Comparing the 2019 American Housing Survey to Contemporary Sources of Property Tax Records: Implications for Survey Efficiency and Quality

Ariel J. Binder[1], Emily Molfino[2], and John Voorheis[1]
[1] U.S. Census Bureau, 4600 Silver Hill Rd, Suitland, MD 20746
[2] U.S. Department of Housing and Urban Development, 451 7th Street S.W., Washington, DC 20410

Abstract
Given rising nonresponse rates and concerns about respondent burden, government statistical agencies have been exploring ways to supplement household survey data collection with administrative records and other sources of third-party data. This paper evaluates the potential of property tax assessment records to improve housing surveys by comparing these records to responses from the 2019 American Housing Survey. Leveraging the U.S. Census Bureau's linkage infrastructure, we compute the fraction of AHS housing units that could be matched to a unique property parcel (*coverage rate*), as well as the extent to which survey and property tax data contain the same information (*agreement rate*). We analyze heterogeneity in coverage and agreement across states, housing characteristics, and 11 AHS items of interest to housing researchers. Our results suggest that partial replacement of AHS data with property data, targeted toward certain survey items or single-family detached homes, could reduce respondent burden without altering data quality. Further research into partial-replacement designs is needed and should proceed on an item-by-item basis. Our work can guide this research as well as those who wish to conduct independent research with property tax records that is representative of the U.S. housing stock.

Keywords: administrative records, third-party data, housing survey, item replacement and supplementation, respondent burden

1. Introduction

As surveys face declining response rates, the U.S. Census Bureau and other statistical agencies have been exploring ways to supplement their surveys with alternative sources of data. For the American Housing Survey (AHS), one source of interest consists of property and tax assessment records. These are maintained by local jurisdictions primarily for administrative and tax purposes, but they also have information that is collected by the AHS and other housing surveys—e.g. on when the structure was built, its lot size, and its number of bedrooms (Molfino et al. 2017; Weinberg 2015). Each jurisdiction follows their own processes to gather and store these data. Although the data are publicly available, this inherent level of disaggregation makes it difficult to compile a dataset that can be used by national surveys. Fortunately, private vendors aggregate and standardize these records, creating data products available for purchase (Weinberg 2015). The Census Bureau has contracted with multiple vendors over a span of several years, yielding a repository of datasets available for internal research.

In this paper, we explore the potential of property tax and assessment records (hereafter referred to as "property tax records" or "property data") to supplement the AHS. Our work is part of an ongoing collaboration between the US Census Bureau and the Department of Housing and Urban Development (HUD), and extends a broader research agenda at the Census Bureau to incorporate administrative records and third-party data into survey production processes (e.g. Brummet, 2015; Dillon, 2019; U.S. Census

Bureau, 2020). It is also informed by recently developed frameworks for assessing data quality (e.g., Federal Committee on Statistical Methodology (FCSM) 2020; Keller et al. 2017; Agafţei et al. 2015), which suggest the evaluation of third-party data sources along several "fitness-for-use" dimensions with respect to household surveys.

Guided by this framework, our paper has two complementary aims. The first aim is to use the AHS as a basis for evaluating the accuracy, reliability, and coherence of property tax records. These are the core dimensions of the *objectivity* domain laid out in the FCSM (2020) framework.[1] If property tax records are sufficiently objective, they can improve survey *efficiency* by removing the need to ask certain questions to certain subsets of respondents. They could also bolster survey *quality* by providing information that can be used for item response editing and allocation,[2] as well as the refinement of existing survey weights (e.g. Eggleston and Westra 2020; Rothbaum et al. 2021). Such data could also address a longstanding concern among housing survey researchers that respondents, especially non-owners, may not provide accurate information about certain characteristics of their housing unit (e.g. lot size). Finally, property data could be used in the production of small area estimates of housing characteristics of interest to HUD, even if they do not replace survey microdata.

Our second (and related) aim is to assess the reliability of information provided by two different property tax data sources The Census Bureau acquires property tax records using a data acquisition process to solicit proposals from property tax vendors. The accepted proposal is typically for a specific number of years, after which the Bureau solicits a new set of proposals for subsequent acquisitions. Between the AHS survey years of 2017 and 2019, the acquisition process resulted in a different vendor providing files to the Census Bureau. As a result, there was interest in comparing the two sources to assess data reliability between vendors and potential impact on AHS production when a different vendor is used. Since vendor transitions may continue to happen in the future, the implications of this analysis may extend to any repeated housing survey seeking to incorporate property data under vendor uncertainty.

We address these aims first by comparing the two property tax data[3] sources to each other and to state-level housing unit counts published by the Census Bureau's Population Estimates Program. Next, using address-based linkages, we compute the fraction of 2019 AHS housing units that link to a unique property record in each property tax data source (i.e. *coverage rates*). We analyze heterogeneity in coverage rates across states, AHS items, and housing characteristics (namely, different structure types and tenures of ownership). Third, we calculate the extent to which the property tax data and the AHS contain the same information (i.e. *agreement rates*) for each of 11 AHS items of interest. We analyze heterogeneity in agreement rates across states, data sources, and housing characteristics. Finally, we show how the inclusion of linkages based on a newly available geographic shapefile for one source affects that source's coverage and agreement rates.

Our results can be summarized as follows. First, the two sources appear to exhibit imperfect overlap: when matched by parcel number, roughly half of the observations in each source can be found in the other, with substantial variation across states. However, we surmise that this is largely due to different coding of parcel

[1] As defined in Table 1 of the FCSM report, "Accuracy measures the closeness of an estimate from a data product to its true value. Reliability…characterizes the consistency of results when the same phenomenon is measured or estimated more than once under similar conditions. Coherence is defined as the ability of the data product to maintain common definitions, classification, and methodological processes, to align with external statistical standards, and the maintain consistency and comparability with other relevant data" (FCSM 2020, p. 4).

[2] For example, year built and lot size information from one of the property tax sources we analyzed has been used in AHS imputation models since 2015 (Molfino2019a; Molfino 2019b). This analysis helps to inform the discussion of multiple data sources on existing AHS imputation processes.

[3] The Census Bureau's Data-Use Agreement with these property tax vendors prohibits the revealing of the names of either vendor.

numbers between the two data sources: the state-level counts of housing units in each property data source are similar to those published by the Census Bureau.

Second, coverage rates of surveyed housing units in the property tax data are high, but not perfect. Over three-quarters of AHS units can be found in each of the property data sources, with limited variation in coverage rates across states or across data sources. Coverage is above 90% for single-family detached units and for owner-occupied units, but is substantially lower for units in multi-family structures and for rented units.

Third, the extent to which property data sources and the AHS contain the same information is moderate: averaged across all 11 variables, the national agreement rate is 49.7% in one data source, and 44.3% in the other. For *high priority* variables[4]—year built, lot size, and property tax bill—the average agreement rates are 60.3% and 58.7%, respectively. Note that this *unconditional* agreement rate treats missing values as disagreements. *Conditional* on the AHS unit being matched to a non-missing property tax record, the average agreement rates for high priority variables are 72.3% and 72.7%.

Fourth, these agreement rates vary substantially across items and states, but do not vary much between data sources. Single-family and owned structures tend to have the highest agreement rates, although conditional agreement rates for high-priority variables do not exhibit much variation across housing characteristics.

Fifth, exploiting geospatial information for one of the vendors in addition to address information improves coverage rates by a few percentage points. However, it does not improve agreement rates (apart from two-unit apartment buildings).

These results yield two main conclusions. First, given the similar coverage and agreement rates for both property tax data sources, surveys should not be affected by the acquisition-related implications that may impact the availability of the two data sources for internal use. Second, coverage and agreement rates indicate that a full replacement of survey responses with property tax records is not likely be feasible for any of the 11 AHS data items analyzed. However, the high coverage and agreement rates in certain states, as well as for single-family or owned structures, suggest potential to improve AHS efficiency with partial-replacement designs, particularly for the following items: year built, lot size, property tax amount, number of units in building, basement type, garage type, and legal subdivision status.

Before presenting the results, we provide some technical detail on the processes of linking the three data sources together and on harmonizing the variable coding. We conclude by envisioning further research on this subject and discussing implications of the results for AHS survey design. Our study can serve as a guide for other exploratory comparisons between survey and administrative record data, and can also help orient those who wish to conduct representative housing research with property tax records.

2. Overview of Data Sources, Items, and Units of Analysis

This research utilized three main data sources: the 2019 AHS Household internal-use file, 2019 vintage data from one property tax data source (hereinafter S1), and 2017 vintage data from a second property tax data source (hereinafter S2). Discussions with subject matter experts at HUD and Census highlighted 11 different items of interest, as shown in Table 1. Three of these items were of particular interest (i.e. "high priority"): year built, lot size, and monthly property tax amount. Year built and lot size information from S2 are currently used in AHS imputation models, so it is important to assess how the switch to S1 may affect these model estimates. The property tax amount item was also seen as high priority because the

[4] These variables were defined as high priority by subject matter experts working with AHS and property tax data. These can be considered "best case" variables for supplementation or replacement with property tax data.

amount of tax paid contributes to the AHS measure of total housing costs. In general, we expected high coverage and agreement rates for all three items.

