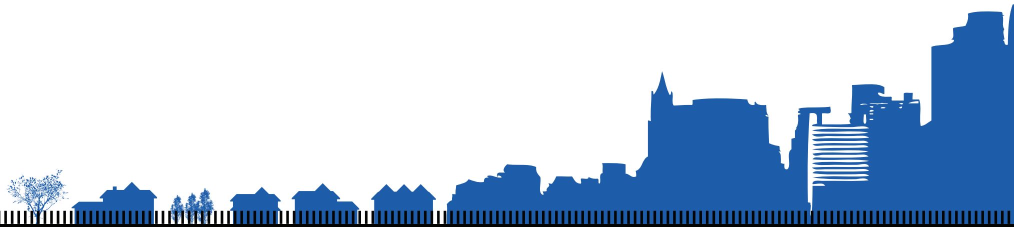
Imputing Year Built with Property Tax Data



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

Imputing Year Built 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 when their housing units were built (hereinafter referred to as year built). There is a high rate of item nonresponse for year built—about 12 percent in the 2015 AHS and about 18 percent in the 2014 ACS. Year built is also not missing at random. Respondents who are more likely not to report a year built are those likely not to know the answer, such as renters and residents in older housing units. Such item nonresponse impacts the total survey error and can result in biased estimates and less statistical power.[2]

To combat the impact of missing data, both the AHS, in survey years 2013 and earlier, and the ACS used hot-deck imputation methodology to fill in missing year-built values. While hot-deck imputation has its advantages, hot-decking can result in biased imputation, especially when missingness is not random and there is a high dimensionality—see Section 6.

Year built 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, roughly 61 percent of AHS and 56 percent of ACS housing units have a matched property tax record with a year-built value. As such, year built could be potentially imputed from property tax records.

This whitepaper presents discussion and analysis that was performed to assess using property tax data to impute year built for respondents who did not provide a valid response. Section 2 provides a general discussion of year built 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 record with a year-built value, concluding that the availability of year built from matched tax assessment records is sufficient to consider it a viable source for imputation.

Sections 4, 5, and 6 lay the evidentiary foundation for using property tax records to replace missing responses—a method known as cold-decking. Section 4 compares the aggregate distribution of year built from respondent-reported values to the aggregate distribution of year built from property tax records, demonstrating that the aggregate distributions are similar. Section 5 evaluates the individual-level correspondence between respondent-reported year-built values and values from property tax records, showing a high level of correspondence. Section 6 presents result from a simulation model comparing hot-deck imputation for year built with the proposed cold-deck imputation methodology for year built, concluding that the cold-deck method outperforms the current hot-deck method with both the AHS and ACS.

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

[2] Weisberg, Herbert F. 2009. *The Total Survey Error Approach: A Guide to the New Science of Survey Research*. University of Chicago Press, and De Leeuw, Edith D. 2001. “Reducing missing data in surveys: An overview of methods.” *Quality and Quantity* 35, no. 2: 147–160.

Section 7 focuses on survey respondents who do not report year built and who do not have a matched property tax report. 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 8 concludes the whitepaper by discussing the findings from the prior sections and the decision to use year built from property tax records for imputing missing AHS responses.

2. Year built as a survey construct and property tax construct

This section presents a discussion of year built in the context of two components of total survey error.[3] First, the section outlines the validity of year built as a construct for both respondents and property tax records. Secondly, it details the reliability of the measurement of year built in both data sources. This whitepaper focuses on the Census Bureau's two largest demographic surveys that collect year-built 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

Both the ACS and AHS define year built as the year the home was first constructed, rather than a year of major remodeling or addition. While the definition of year built is typically unambiguous, respondents can recall the age of their housing unit differently. Besides major additions, the most likely instance of when the year a home was built is subject to interpretation by the respondent is when a home is rebuilt following neglect or natural disaster.

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

For the ACS, the year-built question is presented to respondents using ten-year ranges prior to 2000. For housing units built after 1999, the respondents are asked to provide the specific year. For the AHS, the year-built question is presented to respondents using year ranges prior to 2010. For housing units built after 2009, respondents are asked to provide the specific year. Asking the question in ranges reduces item non-response and reduces the impact of recall bias. Owners can be more likely to forget the exact year built of their unit as time passes following the initial purchase, whereas renters might generally know when their housing unit was built, but not an exact year.[4] Moreover, uncertainty in the exact year built has been shown to result in rounding to a whole or common value by the respondent.[5]

Additionally, the sponsor of the AHS, U.S. Department of Housing and Urban Development (HUD), prefers to use year ranges rather than specific years because HUD believes the

[3] Groves, R.M., F.J. Fowler, Jr, M.P. Couper, J.M. Lepkowski, E. Singer, and R. Tourangeau, R. 2011. *Survey Methodology* (Vol. 561). John Wiley & Sons.
[4] 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.
[5] Manski, Charles F., and Francesca Molinari. 2010. "Rounding Probabilistic Expectations in Surveys." *Journal of Business* & *Economic Statistics* 28.2: 219–231.

reduction in precision from using ranges is offset by the reduction in respondent cognitive and time burden.

Exhibit 2.1 shows the aggregation of year built used in this whitepaper. Homes built in 2010 or later are grouped into one category rather than their individual years, as seen in the data. This allows for results to be directly comparable between the AHS and ACS.

Exhibit 2.1 Year Built Groupings

Year Built
1939 or earlier
1940 to 1949
1950 to 1959
1960 to 1969
1970 to 1979
1980 to 1984
1985 to 1989
1990 to 1994
1995 to 1999
2000 to 2004
2005 to 2009
2010 to 2014

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). In general, the reliability in self-reported year-built responses likely decreases the older the home was at the time of interview because more renovations have occurred and property records have become less dependable.

