes and geography to shed some light onto why correspondence rates are not 100 percent.

5.1 How often does the AHS and ACS respondent-reported lot size category correspond to the property tax record category?

Exhibit 5.1 shows the individual-level correspondence between respondent-reported lot size values and the lot size values from matched property tax records. Only AHS is shown due to the limited categories collected in the ACS.

A direct comparison of a lot size value reported by an AHS respondent with the lot size value from the respondent's property tax record reinforces that AHS respondents more often report lot sizes larger than what their tax record shows. About 53 percent of AHS respondents report lot sizes that correspond their tax records. However, about 30 percent report a larger lot size than their tax record shows, while about 17 percent report a lot size smaller than their tax record shows.

Exhibit 5.1 Lot Size Correspondence Rates Between Respondent-reported Value and Property Tax Records

	2015 AHS		
Reported vs Tax Record	**Owners (%)**	**Renters (%)**	**All Tenures (%)***
Respondent reported more than one category smaller than property tax record	3.8	11.1	5.4
Respondent reported one category smaller than property tax record	9.8	18.1	11.5
Respondent reported category corresponds with property tax record	55.6	44.4	53.0
Respondent reported one category larger than property tax record	18.9	13.6	17.8
Respondent reported more than one category larger than property tax record	11.9	12.8	12.1
Total	100.0	100.0	100.0

Note: Rates are calculated using weights to allow for comparison only and are thus approximations.
* Excludes vacant units
Sources: U.S. Census Bureau, 2015 American Housing Survey, 2015 CoreLogic Property Tax Database

5.2 Do lot size correspondence rates vary by size of lot?

Section 5.1 showed that respondents tend to report their lot sizes as larger than what is seen in the property tax record. A reasonable question to ask is whether the correspondence rates also vary across the size of lot.

Exhibit 5.2 shows the correspondence rate by lot size. In the AHS, there is a high rate of agreement for the smallest and largest categories. Generally, there is less correspondence the more refined a category happens to be. This demonstrates that respondents do not know the exact size of the lot. For the ACS, respondents in lots of 1 acre or less typically agree with their matched property tax records.

Exhibit 5.2 Correspondence Rate Variation over the Size of Lot

	2015 AHS			2014 ACS			
Respondent-reported Lot Size Category	Owners (%)	Renters (%)	All Tenures (%)	**Respondent-reported Lot Size Category**	Owners (%)	Renters (%)	All Tenures (%)

Less than 1/8 acre	89.7	89.3	89.6	Less than 1 acre	97.0	95.8	96.8
1/8 up to 1/4 acre	61.1	41.2	57.5				
1/4 up to 1/2 acre	74.8	49.5	70.3				
1/2 up to 1 acre	73.6	42.1	67.9				
1 up to 5 acres	80.6	53.2	75.6	1 up to 10 acres	75.8	56.7	72.8
5 up to 10 acres	80.7	59.0	76.8				
10 acres or more	91.1	69.1	87.2	10 acres or more	71.8	65.7	70.9
Any size	80.8	59.5	77.0	Any size	92.3	91.7	92.2

Note: Rates are calculated using weights to allow for comparison only and are thus approximations.
Sources: U.S. Census Bureau, 2015 American Housing Survey, U.S. Census Bureau, 2014 1-year American Community Survey, 2015 CoreLogic Property Tax Database

5.3 Do correspondence rates vary across geography?

Lot size values from property tax records must be evaluated to ensure that they are being measured the same across each taxing jurisdiction. One way to conduct the analysis is to calculate the share of respondents whose lot size categories correspond to their property tax records at a state-level, then look for outlier states. Exhibit 5.3 shows that the correspondence rates range from approximately 79 to 98 percent.