Table 1: Summary of Analysis Variables

Variable	AHS name	AHS universe	Priority
Year built	YRBUILT	All	High
Lot size (acres)	LOTSIZE	Single-family units and manufactured homes	High
Monthly property tax	PROTAXAMT	Owner-occupied units	High
Unit size (sqft)	UNITSIZE	All	Other
Bedrooms	BEDROOMS	All	Other
Bathrooms	BATHROOMS	All	Other
Number of units in building	NUNITS	All	Other
Stories in building	STORIES	All	Other
Garage presence and type	GARAGE	All	Other
Basement presence and type	FOUNDTYPE	Single-family units and manufactured homes	Other
In a legal subdivision?	SUBDIV	Single-family unit	Other

There are two critical pre-processing steps in aligning property tax records with household surveys such as the AHS. The first step is to convert property address information from the tax records into Census Bureau address identifiers taken from the Master Address File (MAF).[5] These MAFIDs are assigned by a record-linkage team at the Census Bureau based on exact or probabilistic address text matches between the property data and the MAF. Because household surveys like the AHS directly sample from the MAF, each AHS record already has a MAFID. Therefore, once a unique MAFID has been assigned to a property tax record, that record can be matched directly to a corresponding household survey record.

As shown in Table 2, 59.6% of S1 and 66.6% of S2 records were uniquely matched to a MAF record. This rate is well below 100% because while the MAF consists of *physical* addresses where the housing unit is located, the property data consists of records on any parcel of land where property taxes are paid, including both residential and commercial properties. In contrast, a small share of records (3.1% and 4.2%) were linked to the same MAFID as at least one other record. Most of these duplicate matches are cases in which two adjacent parcels share very similar, or even identical physical addresses (e.g. two condominiums in the same complex, or a household which owns more than one parcel surrounding their domicile). When two or more property tax records shared the same MAFID, we deemed such records as not eligible for matching.

[5] The MAF is the Census Bureau's comprehensive list of residential addresses, which forms the sample universe for all Census-administrated household surveys. It is created by the Geography Division from numerous sources, most importantly the U.S. Postal Service's Delivery Sequence File (DSF) and past decennial censuses.

Table 2: Shares of Property Tax Parcels matched to the Census Master Address File

Matching Status	**Share S1**	**Share S2**
Eligible for matching		
Records matched to MAF and having unique MAFID	59.6%	66.6%
Not eligible for matching		
Record matched to MAF, but shares MAFID with another record	3.1%	4.2%
Record not matched to MAF	37.3%	29.2%
Total	100.0%	100.0%

Source: Third-party property tax records from 2019 (S1) and 2017 (S2); Census Master Address File.

The second pre-processing step involves cleaning and recoding items in the property tax data to align with the coding used by the given household survey. We implemented the following adjustments to align the property data with the coding present in the 2019 AHS.

Numeric variables: year built, lot acreage, property tax amount, unit square footage, bedrooms, bathrooms, and stories.

- Year-built values were rounded to the floor of the nearest decade (e.g. 2017==2010, 1991==1990), with one indicator for 1919 or earlier.
- Lot-acreage values were grouped into the following categories: under 1/8 acre, 1/8 to ¼ acre, ¼ to ½ acre, ½ to one acre, one to five acres, five to ten acres, ten acres or more.
- The AHS published monthly property tax amounts to the nearest dollar, with a top-code for $8300. After converting one source's tax amounts from annual to monthly and imposing a $8300 topcode on the property tax data, we assign agreement based on whether the property data contained a number within 25% of the AHS respondent-reported amount.[6]
- Unit square footage variables were recoded to AHS codes: (<500 sq. ft., 500-749, 750-999, 1000-1499, 1500-1999, 2000-2499, 2500-2999, 3000-3999, 4000+).
- Bedrooms were topcoded at 10, commensurate with AHS.
- For the bathrooms variable, we collapsed the various AHS codes capturing housing units without a full bathroom (codes 7-13) into one code (7). In one property tax data source, bath information was often (but not always) reported in decimal format (e.g. 100==1) and partial bath information was contained in a separate variable. After converting decimal values to regular base-10 values, we constructed the total bath variable of baths + 0.5*partial baths. The other data source contained a clean, base-10-formatted "total calculated baths" variable. For both data sources, we assigned the code 7 if missing or if the actual number of bathrooms was 0.5 or less. Finally, we converted remaining the remaining numerical values into the 1-6 scale used by AHS (1==one full bath, 2==1.5, 3==2, 4==2.5, 5==3, 6==3.5+).
- The stories variable often contained significant digits after the decimal in the property data sources. In one source, there was sometimes trailing text as well that was often a plus sign (e.g. "1.25+") or ("1+A"). We stripped these trailing characters where we could and then rounded the resultant numerical variable to the nearest floor. Then, we translated the resulting integer variable to AHS categories: 1-6 for buildings of up to 6 stories, with a topcode of 7.

Character variables: garage type, basement type and subdivision membership.

[6] This level of agreement used was based on the recommendation of HUD experts and has been used in other HUD agreement analyses based on S1 and S2.

- For all three variables and in both data sources, we assigned the code corresponding to "does not exist" or "not in legal subdivision" if the given information was missing from the given source.
- Both data sources have a subdivision variable that records the subdivision name if one exists: if this information was non-missing, we assigned the code "in a legal subdivision".
- For garage and basement type, the property tax data sources have long and complex sets of codes that do not straightforwardly map to the AHS codes. We were not able to match basement type information for any mobile home unit: codes 5-9 capture foundation types among mobile homes (e.g. 5="mobile home set up on a masonry foundation") yet none of the property tax basement codes mentioned mobile home information.

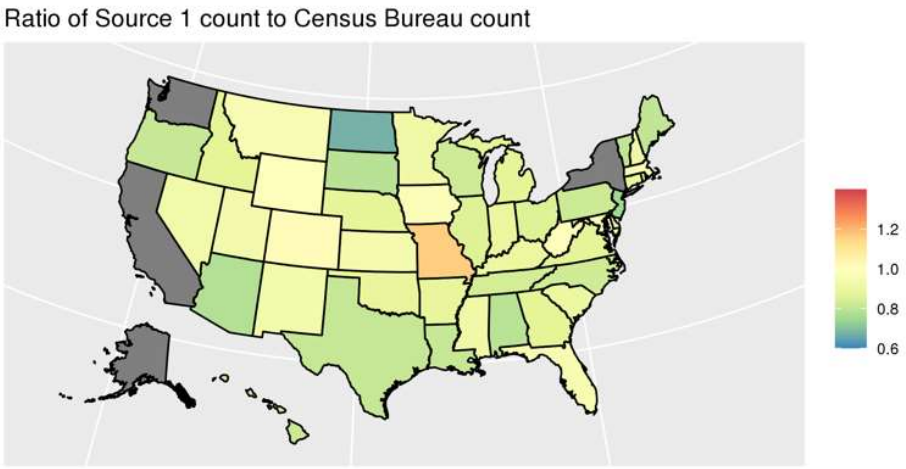
3. Comparing Property Data Sources to Each Other and to the Housing Universe

Before comparing the property data sources to the AHS, we performed two analyses to gauge the property tax data's coverage of the entire housing universe. First, we computed housing unit counts at the state level in each data source and compared these counts to published estimates from the Census Bureau's Population Estimates Program.[7] We compared S1 to 2019 estimates and S2 to 2017 estimates, consistent with each source's vintage.

As noted above, property tax data contain records on any parcel of land where property taxes are paid, including both residential and commercial properties. We used detailed land use codes to construct a binary variable taking the value 1 for residential parcels and 0 otherwise. We multiplied this variable by information on the number of units in the parcel to construct a parcel-level count of the number of residential units. Note that number-of-units information was incomplete: where it was missing or zero for residential parcels, we assigned the value 1. We then summed the number of housing units across parcels to the state level for each data source.

Figure 1 presents the results of this analysis. It shows heat maps recording the ratio of property data housing unit counts to the Census Bureau's published counts, for each of S1 and S2 (where a ratio of 1 indicates that the property data count equals the Census data count). The ratios range between 0.75 and 1 for most states: the median ratio across states is 0.89 for S1 and 0.83 for S2. The few outliers with ratios outside of the 0.6-1.4 range shown on the map are shaded in gray: these include New York (both sources) and California (S1).

[7] See National, State, and County Housing Unit Totals: 2010-2019 (census.gov). 2010 estimates come from the 2010 Decennial Census. Subsequent estimates for 2011-2019 are based on adding in new construction from the Building Permits Survey and the Survey of Construction, adding in new Mobile Homes from the Manufactured Homes Survey, and subtracting out lost housing based on attrition estimates from the American Housing Survey.



Source: Third-party property tax records from 2019 (S1) and 2017 (S2); Census Bureau Housing Unit Totals from 2019 (S1 comparison) and 2017 (S2 comparison).
Notes: States in gray lie outside of the 0.6-1.4 range depicted on the graph and are suppressed for legibility. The national ratio equals the total national count of parcels in the given property data source, divided by the total national Census Bureau count. However, the national ratio for S1 excludes extreme outlier states of Washington and New York.