The universe of both surveys includes owners, renters, and vacant units. Respondents who own their housing units are likely well informed about when their homes were built, especially those who are the first owners or recent buyers of their housing units. The same may not be true of respondents who are renters or individuals responding for vacant housing units.[6] They may be able to infer the year built from their own historical knowledge of the building or from a second-hand source. Nevertheless, a renter or individual responding for a vacant unit would need to try to determine or confirm the year built of their unit by searching through documents they may have or researching the neighborhood. The amount of effort to find the answer is not uniform across respondents.[7]

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

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

2.3 Validity of the construct, from the taxing jurisdiction's perspective

How year built is defined can also vary between taxing jurisdictions. Some jurisdictions record both a year built and an effective year built. Within the property tax data, year built is defined as the construction year of the original building. The effective year built is defined as the first year when a home was assessed with its current components. This is typically used when a home undergoes major additions, but this varies by jurisdiction.

2.4 Reliability of the measurement, from the property tax jurisdiction's perspective

Apart from engraved cornerstones or historical markers on some buildings, taxing jurisdictions cannot directly observe year built. Thus, a tax jurisdiction is dependent on valid and consistent year-built values being recorded. Verifying if the year-built value in a property tax record is correct would require pulling together one or more sources of information to *prove* a home was built in a particular year, such as aerial photography, deed information, permits, or other planning documents.

Measuring the potential impact of these differences and errors at the national level is not possible. Property tax jurisdictions have access to the information and the incentive to make a correct determination. Year built is an important component of the property assessment process, and having an officially recorded year-built value is helpful during a depreciation process.

2.5 Conclusion

While the ACS, AHS, and property tax records intend to capture the year a housing unit was first built, this section has described how this is not always what is captured. There is no easily verifiable "truth" in the measurement of year built. It may be tempting to accept the property tax record as the truth or "more likely to be true" than a respondent-reported value, but the decision to do so would not be based on a statistical assessment. Rather, we have confidence that there is a high level of truth within respondent-reported year-built values.

3. Analysis of the availability of year built in property tax records

A potential imputation technique is to simply replace a non-response with the year-built 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 year built is available from property tax records for sampled AHS and ACS housing units.

All tables in this whitepaper were calculated using the survey's respective household weights. This allows us to compare results between the two surveys, which have different sample sizes and procedures.[8] The weights used in tables are approximations of rates of agreement. Statistical testing is only conducted in Section 4 when comparing estimated distributions.

[8] 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

3.1 Initial review of availability of a year-built value from a matched property tax record

The *availability* of a year-built 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 year-built value on 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 are 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 a 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 a failure 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.[9] For the 2019 AHS, this results 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 year-built missing and year-built availability rates for AHS and ACS records. About 12 percent of AHS records and 18 percent of ACS records do not have a year build response. Yet nearly 61 percent of AHS and 56 percent of ACS records have a matched property tax record with a year-built value. The differences in AHS and ACS year-built availability rates potentially reflect the improved algorithm to match AHS respondents to their property tax records.

Exhibit 3.1 Year Built Availability Weighted Rates

	Percent Without a Respondent-Reported Year-Built Value (%)		**Percent of Respondents Matched to a Property Tax Record with a Year-Built Value (%)**	
	2015 AHS	2014 ACS	2015 AHS	2014 ACS

and https://www.census.gov/programs-surveys/ahs/tech-documentation/def-errors-changes.html for the AHS.

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

All	12.4	18.2	61.0	55.5
Tenure				
Owner	3.0	5.8	77.9	74.4
Renter	27.6	34.7	35.7	28.3
Vacant	13.9	31.0	48.9	41.9
Structure Type				
Single-family detached	6.6	11.4	79.7	75.4
Single-family attached	17.4	17.8	57.7	63.8
Multi-Unit	25.9	33.0	21.2	13.8
Mobile home	10.3	24.1	31.2	28.2

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

Exhibit 3.1 also shows how the year built missing rates and availability rates vary by tenure and by structure type. Owner-occupied units have a lower missing rate of year-built responses and a higher year-built availability rate, primarily due to their concentration in single-family detached housing units, which have a higher response and property tax record matching rate. Vermont did not have year-built data in the property tax records. The results in exhibit 3.1 include Vermont. Thus, all other states have higher rates of availability.

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

Section 3.1 showed that 61 percent of AHS and 56 percent of ACS respondents have matched property tax records with year-built values. This rate indicated that we were capturing a significant portion of survey respondents to consider using tax records to impute missing year-built values. As mentioned in the introduction, about 12 percent of AHS respondents and 18 percent of ACS respondents do not report a year-built value.

Exhibit 3.2 shows the potential cold-deck rates for the AHS and ACS. For the AHS, about 88 percent of respondents report a year-built value, and roughly 5 percent did not report a value, but the year-built value could be imputed with cold-decking. For the ACS, about 82 percent of respondents report a year-built value, and roughly 6 percent did not report a value, but the year-built value could be imputed with cold-decking.

Exhibit 3.2 Potential Cold-Deck Weighted Rates for Year Built

	Owner (%)	**Renter (%)**	**Vacant (%)**	**All Tenures (%)**
2015 AHS				
Respondent-reported value	97.0	72.4	86.1	87.6
Potential cold-deck rate: No respondent-reported value, but matched record with year-built value available	2.1	8.4	6.3	4.7
No respondent reported value or matched record with year-built value available	0.8	19.2	7.6	7.7

Total	100.0	100.0	100.0	100.0
2014 ACS				
Respondent-reported value	94.2	65.3	69.0	81.8
Potential cold-deck rate: No respondent-reported value, but matched record with year-built value available	3.8	8.3	10.5	6.1
No respondent-reported value or matched record with year-built value available	2.0	26.4	20.5	12.2
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

Exhibit 3.2 does include Vermont, which does not have year-built values in the property tax data. Thus, the potential cold-deck rates are higher for all other states.

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 year-built values. While there is variation by structure type and tenure, this analysis is evidence that property tax records can be used to impute missing year-built values.