Exhibit 5.3 Lot Size Correspondence Weighted Rates for ACS by State

State	Rate (%)		State	Rate (%)		State	Rate (%)
Alabama	83.9		Kentucky	87.7		North Dakota	90.9
Alaska	88.7		Louisiana	82.2		Ohio	92.6
Arizona	96.2		Maine	86.9		Oklahoma	89.8
Arkansas	85.5		Maryland	94.1		Oregon	94.7
California	96.0		Massachusetts	91.4		Pennsylvania	93.2
Colorado	95.3		Michigan	90.9		Rhode Island	93.9
Connecticut	90.7		Minnesota	92.2		South Carolina	85.3
Delaware	93.6		Mississippi	82.4		South Dakota	78.9
District of Columbia	96.9		Missouri	92.7		Tennessee	86.1
Florida	95.2		Montana	92.3		Texas	94.2
Georgia	86.1		Nebraska	92.6		Utah	96.1
Hawaii	96.5		Nevada	97.5		Vermont	90.6
Idaho	91.9		New Hampshire	90.4		Virginia	90.6
Illinois	94.5		New Jersey	95.0		Washington	93.5
Indiana	91.4		New Mexico	91.6		West Virginia	89.7
Iowa	92.3		New York	92.9		Wisconsin	90.5
Kansas	94.0		North Carolina	85.8		Wyoming	93.1

Note: Rates are calculated using weights to allow for comparison only and are thus approximations.
Sources: U.S. Census Bureau, 2014 1-year American Community Survey, 2015 CoreLogic Property Tax Database

5.4 Conclusion

The results in Section 5 show that most AHS and ACS respondents provide lot sizes that are in the same category as their matched property tax records. For those that do not agree, respondents tend to provide larger lot sizes than the ones in their matched property tax records. Moreover, the lower rates of correspondence in the more refined categories in the AHS demonstrates that respondents do not necessarily know the exact sizes of the lots.

6. Imputing Missing Lot Size Response Using Local Geographic Distribution of Lot Size

Sections 4 and 5 presented evidence that property tax records are a good source of information for complete replacement of lot size responses via cold decking. However, the cold-deck method is feasible only for respondent addresses that have matched records with lot size values, which account for about 82 percent of AHS and 78 percent of ACS housing units.

This section addresses the remaining roughly18 percent of AHS records and 22 percent of ACS records that do not have a lot size values available from matched records. As mentioned before, the current AHS strategy for missing responses to lot size is to impute through a hot-decking

procedure.[8] In this section, a simple imputation method is proposed based on the best available local cumulative distribution function (CDF) of lot size, which can be derived from the Census block, Census block group, or Census tract. A simulation is then conducted to determine if this approach performs better than the existing AHS approach.

6.1 Review of local variation in lot size

A CDF is the probability that a variable (lot size) takes a value less than or equal to x: $F_x(\chi) = P(X \leq \chi)$. The best available local CDF is the CDF at the lowest level of geography available where data quality and availability to form a CDF meets a certain threshold. The key assumption underpinning the use of best available local CDF is that the best predictor for the lot size of a housing unit is the lot size value from a nearby housing unit. In other words, we assume there is little variation in lot size values for nearby housing units.

To investigate whether this assumption is true, an analysis was conducted on the within-group and between-group variance of lot size values. The overall variance was partitioned by nested Census geographies. The universe of property tax records for the counties in the AHS sample was used, but it was restricted to the universe of property tax records that have a Census block value, which is approximately 95 percent of all property tax records.

Exhibit 6.1 shows that about 21 percent of the overall variation in lot size occurs between housing units within a Census block level, while 10 percent is happening between blocks within a block group, and 22 percent is happening between block groups within a tract. In other words, lot-size values within a block and block group are in fact similar. This stands even when broken down by owner- and renter-occupied units. This result is encouraging because it confirms low variance within smaller levels of geography. Nonetheless, lot size it not normally distributed, so analysis of variance results should be interpreted with caution.