Figure 1: Housing Unit Counts in Property Data, Compared to Census Bureau Estimates

Discrepancies between the property data counts and Census Bureau counts could exist for two reasons: i) a building in one source does not show up in the other source; ii) a building shows up in both sources but the number of housing units in that building differs between sources. We suspect that the second reason is important in explaining why the property data counts tend to be lower than the Census Bureau counts, given that we assigned the value 1 in cases where number-of-units information was missing.

Missing or inaccurate number-of-units information is also likely responsible for the few outlier cases. In S2, the ratio of the property data count to the Census Bureau count is 0.58 for New York. In S1, the ratio is

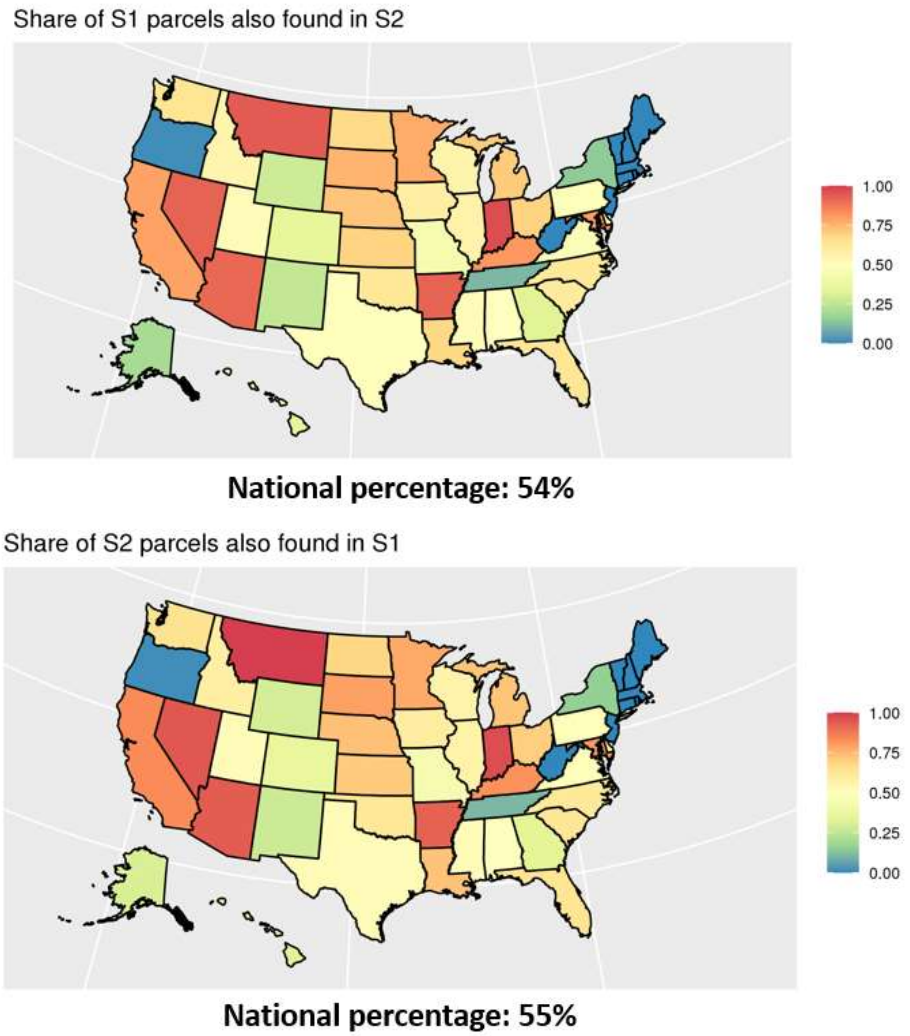
2.17 for Alaska, 1.65 for California, 21.28 for New York, and 3.56 for Washington. As will be seen in Section VI, AHS coverage rates are not nearly as idiosyncratic for these states, suggesting that the dramatic discrepancies seen here is largely an artifact of flawed number-of-units data.

Next, we directly compared the two property data sources by merging them together at the parcel level. We created unique identifiers based on FIPS county codes and Assessor's Parcel Numbers (APNs).[8] We did not use MAFID information for this merge because, as shown above in Table 2, property tax parcels often are assigned missing or nonunique MAFIDs. We then computed overlap rates between the two data sources, i.e. shares of each data source that could be found in the other, at the national and state levels.

Figure 2 presents the results of this analysis. It shows heat maps containing parcel overlap rates between S1 and S2. The top map reports the extent to which S2 overlaps with S1—i.e. the number of parcels found in both data sources as a percentage of the total number of parcels found in S1. The national overlap rate is 54%, but there is substantial heterogeneity across states: overlap rates are zero in all New England states, West Virginia and Oregon, and are low in several other states as well. However, most states exhibit an overlap rate of above 60% (especially in the Midwest), and Indiana, Arkansas, Montana, Nevada and Arizona exhibit near-perfect overlap. The bottom map reports the extent to which S1 overlaps with S2—i.e. the number of parcels found in both sources as a percentage of the number of parcels found in S2. Overlap rates are almost identical at the national and state levels.[9]

[8] Assessor's Parcel Numbers (APNs) are arbitrary parcel identifiers assigned by the county to simplify the identification of parcels. APNs are typically unique at the local level. Yet, there are cases that multiple parcels are under one APN if they are taxed together. Parcels are also static—a parcel will only get a new APN after a significant change to the parcel (e.g., merge or split).

[9] This should be expected, since S1 and S2 contain roughly the same number of parcels at the national and state levels (S1 contains slightly more parcels than S2).



Source: Third-party property tax records from 2019 (S1) and 2017 (S2).

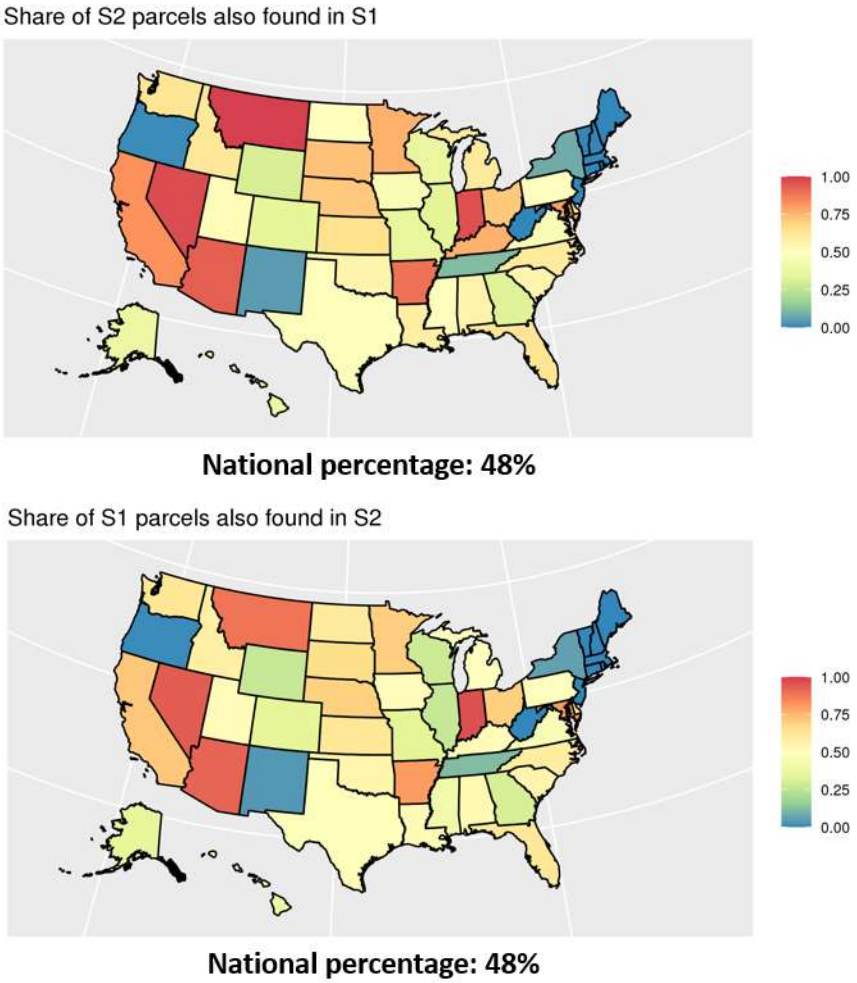
Figure 2: Rates of Parcel Overlap between Property Tax Data Sources

There are two possible reasons for imperfect overlap: i) each source captures different portions of the parcel universe; or ii) each source captures close to the full universe, but parcels are imperfectly matched between the two sources. The prior analysis of housing unit counts already suggests that both sources cover close to the universe of residential parcels. Moreover, both property tax data vendors assemble and standardize information taken straight from public records, so it is reasonable to suspect that the two sources are working with virtually the same underlying information.