4. Aggregate distribution correspondence

Section 3 presented results showing that 5 percent of AHS responses and 6 percent of ACS responses do not report year-built values, but they could have a year-built value imputed with cold-decking. For cold-decking using year-built values from matched property tax records to be acceptable, the property tax records must be an unbiased source of information for year built. 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 year-built values and accumulated year-built values 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 respondents' reported year-built values and the year-built values in their property tax records. 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 year built, while low correspondence rates would suggest some possible systematic disagreement.

4.1 Aggregate distribution of year built

The exhibits below compare the aggregate distribution of year-built values from two sources: the respondent-reported year-built values from the housing surveys (AHS and ACS) and year-built values from the respondents' matched property tax records. Exhibits are presented for both owners (exhibits 4.1 and 4.2) and renters (exhibits 4.3 and 4.4).

The results in exhibits 4.1 and 4.2 show a high degree of agreement in aggregate distributions between respondent-reported values in owner-occupied housing units and property tax records. This is true for each of the year-built ranges as well as the cumulative distribution. For the AHS (exhibit 4.1), the cumulative share of the distribution in the older years (1950 and before) is slightly greater for the distribution based on property tax records (29 percent of housing units) compared to AHS respondent-reported values (28 percent). The AHS and property tax year-built distributions for owner-occupied units are not statistically different, χ^2 p=<.0001.

Exhibit 4.1 AHS Aggregate Weighted Distribution of Year Built for Owner-Occupied Units

Year Built	AHS Share of Respondents (90% Margin of Error) (%)	AHS Cumulative Share (%)		Property Tax Share (%)	Property Tax Cumulative Share (%)	Difference in Cumulative Share (%)
1939 or earlier	11.3 (0.6)	11.3		12.1	12.1	– 0.8
1940 to 1949	4.6 (0.3)	15.9		4.9*	17.0	– 1.1
1950 to 1959	12.1 (0.5)	28.0		11.7*	28.7	– 0.7
1960 to 1969	10.9 (0.6)	38.9		10.7*	39.3	– 0.4
1970 to 1979	14.6 (0.6)	53.5		14.0*	53.3	0.1
1980 to 1989	13.9 (0.6)	67.4		13.1	66.4	0.9
1990 to 1999	15.0 (0.6)	82.3		15.2*	81.6	0.7
2000 to 2004	8.2 (0.5)	90.5		8.7*	90.3	0.2
2005 to 2009	7.1 (0.4)	97.6		7.4*	97.7	– 0.1
2010 to 2014	2.4 (0.4)	100.0		2.3*	100.0	0.0

Notes: * signifies that property tax share is within a 90% confidence interval of AHS respondent share.
The AHS and property tax year-built distributions are not statistically different, χ^2 p=<.0001.
Sources: U.S. Census Bureau, 2015 American Housing Survey; 2015 CoreLogic Property Tax Database

For the ACS (exhibit 4.2), the finding is similar. ACS respondents reported fewer older homes than what were in the property tax records. However, the ACS and property tax year-built distributions for owner-occupied units are not statistically different, χ^2 p=<.0001.

Exhibit 4.2 ACS Aggregate Weighted Distribution of Year Built for Owner-Occupied Units

Year Built	ACS Share of Respondents (90% Margin of Error) (%)	ACS Cumulative Share (%)		Property Tax Share (%)	Property Tax Cumulative Share (%)	Difference in Cumulative Share (%)
1939 or earlier	10.2 (0.1)	10.2		11.1	11.1	– 0.9
1940 to 1949	4.8 (0.0)	15.0		4.7	15.8	– 0.8
1950 to 1959	12.0 (0.1)	27.0		11.6	27.3	– 0.3
1960 to 1969	11.0 (0.1)	38.0		11.0*	38.3	– 0.3
1970 to 1979	14.5 (0.1)	52.4		14.0	52.3	0.2
1980 to 1989	13.6 (0.1)	66.0		13.4	65.7	0.4
1990 to 1999	15.4 (0.1)	81.4		15.5*	81.2	0.2
2000 to 2004	9.1 (0.1)	90.5		9.5	90.7	– 0.2
2005 to 2009	7.4 (0.1)	97.9		7.4*	98.1	– 0.2
2010 to 2014	2.1 (0.1)	100.0		1.9	100.0	0.0

Notes: * signifies that property tax share is within a 90% confidence interval of AHS respondent share.
The ACS and property tax year-built distributions are not statistically different, χ^2 p=<.0001.
Sources: U.S. Census Bureau, 2014 1-year American Community Survey; 2015 CoreLogic Property Tax Database

While the results are similar for renters, the disparity is larger. For the AHS (exhibit 4.3), 36 percent of renters report year-built values of 1959 or earlier, compared to 42 percent of property tax records. Year-built distributions between the AHS/ACS and property tax records for renters are not statistically different, χ^2 p=<.0001.

Exhibit 4.3 AHS Aggregate Weighted Distribution of Year Built for Renter-Occupied Units

Year Built	AHS Share of Respondents (90% Margin of Error) (%)	AHS Cumulative Share (%)		Property Tax Share (%)	Property Tax Cumulative Share (%)	Difference in Cumulative Share (%)
1939 or earlier	16.0 (1.3)	16.0		22.0	22.0	– 6.0
1940 to 1949	7.0 (0.9)	23.0		7.2*	29.3	– 6.3
1950 to 1959	12.8 (1.0)	35.8		12.5*	41.7	– 5.9
1960 to 1969	12.4 (1.2)	48.2		10.2	52.0	– 3.7
1970 to 1979	13.3 (1.1)	61.6		11.6	63.6	– 2.0
1980 to 1989	13.0 (1.1)	74.5		11.9	75.5	– 1.0
1990 to 1999	10.4 (1.0)	84.9		9.7*	85.2	– 0.3
2000 to 2004	6.7 (0.8)	91.6		6.7*	91.9	– 0.2
2005 to 2009	6.4 (0.6)	98.0		6.8*	98.7	– 0.7
2010 to 2014	2.0 (0.5)	100.0		1.3	100.0	0.0

Notes: * signifies that property tax share is within a 90% confidence interval of AHS respondent share. The AHS and property tax year-built distributions are not statistically different, χ^2 p=<.0001.