[8] In hot-deck imputation, a household with a missing value for an item (recipient) “borrows” a value from another household who provided a valid response for that item (donor). The hot deck imputation procedure is implemented in a way that attempts to match a recipient household with a donor household based on a common set of characteristics, referred to as the hot deck. In the AHS, the variables that define the hot deck are chosen because they are expected to be correlated, or more generally, they are associated, with the variable being imputed. Before imputation, all records are sorted by an internal variable that contains some geographic information (state and county). This sorting keeps donor and recipient records geographically close to each other.

Exhibit 6.1 Partition of Variance in Lot Size Using Property Tax Data

Geography	Percent of Variance		
	All Units (%)	Owner Occupied Units (%)	Renter Occupied Units (%)
Block	20.8	21.7	20.0
Block Group	10.3	11.0	9.4
Tract	22.4	22.9	21.5
County	22.5	19.8	25.9
Error	24.0	24.6	23.2

Note: Lot size was standardized in logged form.
Source: 2015 CoreLogic Property Tax Database

6.2 Simulation using random draw from best available local CDF

Given the results above, it can now be determined whether imputation based on the best available local CDF performs better than the existing AHS hot-deck approach. One initial test is to simulate how often a random draw from the best available local CDF makes the correct assignment of lot size value. This simulation uses the seven categories as seen in the AHS.

To conduct this simulation, the following steps were performed:

1. Use all property tax assessment records that have a lot size and Census block value and are in counties where there is at least one AHS record (62.2 million property tax records).
2. Calculate the CDFs for lot size (for geographies with >=5 records)[9] for each Census block, block group, and tract.
3. Merge the CDFs to the property tax records.
4. Calculate an imputed lot size for each record by drawing a random number on the uniform distribution and selecting the lot size category corresponding to where the random number falls within the block-level cumulative distribution.
5. Repeat Step 4 process for Census block group and Census tract.

Exhibit 6.2 below shows the results of the simulation broken down by geography. For about 57 percent of Census blocks and 42 percent of Census tracts, the imputed value for lot size, which is based on a random draw from the cumulative distribution function of administrative records in the same geography as the sample unit receiving the imputation, is equal to the actual value.

[9] In other words, CDFs are calculated only when the geography (block, block group, tract) has at least five records. This threshold was chosen based on a series of tests to find the lowest threshold that resulted in the largest improvement of correct imputation in the simulation.

Exhibit 6.2 Results of Simulation of Imputation by Geography Type

Geography	Percent of all property tax records where a CDF is feasible (%)	Percent of records with imputed value equal to actual value (%)
Block	90.3	56.7
Block Group	95.9	45.8
Tract	99.2	42.2

Note: Rates are unweighted.
Sources: U.S. Census Bureau, 2014 1-year American Community Survey, 2015 CoreLogic Property Tax Database

6.3 Conclusion

There is low variation in lot sizes for property tax records in the same block. The results in Section 6 demonstrate that a simple imputation method using the best local CDF does offer good results for cases that do not have matches to property tax records.

7. Whitepaper Conclusion

This whitepaper describes an approach to improve data quality on lot size by performing full replacement cold deck imputation. Section 3 provided evidence that using matched property tax records to impute missing lot sizes is feasible. Sections 4 and 5 demonstrated that, while there are differences between the survey responses and the property tax data, these differences are drive by respondents reporting larger sizes than what are in their matched records and rounding. Section 6 introduced a new method for imputing missing lot size values for AHS records that did not have matched property tax records and demonstrated that this new method provides accurate results.

Given the results of this analysis, HUD elected to develop a new imputation process for lot size for the 2015 AHS and subsequent iterations of the survey. This process is a sequential imputation of lot size values based on the steps below:

- Step 1: Use the exact lot size value from a matching tax record. If not available, then…
- Step 2: Use the imputed value from the local CDF. If not available, then…
- Step 3: Use the respondent-reported values. If not available, then…
- Step 4: Use a hot-deck method, where all prior imputed values are considered valid candidates for the hot deck.

U.S. Department of Housing and Urban Development
Office of Policy Development and Research
Washington, DC 20410-6000

December 2021