To further distinguish between cases i) and ii), we implemented a partial merge, in cases with unique MAFID information, for three states: MA and OR (zero-overlap states) and IN (near-perfect overlap state). We then visually inspected APN strings in these MAFID matches: these are pairs of records that *should* have the same APN information. These inspections revealed no discernible patterns: sometimes the APNs agreed between the two sources and sometimes they did not, and disagreements did not take a consistent form (e.g. a string of 4 numbers in one source would be missing in the other, or would be replaced by a

letter). This further suggests that each data source captures the great majority of the housing universe, and imperfect overlap reflects differences in APN encoding between the two sources.[10]

Because the two data sources differ in their vintage, we reexamined overlap rates within a consistently-defined universe: structures with non-missing year-built values that were built in 2016 or earlier. Figure 3 presents the results. Overlap rates change little—in fact, they decline slightly, perhaps because overlap rates are slightly higher among structures with missing year-built values. The same patterns of state heterogeneity persist as well.



Source: Third-party property tax records from 2019 (S1) and 2017 (S2).

Figure 3: Rates of Parcel Overlap between Property Tax Data Sources, Conditional on Year Built Non-missing and before 2017

4. Aside: Do Property Tax Data Sources Capture New Construction?

S1 has a 2019 vintage and thus could potentially contain all housing units built up through 2019. S2, on the other hand, would miss buildings constructed after 2017. Because the overlap rates changed so little when

[10] Census does not have information from the data vendors on how they process APN beyond what is listed in their respective data dictionary. S1 has only one APN field while S2 has three APN fields: formatted, unformatted, and original. In the APN-based merge, we allowed for matches between S1's APN field and *any* of S2's three fields.

the sample universe was restricted to a range that excluded new construction, we analyzed the coverage of new construction in S1.

According to the Census Bureau's data on new residential construction,[11] a total of 3.59 million housing units were newly constructed in 2017-2019. Among structures with non-missing year-built values in S1, approximately 1.2% had year-built values ranging from 2017 to 2019. The estimated number of housing units in the United States in 2019, according to the Population Estimates Program, was 139.7 million, and S1 contained 97 percent of this number (Figure 1). Then, assuming year built is missing at random, the total number of structures in S1 built between 2017 and 2019 is 139.7 million*0.97*.012=1.63 million. This number is less than half of the official Census Bureau estimate. Thus, it appears that a given vintage of the property data does not accurately reflect new constructions completed within two calendar years of the vintage year.

Part of the reason for this discrepancy could be because S1 and S2 do not collect annual data from all jurisdictions. For example, while S1 is a 2019 vintage, over 7% of the data deliveries (across counties and county-equivalent jurisdictions) came from years earlier than 2019. However, these earlier-than-2019 deliveries are generally from small counties that contain only about 1% of the national population. Therefore, it must be the case that even data deliveries occurring within the vintage year lack currency. We suspect that this is due to lags between when new housing is completed and: i) when property taxes begin to be paid, and ii) when the new property tax record is updated by the County Assessor. One implication of this finding is that these data are not well-suited to supplement survey responses for respondents living in newly constructed housing.

In any case, even supposing that S1 captured only half of new construction, the pre-2017 stock of structures is massive enough that including all new construction should have only a trivial effect on overlap rates.[12]

5. Linking Property Tax Records to AHS Units

As outlined in Section II, linking property tax records to housing units is challenging due to the conceptual differences between the two. The AHS is a survey of *housing units*—i.e., any house, townhouse, apartment, mobile home or trailer, single room, group of rooms, or other location that is intended for occupancy as separate living quarters. Separate living quarters are those in which the occupants live separately from any other people in the structure and have direct access from the outside of the structure or through a common hall, lobby, or vestibule that is used or intended for use by the occupants of more than one unit or by the public.[13] Property tax records, on the other hand, contain all *parcels* of land—i.e., pieces of real estate identified for ownership purposes—that are subject to property taxes. Local jurisdictions record the characteristics of both the parcel and structures on the parcel to determine the correct property tax amount. As such, property tax data contain one record for each parcel.

The difference between the two is best illustrated with a fictitious example of a sampled apartment: 123 Main St #45. The AHS asks the respondent about their specific housing unit. The sample apartment is in a complex that has two buildings with 50 apartments on one parcel, and the physical address of the parcel is 123 Main St. The property tax record for this parcel will include information on the entire lot and on both structures: it will *not* have information specific to the sampled apartment unit.

[11] See New Residential Construction > Historical Data (census.gov)
[12] I.e. under this assumption, the total size of S1 would increase only by 1.2% if all new structures were accounted for. Therefore, the overlap rate of S2 in S1 could decline by at most 1.2%.
[13] See https://www.census.gov/programs-surveys/ahs/about/ahs-introduction-history.html#:~:text=The%20Census%20Bureau%20conducted%20the,to%20the%20American%20Housing%20Survey

Based on this, we developed a linking routine to improve linking AHS sampled housing units to their property tax record. The first stage in the matching process is the direct MAFID match, as described in Section II. The second stage is applied to any record that did not have a MAFID match from the first stage matching process. Matches are made using Census Tract, House Number, and Street Name. The matching process is "fuzzy" in that the match assignments are driven by an algorithm that calculates the likelihood that two text strings match.

This linking routine produced a S1-AHS crosswalk for all AHS units sampled in the 2019, 2017 and 2015 AHS, and a S2-AHS crosswalk for all AHS units sampled in the 2017 and 2015 AHS. S2 was not linked to the 2019 AHS because the Census Bureau had switched to working with S1 at the time of production. Thus, to provide a consistent comparison between S1 and S2, we used the following analysis sample: *all 2019 AHS units that were also sampled in either the 2017 or 2015 survey.* Because the AHS is a longitudinal survey, very few 2019 AHS units were not sampled in prior surveys: the analysis sample contains roughly 95% of the full 2019 AHS sample.[14]

6. Property Data Coverage of the AHS Sample

We now turn to the main analyses of this paper. Having found in Section III that the property tax data sources are not exact replicas of each other, we assess how well each source represents the areas included in the 2019 AHS National sample. We begin by examining *coverage rates*, i.e. the share of AHS units that we can match to a corresponding property tax record. What constitutes good coverage depends on the specific use case. For example, a higher coverage rate is needed when the property data will be used for complete replacement of a survey question versus when the property data is used for response editing or imputation for a specific subgroup of sampled units.

Table 3 records coverage rates by structure type (as classified by the 2019 AHS) and data source. As expected, given the above discussion, coverage rates are highest for single family detached structures and lowest for multi-unit buildings. Moreover, the MAFID match rate is particularly high for single family detached structure (86.0% for S1), with only a small improvement made from the inclusion of the fuzzy address matching algorithm (7.6%, for a total coverage rate of 93.6%). On the other side of the spectrum, the MAFID match rate is quite low for 5+ unit buildings (13.1%), and the fuzzy address matching algorithm raised the coverage rate considerably (by 34.4%, for a total coverage rate of 47.4%). While these patterns and corresponding coverage rates are generally similar across the two sources, S1 does cover a larger share of AHS housing units, particularly in multi-unit structures (S1 covers 6%-8% more units than does S2 in 2-unit, 3-4 unit, and 5+ unit structures).

Table 4 presents coverage rates by tenure type (as classified by the 2019 AHS) and data source. It stands to reason that if an individual housing unit is owned, that housing unit should have a corresponding tax record representing only that unit. The same may not be true of housing units that are rented. Indeed, most apartment buildings are deeded at the building or complex level, rather than the unit level, as indicated in the above example. Accordingly, Table 4 shows that the coverage rate is much larger for owner-occupied units than for renter-occupied units (93.1% versus 57.3% in S1). This pattern matches the heterogeneity by building type reported in Table 3. Moreover, just as with multi-unit structures, the coverage of renter-occupied housing units benefited considerably from the fuzzy address match.

[14] The rate is below 100% because a new sample is added to every AHS wave to account for new construction.

Table 3: AHS Coverage Rates by AHS Structure Type and Data Source

	S1			S2		
Structure Type	**MAFID**	**Address**	**Total**	**MAFID**	**Address**	**Total**
Single Family Detached	86.0%	7.6%	93.6%	90.5%	2.1%	92.6%
Single Family Attached	70.1%	10.3%	80.4%	72.4%	5.0%	77.5%
Multifamily, 2 unit	20.5%	32.1%	52.6%	24.3%	21.9%	46.2%
Multifamily, 3-4 unit	13.4%	35.8%	49.2%	14.3%	26.7%	41.0%
Multifamily, 5+ unit	13.1%	34.4%	47.4%	13.5%	28.2%	41.6%
Other	42.8%	17.0%	59.9%	49.3%	8.5%	57.8%
All Types	62.1%	16.0%	78.1%	65.5%	9.9%	75.3%