Sources: 2015 AHS internal use microdata; 2015 CoreLogic Property Tax Database

Exhibit 4.4 ACS Aggregate Weighted Distribution of Year Built for Renter-Occupied Units

Year Built	ACS Share of Respondents (90% Margin of Error) (%)	ACS Cumulative Share (%)		Property Tax Share (%)	Property Tax Cumulative Share (%)	Difference in Cumulative Share (%)
1939 or earlier	12.8 (0.1)	12.8		17.4	17.4	– 4.6
1940 to 1949	6.8 (0.1)	19.5		7.2	24.6	– 5.1
1950 to 1959	13.3 (0.1)	32.9		13.4*	38.0	– 5.1
1960 to 1969	11.9 (0.2)	44.8		10.3	48.3	– 3.5
1970 to 1979	16.0 (0.2)	60.8		13.0	61.4	– 0.5
1980 to 1989	13.5 (0.2)	74.3		12.9	74.3	0.1
1990 to 1999	11.8 (0.2)	86.1		10.2	84.5	1.6
2000 to 2004	6.0 (0.1)	92.1		7.4	91.9	0.1
2005 to 2009	6.7 (0.1)	98.7		7.1	99.1	– 0.3
2010 to 2014	1.3 (0.1)	100.0		0.9	100.0	0.0

Notes: * signifies that property tax share is within a 90% confidence interval of AHS respondent share. The ACS and property tax year-built distributions are not statistically different, χ^2 p=<.0001.
Sources: U.S. Census Bureau, 2014 1-year American Community Survey; 2015 CoreLogic Property Tax Database

Comparing AHS results (exhibits 4.1 and 4.3) to ACS results (exhibits 4.2 and 4.4), the share of property tax records in each category does fall in the 90 percent margin of error more frequently than the ACS. Nonetheless, the ACS and property tax records for both renters and owners are not statistically different.

4.2 Aggregate distributional impact under "full replacement where available" scenario

Although the focus of this analysis is using year-built values from property tax records as a source for imputing missing year-built responses, it is important to acknowledge an alternative scenario in which year-built values from property tax records fully replace respondent-reported values where they are available. Exhibit 3.1 showed that 61 percent of AHS respondents and 56 percent of ACS respondents can be matched to property tax records with valid year-built values. Moreover, exhibits 4.1 through 4.4 show that aggregate distributions of year-built values from matching property tax records are similar to the aggregate distributions of property tax records from respondent-reported values.

Exhibits 4.5 and 4.6 show the aggregate distribution of respondent-reported year-built values and year-built values from the respondents' property tax records for the AHS and ACS, respectively. They include owners, renters, and vacant households. Because full replacement would only occur when there is match to the property tax data, only those records are shown.

Exhibits 4.5 and 4.6 show the aggregate distributions to be very similar—as expected given the results in exhibits 4.1 through 4.4. This is true for each of the year-built ranges, as well as the cumulative distribution. For the 61 percent of all AHS respondents where property tax records with valid year-built values are available, replacing the respondent-reported values of year built with the year built from their matched property tax records results in similar aggregate distributions of year built. The same is true for the 56 percent of ACS respondents with matching property tax records and valid year-built values. The difference between the AHS and property tax share of respondents are not statistically different, χ^2 p=<.0001.

Exhibit 4.5 AHS Weighted Aggregate Distribution of Year Built for All Units

Year Built	AHS Share of Respondents (90% Margin of Error) (%)	AHS Cumulative Share (%)		Property Tax Share (%)	Property Tax Cumulative Share (%)	Difference in Cumulative Share (%)
1939 or earlier	12.4 (0.6)	12.4		13.8	13.8	– 1.5
1940 to 1949	5.2 (0.3)	17.5		5.3*	19.2	– 1.6
1950 to 1959	12.1 (0.5)	29.6		11.8*	30.9	– 1.3
1960 to 1969	11.0 (0.5)	40.6		10.6*	41.5	– 0.9
1970 to 1979	14.5 (0.5)	55.1		13.6	55.1	0.0
1980 to 1989	13.6 (0.5)	68.8		12.9	68.0	0.7
1990 to 1999	13.9 (0.5)	82.7		14.2*	82.2	0.4
2000 to 2004	7.9 (0.4)	90.6		8.4	90.6	0.0
2005 to 2009	7.1 (0.4)	97.7		7.3*	97.9	– 0.2
2010 to 2014	2.3 (0.4)	100.0		2.1*	100.0	0.0

Notes: * signifies that property tax share is within a 90% confidence interval of AHS respondent share. The AHS and property tax year-built distributions are not statistically different, χ^2 p=<.0001.