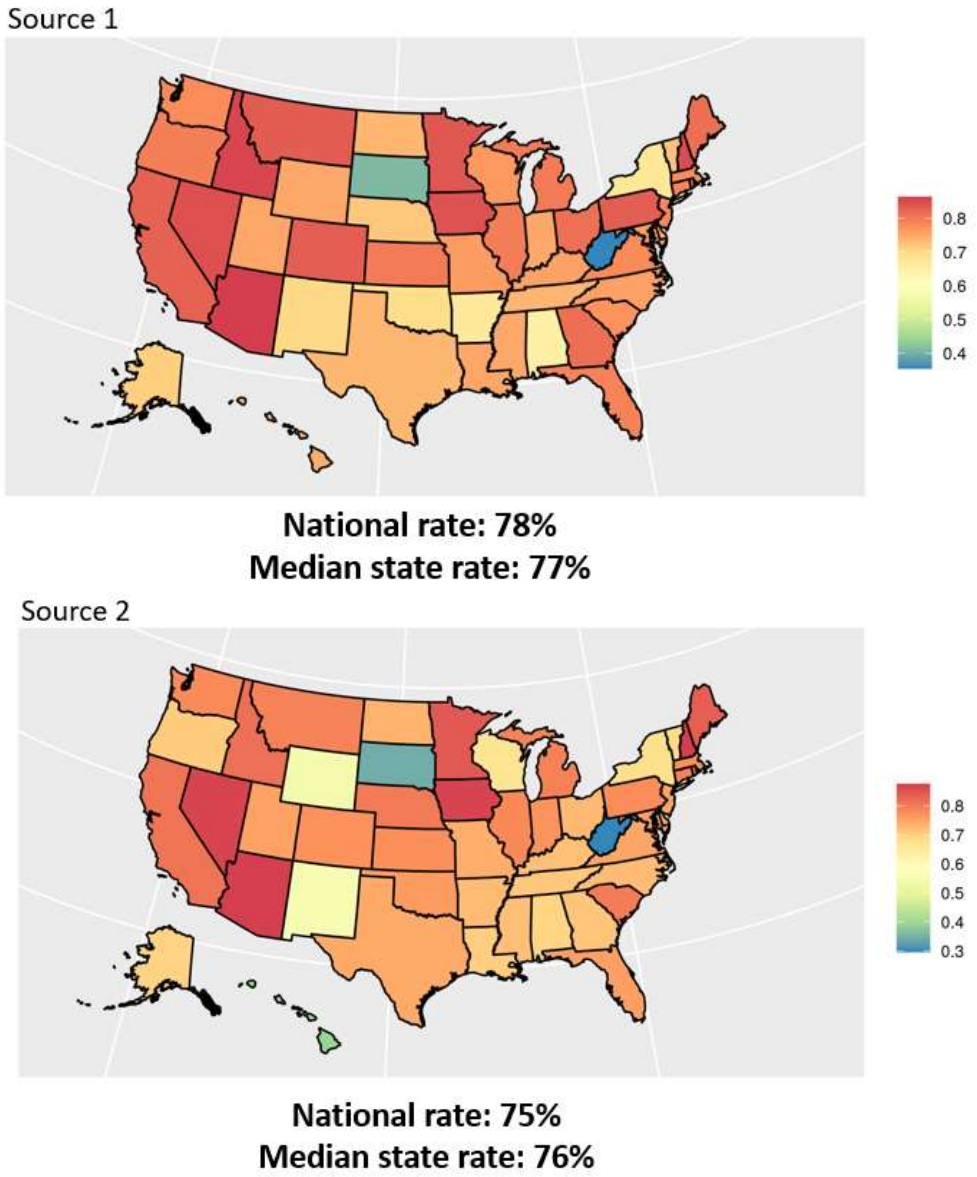
Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).
Notes: "Other" includes mobile homes and structures (boat, RV, van, etc). The coverage rate is defined as the share of AHS housing units that were matched to a unique property tax record. Matches occur in the assignment of property tax parcels to MAFIDs, or via the fuzzy address text matching algorithm.

Table 4: AHS Coverage Rates by AHS Tenure and Data Source

	S1			S2		
Tenure	**MAFID**	**Address**	**Total**	**MAFID**	**Address**	**Total**
Owners	87.5%	5.5%	93.1%	89.3%	2.2%	91.6%
Renters	27.0%	30.2%	57.3%	31.7%	21.1%	52.8%
Vacant/DK/Refused	45.4%	23.2%	68.5%	52.3%	13.1%	65.4%
All Tenures	62.1%	16.0%	78.1%	65.5%	9.9%	75.3%

Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).

Turning to geographical heterogeneity, Figure 4 displays state-level maps of the total coverage rate for each data sources. The coverage rates are reported in tabular format in Appendix Table A1. There is relatively little heterogeneity across states or data sources: most states have coverage rates of at least two-thirds in both data sources (with a few small-state outliers: West Virginia, South Dakota, and Hawaii).

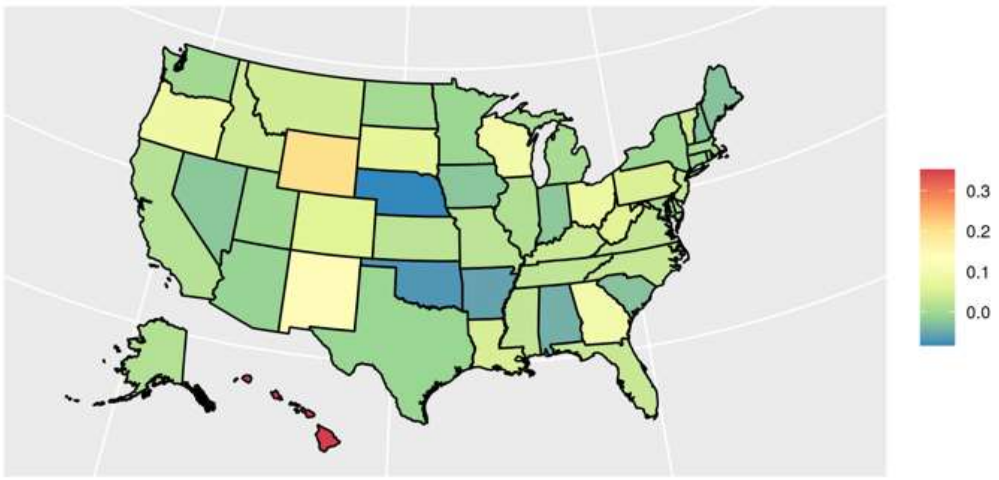


Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).
Note: The coverage rate is defined as the share of AHS housing units that were matched to a unique property tax record.

Figure 4: AHS Coverage Rates by State and Data Source

Figure 5 presents coverage rate differences, expressed as the coverage rate in S1 minus coverage rate in S2. It illustrates substantial agreement between the two sources: the differences range from -0.1 to 0.1 for the vast majority of states (again, with a few small-state outliers).[15]

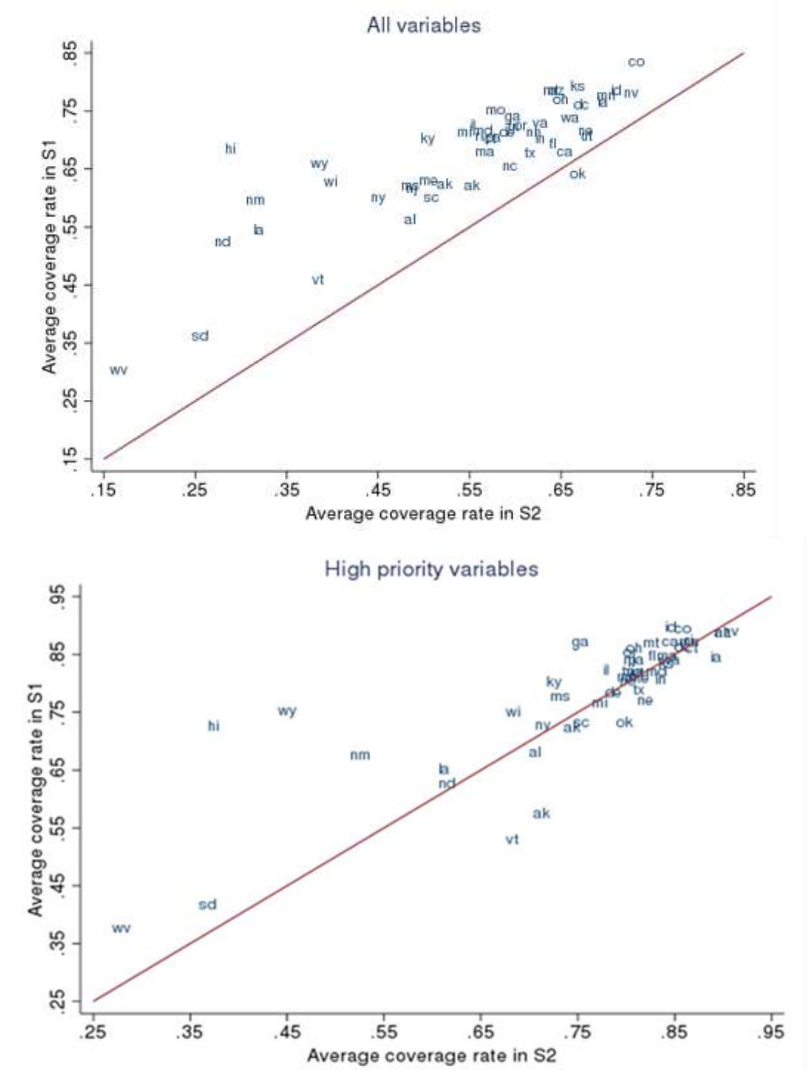
[15] Note that the median difference in coverage rates is not necessarily the same as the difference in median coverage rates. However, the fact that the two statistics are nearly identical and are close to zero (2 percent versus 1 percent) indicates that the two sources have very similar coverage rate distributions.



Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).

Figure 5: Coverage Rate Differences by State and Data Source

Next, we computed coverage rates for each of the 11 survey items displayed in Table 2. For a given state, each individual item could have a coverage rate as high as the housing unit coverage rate reported in Figure 4. However, if some local jurisdictions do not record a certain item or report missing data for another reason, then that item's coverage rate would fall below the unit coverage rate. To summarize this exercise, the top panel of Figure 6 plots S1 coverage rates against S2 coverage rates for each state, where coverage rates are averaged across all 12 variables (11 survey items plus housing unit coverage). The bottom panel repeats the exercise for the 4 high priority variables (3 survey items plus housing unit coverage). Notice that the scales range from .15-.85 for all variables, but .25-.95 for high priority variables.



Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).

Figure 6: Average Coverage Rates for All Variables and for High-Priority Variables

The results illustrate considerable variation in coverage rates across states, but minimal variation across data sources. For example, using S1 as a benchmark, most states lie in the 55%-80% coverage range, although there are several outliers. For high-priority variables, most states lie in the 75%-90% coverage range, again with several outliers.[16] S1 has consistently higher coverage rates than S2 (i.e. most points lie above the 45 degree line), although coverage rate differences are fairly small. This is especially true for high-priority variable coverage rates, where S1 and S2 have equivalent coverage rates across many states.

[16] Notice that average coverage rates for high-priority items often exceed housing unit coverage rates alone. This is not possible if all high-priority survey questions are asked of all survey respondents, since housing unit coverage is a prerequisite for coverage of a survey item associated with that housing unit. However, 2 of the 3 high-priority questions have a limited sample universe: lot size is asked only for mobile and single-family homes, and monthly property tax amount is asked only for owner-occupied units. Because housing unit coverage rates are particularly high for single-family homes and owner-occupied units (Tables 3 and 4), the result is a higher average coverage rate for high-priority survey items than for housing units in general.

Table 5 presents national coverage rates for each of the 11 AHS items. National coverage rates range from 39% to 91% in Source 1 and 11% to 90% in Source 2. Within each variable, there is substantial coverage rate heterogeneity across states, as shown by the state min(imum), state med(ian), and state max(imum) columns. Property tax records in a few small states, when combined with small numbers of AHS sample housing units, did not contain any information on most of the survey items of interest—resulting in minimum coverage rates of zero. However, most median coverage rates are above 60%, and maximum coverage rates tend to be above 80%.

Table 5: Coverage Rates for 11 AHS Items, by Property Tax Data Source

	Source 1				Source 2			
	Nat'l rate	State min	State med	State max	Nat'l rate	State min	State med	State max
Year built	0.70	0.00	0.73	0.84	0.65	0.00	0.69	0.86
Lot size	0.91	0.37	0.90	0.97	0.88	0.31	0.88	0.96
Tax amount	0.88	0.00	0.89	0.97	0.90	0.36	0.90	0.98
Unit Size	0.79	0.40	0.78	0.87	0.68	0.00	0.71	0.86
Bedrooms	0.78	0.35	0.77	0.86	0.47	0.00	0.52	0.81
Bathrooms	0.78	0.35	0.77	0.86	0.20	0.00	0.03	0.66
Units in building	0.78	0.35	0.77	0.86	0.69	0.04	0.72	0.87
Stories in building	0.60	0.00	0.67	0.83	0.58	0.00	0.62	0.85
Basement type	0.39	0.00	0.46	0.89	0.11	0.00	0.03	0.83
Garage type	0.43	0.00	0.40	0.69	0.42	0.00	0.38	0.66
Subdivision	0.54	0.00	0.57	0.97	0.69	0.00	0.74	0.96

Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).

7. Item Agreement Rates

The coverage analyses suggest that while full national replacement of survey items by property tax information is infeasible, there is potential to supplement certain items with property data through partial replacement. To assess this possibility further, it is important to analyze the extent to which property tax data sources contain the same information as the AHS—and how these rates of agreement vary across building type, geography, and survey items. If they do not agree, a partial-replacement design could introduce inconsistencies and spurious variation between AHS units that are surveyed versus those for which information is filled in from property tax records. This is especially true of there is an underlying pattern to the partial-replacement design, e.g. if certain areas or building types are systematically more likely to receive property tax data values. Note that this issue is akin to bias arising from nonrandom non-response.

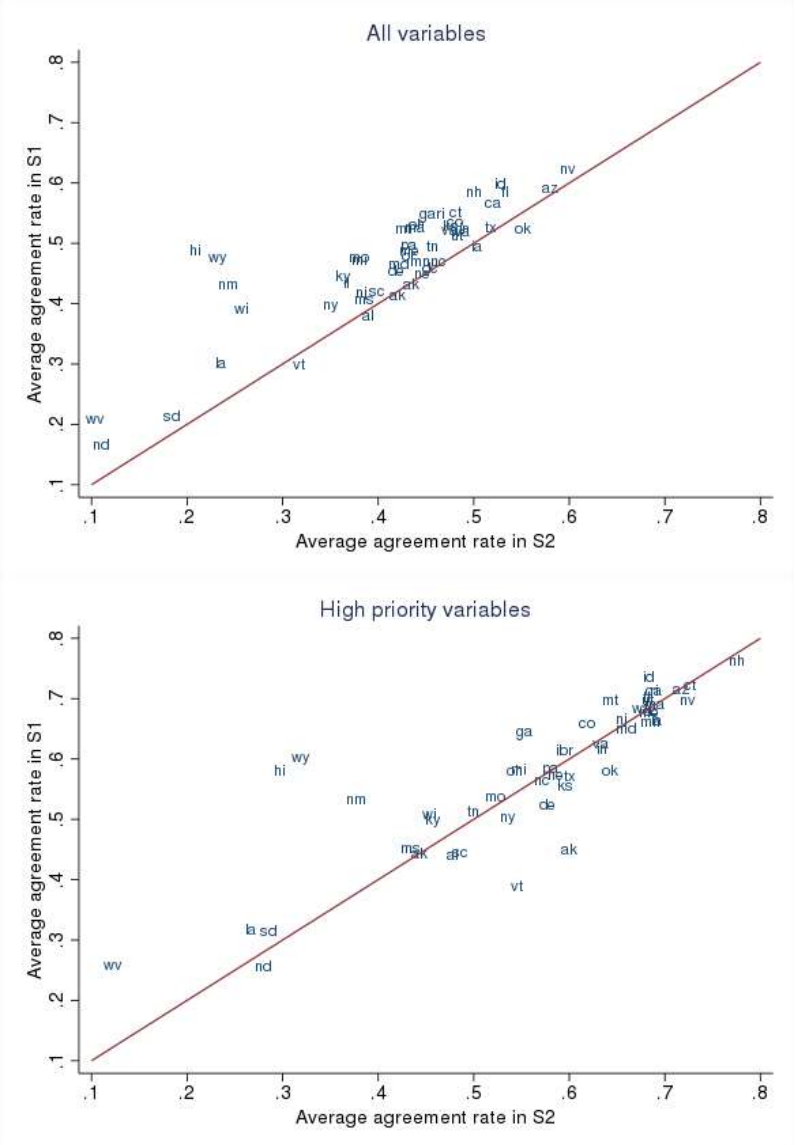
We began by aligning the coding of the property tax information with the schemas used by AHS, as described in Section II. With these alignments in place, we prepared two sets of agreement rates: *unconditional* and *conditional* agreement rates. For a given item and data source, the *unconditional* agreement rate is the share of nonmissing AHS cases where the property tax record contains the same coded value as the AHS record. The *conditional* agreement rate is the share of nonmissing AHS cases in which the AHS item and the property tax item are the same, conditional on the property tax record also being nonmissing. That is, if the coverage rate is c and the unconditional agreement rate is u, the conditional agreement rate, r, is equal to u/c.

There are multiple reasons why disagreement between survey and property data may occur: the respondent-reported value may be wrong, the tax assessment record may be entered or parsed incorrectly,[17] or the wrong property record may be linked to a given AHS housing unit. Without a "ground truth" source of information, it is impossible to distinguish among these reasons. No matter the reason, disagreement would alter AHS estimates if survey responses were replaced by property tax records. The purpose of comparing agreement rates by source is not to assess which source is "more correct," but to assess if similar rates of coverage and disagreement are seen. If so, then changing from one source to another should not measurably affect AHS estimates. If not, then caution must be taken in supplementing a longitudinal survey like AHS with property data, because the Census Bureau may continue to acquire different property data sources in the future.

The top panel of Figure 7 plots S1 unconditional agreement rates against S2 unconditional agreement rates for each state, where agreement rates are averaged across all 11 variables. The bottom panel repeats the exercise for the 3 high priority variables. There is substantial heterogeneity in unconditional agreement rates across states. Using S1 as a benchmark, most states lie in the 40%-55% agreement range, with several small-state outliers. Agreement rates for S1 are generally higher than they are for S2. Average agreement rate differences are small, however, ranging from 0%-10% for most states. High-priority agreement rates are higher than average agreement rates for every single state—most states lie within a 55%-70% band (with the same small-state outliers as before). S1 and S2 contain very similar high-priority agreement rates; most points cluster near the 45-degree line, although there are four cases in which S1 has substantially higher coverage than S2 (Hawaii, Wyoming, New Mexico, West Virginia).