Sources: 2015 AHS internal use microdata; 2015 CoreLogic Property Tax Database

Exhibit 4.6 ACS Weighted Aggregate Distribution of Year Built for All Units

Year Built	ACS Share of Respondents (90% Margin of Error) (%)	ACS Cumulative Share (%)		Property Tax Share (%)	Property Tax Cumulative Share (%)	Difference in Cumulative Share (%)
1939 or earlier	10.8 (0.1%)	10.8		12.0	12.0	– 1.2
1940 to 1949	5.1 (0.0)	16.0		5.0*	17.0	– 1.1
1950 to 1959	12.1 (0.1)	28.0		11.8	28.9	– 0.8
1960 to 1969	11.1 (0.1)	39.1		10.9	39.7	– 0.6
1970 to 1979	14.7 (0.1)	53.8		13.8	53.6	0.2
1980 to 1989	13.6 (0.1)	67.4		13.3	66.9	0.5
1990 to 1999	14.7 (0.1)	82.1		14.8*	81.7	0.4
2000 to 2004	9.4 (0.0)	91.4		9.2	90.9	0.6
2005 to 2009	6.8 (0.0)	98.2		7.4	98.2	0.0
2010 to 2014	1.8 (0.0)	100.0		1.8*	100.0	0.0

Notes: * signifies that property tax share is within a 90% confidence interval of AHS respondent share. The ACS and property tax year-built distributions are not statistically different, χ^2 p=<.0001.
Sources: U.S. Census Bureau, 2014 American Community Survey; 2015 CoreLogic Property Tax Database

4.3 Conclusion

The results in Section 4 show that using matched property tax records as a source of imputation for missing year-built responses would result in potentially minimal impacts on the aggregate distribution of year-built values. In fact, fully replacing respondent-reported values with property tax record values for the AHS, where available, would have little impact on the aggregate distribution of year-built values. The same cannot be said of the ACS, because few year-built categories have a share of property tax records in each category that falls within the 90% confidence interval. This analysis provides evidence that there are low levels of systematic disagreement between property tax and respondent-reported values.

5. Individual-level correspondence

The aggregate distribution results in Section 4 provide some evidence that matched property tax records may be a good source of data for cold-deck imputation. However, the aggregate distribution results 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 year-built data do not focus solely on the aggregate distribution of year built. For instance, a researcher may be interested in how year-built impacts housing values or rents, or how the age of a structure affects the incidence of water leakages or major systems failures. A low level of individual-level correspondence, while not impacting the aggregate distribution, could bias joint distributions between year built and other variables of interest.

This section evaluates the individual-level correspondence between respondent-reported year-built values and the year-built values from matched property tax records. Results are presented across both the age of buildings and geography to shed some light on why correspondence rates are not 100 percent.

5.1 How often does the respondent-reported year-built category correspond to the property tax record category?

Exhibit 5.1 shows the individual-level correspondence between respondent-reported year-built values and the year-built values from matched property tax records. About 75 percent of AHS and ACS respondents who own their homes reported year-built categories that correspond to their property tax record categories. Moreover, for owners, there are a similar number of AHS and ACS respondents who report their housing units as younger than the tax records (11 percent) as respondents who report their housing units are older than the tax records (10 percent). This indicates, for owners, that non-correspondence appears both small and random.

For renters, 58 percent of AHS respondents and 56 percent of ACS respondents report year-built categories that correspond to their property tax records. However, for AHS and ACS renters whose reported year-built responses do not correspond to the tax records, their responses are slightly skewed towards reporting their units as older than the property tax records indicate.

Exhibit 5.1 Correspondence Weighted Rates Between Respondent-Reported Year Built and Property Tax Records

	2015 AHS			2014 ACS		
Reported vs Property Tax Record	Owners (%)	Renters (%)	All Tenures (%)*	Owners (%)	Renters (%)	All Tenures (%)*
Respondent reported more than one category younger than property tax record	3.1	4.5	3.4	3.2	6.0	3.6
Respondent reported one category younger than property tax record	6.7	11.1	7.4	6.9	12.2	7.6
Respondent reported category corresponds with property tax record	79.1	58.0	75.0	78.6	56.2	75.4
Respondent reported one category older than property tax record	7.6	13.9	8.7	7.7	14.5	8.7
Respondent reported more than one category older than property tax record	3.5	12.4	5.1	3.6	11.0	4.7
Total	100.0	100.0	100.0	100.0	100.0	100.0

Note: Rates are calculated using weights to allow for comparison and are thus approximations.

* Excludes vacant units

Sources: U.S. Census Bureau, 2015 American Housing Survey; U.S. Census Bureau, 2014 1-year American Community Survey; 2015 CoreLogic Property Tax Database

5.2 Do year-built correspondence rates vary by age of building?

Section 5.1 showed that 75 percent of AHS respondents report year-built values corresponding with the values in their property tax records. There was also some variation in this rate based on tenure, with owners having a higher correspondence rate than renters. A reasonable question to ask is whether the correspondence rates also vary across the ages of buildings.

It may be reasonable to expect that the rate of year-built correspondence between respondents and their property tax records decreases over time. This would reflect a greater difficulty for respondents in correctly estimating the age of their structure or a lower quality of property tax record information for older structures. Exhibit 5.2 shows that the correspondence rates for structures built in the 1940s (as reported by the respondents), ranges from 39 to 60 percent, while the correspondence rates for structures built after 1950 are relatively stable and range from 68 to 82 percent. The high correspondence rates for structure built before 1940 are a result of the large grouping.

Exhibit 5.2 Correspondence Weighted Rate Variation by the Age of Buildings

	2015 AHS			2014 ACS		
Respondent-Reported Year-Built Category	Owners (%)	Renters (%)	All Tenures (%)	Owners (%)	Renters (%)	All Tenures (%)
1939 or earlier	88.3	89.0	88.4	85.8	81.2	85.1
1940 to 1949	60.1	41.1	56.7	55.3	38.7	52.9
1950 to 1959	74.2	49.2	69.8	70.8	51.0	68.0
1960 to 1969	72.9	41.7	67.4	71.4	41.4	67.2
1970 to 1979	79.3	52.2	74.5	77.3	49.3	73.3
1980 to 1989	77.4	56.5	73.7	79.6	56.2	76.3
1990 to 1999	85.7	60.5	81.2	85.7	56.0	81.5
2000 to 2009	81.6	61.3	78.0	84.4	69.2	82.3
2010 to 2014	80.6	42.2	73.9	81.7	52.2	77.6
Any Year	79.1	58.0	75.4	78.6	56.2	75.4

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?

Year built values from property tax records must be evaluated to ensure that they are being measured the same way across each taxing jurisdiction, even if there is little reason to believe they could be measured differently.

One way to conduct the analysis is to calculate the share of respondents whose year-built categories correspond to their property tax records at a state-level, and then look for outlier states. However, as shown in Section 5.2, the rate of correspondence between respondents and tax records generally diminishes as homes get older. As such, the age of the housing stock within a state will influence the overall rate of correspondence between respondents and their property tax records. Given that the housing stock is younger in some areas of the country, such as the South and West, some adjustments to the state-level analysis are necessary. To control for differences in the age of the housing stock at a state-level, the analysis is conducted using homes built after 1980 as listed in the property tax records.

Exhibit 5.3 shows that the correspondence rates for ACS respondents range from 69.8 percent in Louisiana to 84.2 percent in Connecticut and North Dakota. As previously mentioned, Vermont does not have any year-built values in the property tax records. This compares to a correspondence rate of 81% nationally.

Exhibit 5.3 Year Built Correspondence Weighted Rates for ACS by State for Housing Units Built after 1980

State	Rate (%)		State	Rate (%)		State	Rate (%)
Alabama	77.7		Kentucky	76.2		North Dakota	84.2
Alaska	78.0		Louisiana	69.8		Ohio	83.3
Arizona	80.3		Maine	79.6		Oklahoma	78.7
Arkansas	74.4		Maryland	83.4		Oregon	81.2
California	79.0		Massachusetts	82.2		Pennsylvania	82.4
Colorado	82.1		Michigan	83.2		Rhode Island	79.5
Connecticut	84.2		Minnesota	84.3		South Carolina	77.5
Delaware	81.9		Mississippi	76.3		South Dakota	81.1
District of Columbia	75.8		Missouri	79.6		Tennessee	79.8
Florida	81.0		Montana	78.0		Texas	79.8
Georgia	79.3		Nebraska	84.0		Utah	82.9
Hawaii	71.5		Nevada	80.0		Vermont	--
Idaho	80.0		New Hampshire	80.6		Virginia	83.2
Illinois	80.2		New Jersey	79.2		Washington	81.4
Indiana	81.0		New Mexico	78.8		West Virginia	76.6
Iowa	81.3		New York	78.0		Wisconsin	80.0
Kansas	81.2		North Carolina	81.3		Wyoming	76.8

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 a high-level individual-level correspondence between respondent-reported year-built values and the year-built values from matched property tax records. Section 5.2 also illustrates that lower individual-level correspondence rates for renters reflects the known concern that renters are less likely to know when their units were built.

6. Cold-deck versus hot-deck imputation methods

Section 4 showed that year-built values from respondents have a similar aggregate distribution as year-built values from the respondents' matched property tax records. Section 5 showed agreement in the level of individual-level correspondence between the two sources and that potential systematic disagreements are minimal. Taken together, Sections 4 and 5 provide evidence that property tax records are good sources of data for imputing missing year-built values.

An alternative interpretation of the results in Sections 4 and 5 is that while the aggregate distribution does not show systematic disagreement, there are other sources of systematic disagreement not accounted for at the individual or local level. However, from the perspective of a survey manager, a relevant question is not whether there is systematic disagreement in the property tax records. The more relevant question is whether the use of property tax increase the probability of producing the true value (or what the respondent would have provided) as compared to other methods.

When not using auxiliary data, survey managers often rely on hot-decking for imputation.[10] In fact, the AHS imputed missing year-built values using hot-decking for survey years 2013 and earlier. Hot-deck imputation assigns a missing value from a record from the same survey with similar characteristics. Hot-decking has the advantage of not biasing the aggregate distribution, while also being consistent, easy to implement, and easy to convey to users. Moreover, hot-decking can be implemented in such a way that it preserves some joint distributions with other variables of interest.

However, there is no guarantee that hot-decking results in imputing the correct year-built value for any respondent. Correct values are defined as what the respondents would have provided if they had answered the question. We know year built is not missing at random, and drawing from respondent data as donors, as hot-decking does, might bias the results toward housing unit respondents more likely to know and respond to the year-built question. As such, hot-decking results could be biased, even if the aggregate distribution is not biased.[11]

Moreover, hot-deck imputation faces dimensionality problems. Dimensionality problems occur when there too many cells with small sizes, which increases the chance that a single donor will be used multiple times. To resolve this, one must limit the number of variables included in the process, even if all relate to year built.[12] Thus, hot-deck imputation can risk not all possible correlations being accounted for between the variable being imputed and other confounding variables.

It is important to compare hot-decking to cold-decking using matched property tax records to determine whether cold-decking using property tax records results in more accurate (results with less systematic differences) than hot-decking, as measured by individual-level correspondence. Better accuracy rates for cold-decking would strengthen the record to use such a procedure to impute year built where needed.

[10] 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.

[11] Andridge, Rebecca R., and Roderick J.A. Little. "A Review of Hot Deck Imputation for Survey Non-Response." 2010. *International statistical review* 78, no. 1: 40–64.

[12] Susin, Scott. 2005. "Imputation via Triangular Regression-Based Hot Deck." U.S. Census Bureau. https://www.census.gov/programs-surveys/ahs/research/publications/hotdeck.html.

6.1 A simulation model

To test whether cold-decking using property tax records yields more accurate individual-level values for year built as compared to hot-decking, a simulation model was developed to compare what would happen with cold-decking versus hot-decking. The simulation model was based on the ACS records with both a respondent-reported value for year built and a matching property tax record with a valid year-built value. In the model, the following steps were performed:

- Used the records in AHS sampled counties.
- Drew 250 samples (with replacements) of ACS records with a respondent-reported year-built value and a matched tax record with a year-built value. Each sample was equivalent in size to the AHS sample.
- Changed to non-response a randomly selected, fixed percentage of the ACS records for each sample, meaning the respondent-reported value for year built was erased. The fixed percentage was equal to the AHS nonresponse rate for owners (3 percent) and renters (28 percent). Another reason for setting nonresponse rates for owners and renters separately was that we know the match rates to the property tax records is not random between these two groups.
- Imputed year-built value for the simulated nonresponse records using a hot-decking procedure similar to the AHS procedure.
- Imputed a second year-built value for the simulated nonresponse records using cold-decking—the year-built value from the matched property tax record.
- Compared the original respondent year-built value for the simulated nonresponse record to both the imputed value using the hot-deck method and the imputed value using the cold-deck method.