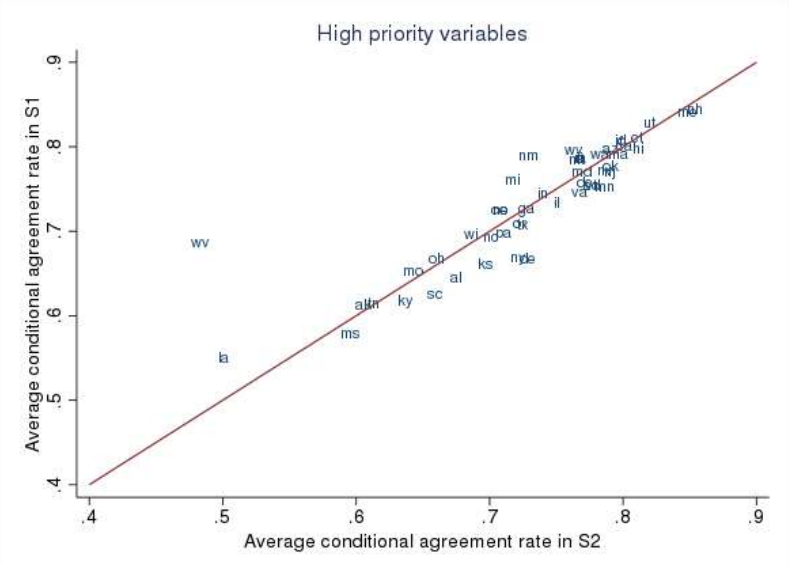
Figure 8 plots average conditional agreement rates, which equal the ratios of the unconditional agreement rates to the coverage rates. Recall that for most non-high priority variables, coverage rates were zero for several small state outliers (Table 6), resulting in missing conditional agreement rates. To ensure a consistent population of states and variables, the figure only shows average conditional agreement rates for high priority variables, and exclusive of North Dakota and Alaska (which have zero coverage rates for at least one high priority variable). Relative to unconditional agreement rates (Figure 6), conditional agreement rates are higher and more narrowly distributed, with most states in the 70%-85% range. Nonetheless, conditional agreement rates are still well below 1, and moderate heterogeneity still exists across states. Therefore, imperfect agreement between the AHS and property tax data stems from a combination of imperfect coverage as well as imperfect agreement conditional on coverage.

[17] For example, as mentioned in Section II, the raw data sometimes contain text characters in fields where one would only expect numbers to be present. In addition, the raw data for numerical variables sometimes occur in decimal format (e.g. 100==1).



Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).

Figure 7: Average Unconditional Agreement Rates for All Variables and for High Priority Variables



Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).

Figure 8: Average Conditional Agreement Rates for High Priority Variables (excluding AK and ND)

Table 6 presents national agreement rates for each of the 11 AHS items, together with state mins and maxes. Unconditional national agreement rates range from 32% to 72% in Source 1 and 11% to 72% in Source 2. Looking at conditional rates and restricting focus to high priority variables only (and excluding North Dakota and Alaska), national agreement rates are 76% for year built, 79% for lot size, and 62% for tax amount. These rates, as well as state median agreement rates, vary little across data sources. While agreement is still well below 100%, these rates are substantial in size and suggest that it is feasible to use property tax information to supplement these high-priority survey items. State maximum unconditional agreement rates are at least 75% for most items investigated, suggesting scope for supplementation of all 11 items with property tax data in certain states of the country.

Table 6: Agreement Rates for 11 AHS Items, by Property Tax Data Source

	Source 1				Source 2			
	Nat'l rate	State min	State med	State max	Nat'l rate	State min	State med	State max
Panel A. Unconditional Agreement Rates								
Year built	0.54	0.00	0.54	0.69	0.50	0.00	0.52	0.67
Lot size	0.72	0.35	0.73	0.95	0.72	0.10	0.77	0.94
Tax amount	0.55	0.00	0.54	0.75	0.54	0.17	0.55	0.76
Unit size	0.35	0.02	0.31	0.54	0.34	0.00	0.30	0.54
Bedrooms	0.34	0.00	0.36	0.54	0.33	0.00	0.35	0.53
Bathrooms	0.36	0.00	0.34	0.55	0.11	0.00	0.10	0.23
Units in building	0.65	0.32	0.65	0.80	0.56	0.00	0.59	0.73
Stories in building	0.32	0.00	0.24	0.64	0.32	0.00	0.22	0.74
Basement type	0.52	0.09	0.48	0.83	0.38	0.01	0.24	0.82
Garage type	0.57	0.04	0.59	0.77	0.56	0.04	0.57	0.75
Subdivision	0.55	0.03	0.52	0.95	0.51	0.11	0.54	0.95
Panel B. Conditional Agreement Rates (excluding ND and AK)								
Year built	0.76	0.64	0.77	0.86	0.76	0.60	0.75	0.84
Lot size	0.79	0.44	0.86	0.99	0.82	0.31	0.89	0.99
Tax amount	0.62	0.46	0.64	0.79	0.60	0.39	0.62	0.85

Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).
Note: The unconditional agreement rate, for a given survey item, is defined as the share of AHS housing units for which there is a corresponding property tax record with the same value for that item (where lack of coverage counts as disagreement). The conditional agreement rate considers only those AHS units that were matched to a nonmissing property tax record: the conditional rate is the unconditional rate divided by the coverage rate. North Dakota and Alaska are excluded from Panel B due to zero coverage for at least one of the considered survey items.

Next, we investigated heterogeneity in agreement rates by building and tenure type, guided by the coverage analysis reported in the previous section. Table 7 presents unconditional and conditional agreement rates by data source and structure type. Within each data source, the first column reports average agreement rates across all 11 survey items of interest, while the second column reports average agreement rates across the 3 high-priority items. Once again, unconditional agreement rates are highest for single-family housing units. For detached structures, average unconditional agreement with S1 is nearly 60% across all survey items, and is nearly 70% for high-priority items. For units in multi-unit structures, average unconditional agreement hovers around 15% for all items and 30% for high-priority items. Agreement rates are similar for S2.

It is reasonable to expect this pattern of results, given that missing values count as disagreements and coverage rates were by far the highest among single-family units. Panel B controls for variation in coverage across structure types by reporting conditional agreement rates. Indeed, there is less variation in conditional agreement rates than in unconditional agreement rates. This is especially true for high-priority items, where conditional agreement rates are around 75% for single-family units and 55%-70% for multi-unit buildings. However, taking all survey items into account, average conditional agreement rates are around 66% for single-family units but only 31%-35% for multi-unit buildings. While these results are encouraging for single-family units, they indicate a need for better property-to-housing-unit linkage for multi-unit structures.

Table 7: Agreement Rates by Structure Type and Property Tax Data Source

Structure Type	Share	Source 1		Source 2	
		All	High priority	All	High priority
		Panel A. Unconditional Agreement Rates			
Mobile home	4.5%	0.255	0.343	0.217	0.313
1-family, detached	60.2%	0.597	0.682	0.537	0.677
1-family, attached	6.0%	0.523	0.613	0.467	0.581
2 apts	3.4%	0.178	0.381	0.139	0.328
3 to 4 apts	4.6%	0.162	0.364	0.122	0.297
5 to 9 apts	5.4%	0.140	0.310	0.113	0.269
10 to 19 apts	5.1%	0.129	0.296	0.107	0.259
20 to 49 apts	4.4%	0.160	0.361	0.128	0.305
50+ apts	6.4%	0.171	0.378	0.137	0.324
		Panel B. Conditional Agreement Rates			
Mobile home	4.5%	0.462	0.650	0.516	0.648
1-family, detached	60.2%	0.655	0.756	0.674	0.761
1-family, attached	6.0%	0.658	0.755	0.683	0.747
2 apts	3.4%	0.343	0.706	0.360	0.684
3 to 4 apts	4.6%	0.330	0.644	0.342	0.582
5 to 9 apts	5.4%	0.321	0.567	0.342	0.529
10 to 19 apts	5.1%	0.320	0.575	0.344	0.525
20 to 49 apts	4.4%	0.325	0.629	0.345	0.563
50+ apts	6.4%	0.315	0.623	0.348	0.613

Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).

Table 8 presents analogous agreement rates by tenure type. Unconditional agreement rates are much larger in owned than in rented housing units, consistent with the fact that owned units are much more likely to be linked to a unique property tax record. Unconditional agreement rates are around 58% for all items and 66% for high-priority items among owned units, but only 30% for all items and 49% for high-priority items among rented units (S1). In contrast, conditional agreement rates are much narrowly distributed across tenure types, even more so than the patterns reported in the previous Table. Conditional agreement rates are around 64% for all items and 74% for high-priority items among owned units, compared to 52% for all items and 75% for high-priority items among rented units (S1). The similarity in conditional agreement across tenure types can partially be explained by the fact that although owners are likelier to live in single-family homes than renters, they are also likelier to live in mobile homes which tend to have lower agreement rates.