The results are presented in exhibit 6.1. The first thing to note is that the correspondence rates from cold-decking (80, 56, and 67 percent) are similar to the findings in exhibit 5.1, as expected. Exhibit 5.1 showed that, for AHS respondents who were owners and had matched property tax records with year-built values, the respondent-reported values agreed with the property tax records 79 percent of the time. Note that although the simulation used ACS data, it was based on only the counties in the AHS sample, so the cold-deck results should be consistent with AHS results in exhibit 5.1.

The second important result to note is how poorly the hot-deck method performs relative to the cold-deck method. Recall that this simulation was conducted with ACS records with a respondent-reported value for year built. If that value is treated as “correct,” the hot-deck method will impute a correct response only 12 percent of the time, whereas the cold-deck method will impute a correct response 76 percent of the time.

Exhibit 6.1 Results of ACS-based Simulation of Nonresponse by Tenure

	Percent of Weighted Records with Imputed Value Equal to Actual Value

Imputation Method	Owners (%)	Renters (%)	All (%)
Hot-decking	12.2	12.4	11.8
Cold-decking	79.6	57.3	76.4

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

Similar to renters, respondents in multi-unit structures are more likely not to provide a year built and thus require imputation. We ran similar simulations as described above, but fixed missing rates by structure type. Exhibit 6.2 shows the results of these simulations. Cold-decking performed worse for multi-unit and mobile home records.

Exhibit 6.2 Results of ACS-based Simulation of Nonresponse by Structure Type

	Percent of Weighted Records with Imputed Value Equal to Actual Value				
Imputation Method	Single Family Detached (%)	Single Family Attached (%)	Multi-Unit (%)	Mobile Homes (%)	All (%)
Hot-decking	12.7	14.2	15.3	22.4	11.8
Cold-decking	77.4	75.0	65.6	64.6	76.4

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

There are limitations to this simulation. First, we employed a plausible hot-deck procedure and not the exact procedure used by either survey. Our hot-deck procedure uses a matrix of seven housing characteristics resulting in 127 different donor cells. In practice, the matrix used for hot-decking could be refined with more cells with greater variance. Nonetheless, the difference between hot-decking and cold-decking results in exhibit 6.1 is about 65 percent. It is doubtful that such improvements would close the gap all the way. Second, we only include ACS records that had matched property tax records. In practice, this would only account for one-third of missing results. We address this in Section 7. Third, this process assumes that the actual value given by the respondent is the true value. Because exhibits 6.1 and 6.2 only include matched records, the lower performance of cold-decking for renters and respondents in multi-unit structures demonstrates that there might also be some respondent error occurring.

7. Imputing Missing Responses Using Local Geographic Distribution of Year Built

Sections 4 and 5 presented evidence that property tax records are a good source of information for cold-decking missing responses to the year-built question due to low levels of systematic disagreement. Section 6 showed that imputing missing responses by cold-decking with matched property tax records easily outperformed the traditional existing hot-deck method by imputing a value that matched the true value. However, the cold-deck method is feasible only for non-respondents who have matched records with year-built values. Overall, that amounts to only 5 percent of AHS records and 6 percent of ACS records.

This section addresses the remaining 8 percent of AHS records and 10 percent of ACS records. These records do not have respondent-reported values for year built and do not have year-built values from a matched records. As mentioned before, the AHS strategy for missing responses to year built was to impute via a hot-decking procedure. In this section, a simple imputation method is proposed based on the best available local cumulative distribution function (CDF) of year built, which is derived from the Census block, block group, or tract. A simulation is conducted to determine if this approach performs better than the existing AHS approach.

7.1 Review of local variation in year built

A CDF is the probability that a variable (here year built) takes a value less than or equal to x: $F_x(\chi) = P(X \le \chi)$. The best available local CDF would be at the lowest level of geography available where data quality and availability to form a CDF meet a certain threshold. The key assumption underpinning the use of the best available local CDF is that the best predictors for the year built of a housing unit are the year-built values from nearby housing units. In other words, we assume there is little variation in year-built 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 year-built 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 Census block values, which is approximately 95 percent of all property tax records.

Exhibit 7.1 below shows that 15 percent of the overall variation in year built occurs between housing units within a Census block level, while 7 percent is happening between blocks within a block group, and 25 percent is happening between block groups within a tract. In other words, year-built 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, year built is not normally distributed, so analysis of variance results should be interpreted with caution.

Exhibit 7.1 Partition of Variance in Year Built Using Property Tax Data

	Percent of Variance		
Geography	**All Units (%)**	**Owner-Occupied (%)**	**Renter-Occupied (%)**
Block	15.1	15.2	14.9
Block Group	6.7	6.8	6.7
Tract	25.1	26.4	23.2
County	23.1	20.1	26.8
Error	30.0	31.5	28.4

Source: 2015 CoreLogic Property Tax Database

7.2 Simulation using random draw from best available local CDF

Given the results above, we can now determine 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 year-built value.

To conduct this simulation, the following steps were performed:

1. Use all property tax assessment records that have a year built and Census block value and are located in counties where there is at least one AHS record (62.2 million property tax records).
2. Calculate the CDFs for year built (for geographies with >=5 records)[13] for each Census block, block group, and tract.
3. Merge the CDFs to the property tax records.
4. Calculate an imputed year built for each record by drawing a random number on the uniform distribution and selecting the year-built 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.

For about 51 percent of all property tax records, the imputed value for year built, which is based on a random draw from the cumulative distribution function of admin records in the same geography as the sample unit receiving the imputation, is equal to the actual value. Exhibit 7.2 below shows the results of the simulation broken down by geography.

[13] 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 7.2 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 Weighted Equal to Actual Value (%)
Block	92.7	55.4
Block Group	96.2	40.6
Tract	99.2	36.8

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

7.3 Simulation comparing hot-decking to an imputation base on the best available local CDF

The final test to determine if imputation based on the best available local CDF performs better than the existing AHS approach is to simulate both approaches on actual survey data. Recall from Section 6 that a simulation implementing the current AHS hot-decking produced an imputed year-built value that matched the actual year-built value for only 12 percent of housing units.

This simulation used 2014 ACS data matched with both the individual property tax records and the local CDFs for Census block, block group, and tract. The simulation worked as described below:

- Step 1. Extracted ACS data for counties in the AHS.
- Step 2. Merged ACS housing units with local CDFs for their Census block, block group, and tract.
- Step 3. Drew a sample of ACS records with either a respondent-reported year built or a matched record, which was used if the respondent did not report year built. The sample was equivalent in size to the AHS sample of occupied units (55,000).
- Step 4. Randomly assigned 3 percent of owners and 27.6 percent of renters to be non-respondents, which is equal to the AHS nonresponse rates for owners and renters.
- Step 5. Implemented the current AHS hot-deck procedure for the simulated non-respondents.
- Step 6. Implemented the best local CDF imputation procedure for the simulated non-respondents.
- Step 7. For the simulated non-respondents, compared the respondent-reported (or matched) year-built values to their imputed year-built values from hot-decking and from the best local CDF.
- Step 8. Repeat steps 3 through 7 a total of 250 times.

The best local CDF imputation procedure (step 6) in this simulation differed slightly from how it was implemented in Section 7.2, which only required there to be at least five records in the geography used for the CDF. In addition to requiring at least five records, this simulation also required at least 75 percent of the records in the CDF geography to have non-missing year-built values. In the rare record in which there are not enough eligible admin records at the tract level, the AHS hot-deck value was used.

The results of the simulation are shown in exhibit 7.3. The results show that the best available CDF methodology greatly outperforms the current hot-deck procedure by accurately imputing a missing year-built value two to three times more than hot-decking.

Exhibit 7.3 Results of Simulation by Imputation Method

	Percent of Weighted Records with Imputed Value Equal to Actual Value	
Imputation Method	Owners (%)	Renters (%)
Hot-decking only	16.5	16.9
Best available CDF, then hot-decking	44.5	35.8

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

We ran similar simulations as those described above, but fixed missing rates by structure type. Exhibit 7.4 shows the results of these simulations. Because respondents in multi-unit structures also tend to be renters, simulation results are similar for respondents in multi-unit structures and renters in exhibit 7.3.

Exhibit 7.4 Results of ACS-based Simulation of Non-Response by Structure Type

	Percent of Weighted Records with Imputed Value Equal to Actual Value			
Imputation Method	Single Family Detached (%)	Single Family Attached (%)	Multi-Unit (%)	Mobile Homes (%)
Hot-decking	16.7	22.1	19.6	19.6
Best available CDF, then hot-decking	44.0	52.3	32.6	32.4

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

While the AHS and ACS are sampled at the housing unit level, exhibits 7.3 and 7.4 do not weight the CDF by number of units in the building. At the block level, a single-family housing unit built in 1950 is just as likely to be selected as an apartment building with 50 housing units built in 2015. This could be problematic in areas with that have undergone significant redevelopment. We tested if this was the record by weighting the CDF by the number of units in in the parcel. The results of these are shown in exhibits 7.5 and 7.6. There is little difference in the results when weighting by the number of units versus not weighting.

Exhibit 7.5 Results of Simulation by Imputation Method

	Percent of Weighted Records with Imputed Value Equal to Actual Value	
Imputation Method	Owners (%)	Renters (%)
Hot-decking only	16.2	16.9
Best available CDF, then hot-decking	44.4	36.3

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

Exhibit 7.6 Results of ACS-based Simulation of Non-Response by Structure Type

	Percent of Weighted Records with Imputed Value Equal to Actual Value			
Imputation Method	Single Family Detached (%)	Single Family Attached (%)	Multi-Unit (%)	Mobile Homes (%)
Hot-decking	16.6	22.4	19.5	19.5
Best available CDF, then hot-decking	43.8	52.3	33.8	33.4

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

8. Whitepaper Conclusion

This whitepaper describes an approach to combat missing data and improve total survey error using cold-deck imputation and auxiliary data. Section 3 showed there was nonresponse for the year-built question in the AHS and the ACS and demonstrated that using matched property tax records to impute missing values is feasible. Sections 4 and 5 provide evidence that there are low levels of systematic disagreement between property tax records and respondent reported values. Section 6 showed that for both the AHS and ACS, imputing missing year-built values using matched property tax records outperforms the hot-deck method used in AHS prior to 2015. Section 7 introduced a new method for imputing missing year-built values for AHS records that did not have matched property tax records and demonstrated that this new method outperforms the hot-deck method used in surveys prior to 2015.

Of primary concern is that survey respondents in multi-unit structures are less likely to be linked to property tax records. Section 2 showed how about 5 percent of all AHS records in 2015 did not provide year-built values and had matching property tax records, while 8 percent did not provide year-built values and did not have matching property tax records. In a realm of full replacement, where available, this would be problematic. Section 7 showed how imputing via using the local geographic distribution of year built does result in a value more likely to agree with what the respondent would have responded with.

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

- Step 1: Use the respondent-reported values. If not available, then…
- Step 2: Use the exact year built value from a matching tax record. If not available, then...
- Step 3: Use the imputed value from the local CDF. 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

EQUAL HOUSING
OPPORTUNITY

December 2021