Table 8: Agreement Rates by Tenure of Ownership and Property Tax Data Source

Tenure Type	Share	Source 1		Source 2	
		All	High priority	All	High priority
		Panel A. Unconditional Agreement Rates			
Owners	54.0%	0.581	0.659	0.517	0.646
Renters	33.1%	0.300	0.491	0.268	0.467
Vacant/DK/Refused	12.9%	0.374	0.581	0.337	0.563
		Panel B. Conditional Agreement Rates			
Owners	54.0%	0.643	0.739	0.661	0.741
Renters	33.1%	0.523	0.756	0.586	0.765
Vacant/DK/Refused	12.9%	0.558	0.823	0.618	0.847

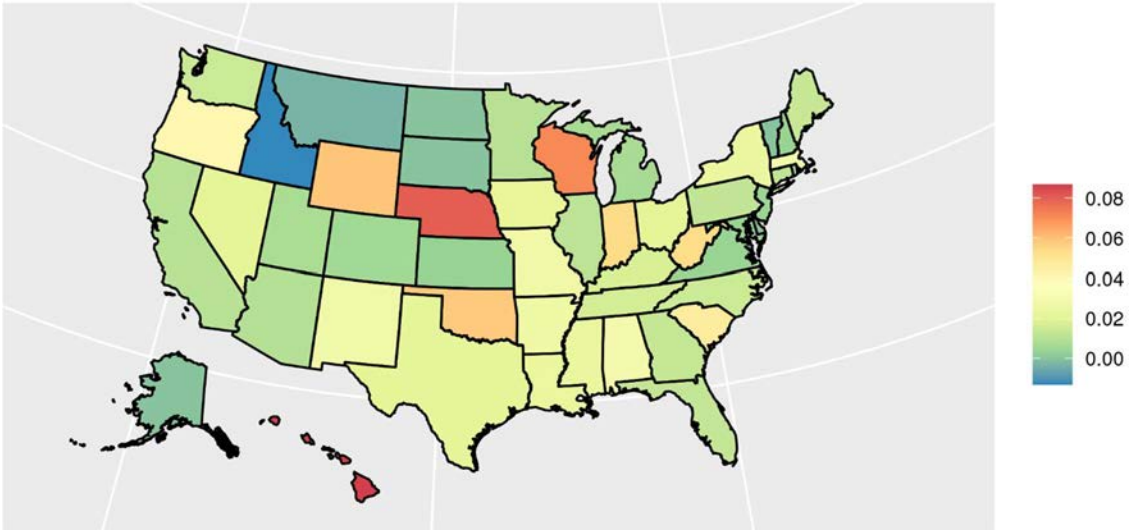
Source: 2019 American Housing Survey combined with property tax records from 2019 (S1) and 2017 (S2).

8. Incorporating Geospatial Data into the Linking Routine

The linking method up to this point relies on address information only. As we saw, this resulted in low coverage rates for multi-unit buildings, because an AHS housing unit in a single-building apartment complex often has a different address than the one listed for the property parcel. Though the fuzzy address matching routine increased coverage rates particularly for units in these multi-unit buildings, coverage was still very imperfect.

In this section, we analyze how coverage and agreement rates improve when we incorporate geospatial data into the linking routine. S1 recently made available a geographic shapefile, which allows for geospatial coordinate matches to be made to AHS in addition to address matches. This is done by assessing whether a given AHS unit, using its latitude and longitude coordinates, lies inside a given parcel's geospatial polygon from the shapefile. A match is declared when the latitude-longitude coordinate point falls anywhere inside the polygon's boundaries. These geospatial matches were then validated by comparing the Basic Street Address (BSA) of the parcel's property address to the BSA of the AHS unit. Buffers of varying distances were used to account for error in the geographic coordinates and parcel boundaries. Census Bureau staff incorporated this information into a new linking routine, which constructs S1-AHS links first based on geospatial information. If no exact geospatial match exists, matches were then identified from MAFID and address information as before.

Figure 9 plots the increase in S1's coverage rate that this augmented linking routine confers over the address-only linking routine. For most states, coverage rate differences are quite small, although for a few, the addition of geospatial data increases coverage by 7-9 percentage points.



Source: 2019 American Housing Survey combined with property tax records from 2019 (S1).
Note: The graph records the difference in S1 coverage rates between the enhanced algorithm that includes geospatial matches in addition to address-based matches, and the simpler algorithm that just includes address-based matches (shown in Figure 4).

Figure 9: Additional Coverage Contributed by Geospatial Joins in SI

Tables 9 and 10 compare S1's coverage rate with and without the geospatial data across structure and tenure types. Just as in Figure 9, coverage rate improvements are generally minimal. For single-family homes, coverage rate gains are only around 1%-2%, likely because these structures had a very address-based coverage rate to begin with. Duplexes did see a moderate improvement in coverage of nearly 8%, which suggests a potential use case for geospatial data. On the other hand, for AHS units in buildings of 10 or greater units, the enhanced linking routine provided almost no increase in coverage rate. This could be a symptom of how the geospatial join validation is done, as larger structures are likelier to have a property address with a different BSA than that of the sampled housing unit. Improvement in how geospatial joins are validated may help further increase the property tax data's coverage of multi-unit structures.

Table 9: S1 Coverage Rates by Structure Type, With and Without Geospatial Joins

	Share	Without Spatial Join	With Spatial Join	Difference
Mobile home	4.5%	59.6%	63.2%	3.5%
1-family, detached	60.2%	93.2%	95.4%	2.2%
1-family, attached	6.0%	81.2%	82.2%	1.0%
2 apts	3.4%	52.7%	60.3%	7.6%
3 to 4 apts	4.6%	49.4%	52.5%	3.1%
5 to 9 apts	5.4%	42.6%	44.2%	1.5%
10 to 19 apts	5.1%	38.8%	39.6%	0.8%
20 to 49 apts	4.4%	49.7%	50.2%	0.4%
50+ apts	6.4%	55.1%	55.6%	0.5%

Source: 2019 American Housing Survey combined with 2019 property tax records.

Table 10: S1 Coverage Rates by Tenure, With and Without Geospatial Joins

	Share	Without Spatial Join	With Spatial Join	Difference
Owners	54.0%	93.1%	94.5%	1.4%
Renters	33.1%	57.3%	59.7%	2.4%
Vacant/DK/Refused	12.9%	68.5%	71.3%	2.8%

Source: 2019 American Housing Survey combined with 2019 property tax records

Finally, we investigated whether these slight coverage improvements translated into improvements in item agreement rates. Table 11 records average agreement rates for all items, and for high-priority items, by structure type. Panel A shows that inclusion of spatial joins raises unconditional agreement rates by a few percentage points in small apartment buildings (2-4 units), for both all variables and high-priority variables, but otherwise exerts no effect. Panel B reports that conditional on coverage, agreement rates slightly *worsen* once geospatial join links are included. Recall that the linking routine prioritized geospatial links first, and resorted to address-based links in the event of no geospatial match. This pattern of results suggests that although geospatial links help raise coverage rates to a modest extent, they may be slightly less accurate on average than address-based links—resulting in slightly lower conditional agreement rates. Table 12 records average agreement rates by tenure, with and without geospatial join links, and displays a similar pattern of results.

Table 11: S1 Average Agreement Rates by Building Type: With and Without Spatial Joins

Stucture Type	Share	All items		High-priority items	
		Without Spatial Join	With Spatial Join	Without Spatial Join	With Spatial Join
		Panel A. Unconditional Agreement Rates			
Mobile home	4.5%	0.255	0.258	0.343	0.341
1-family, detached	60.2%	0.597	0.605	0.682	0.689
1-family, attached	6.0%	0.523	0.525	0.613	0.615
2 apts	3.4%	0.178	0.202	0.381	0.434
3-4 apts	4.6%	0.162	0.171	0.364	0.380
5-9 apts	5.4%	0.140	0.144	0.310	0.317
10-19 apts	5.1%	0.129	0.129	0.296	0.295
20-49 apts	4.4%	0.160	0.160	0.361	0.358
50+ apts	6.4%	0.171	0.170	0.378	0.373
		Panel B. Conditional Agreement Rates			
Mobile home	4.5%	0.462	0.455	0.650	0.638
1-family, detached	60.2%	0.655	0.653	0.756	0.754
1-family, attached	6.0%	0.658	0.655	0.756	0.752
2 apts	3.4%	0.343	0.340	0.705	0.701
3-4 apts	4.6%	0.330	0.330	0.645	0.650
5-9 apts	5.4%	0.320	0.321	0.567	0.569
10-19 apts	5.1%	0.321	0.320	0.574	0.571
20-49 apts	4.4%	0.325	0.323	0.629	0.611
50+ apts	6.4%	0.315	0.315	0.623	0.615

Source: 2019 American Housing Survey combined with 2019 property tax records.

Table 12: S1 Average Agreement Rates by Tenure: With and Without Spatial Joins

Stucture Type	Share	All items		High-priority items	
		Without Spatial Join	With Spatial Join	Without Spatial Join	With Spatial Join
		Panel A. Unconditional Agreement Rates			
Owners	54.0%	0.581	0.585	0.659	0.662
Renters	33.1%	0.300	0.307	0.491	0.500
Vacant/DK/Refused	12.9%	0.374	0.383	0.581	0.592
		Panel B. Conditional Agreement Rates			
Owners	54.0%	0.643	0.641	0.739	0.735
Renters	33.1%	0.523	0.520	0.756	0.748
Vacant/DK/Refused	12.9%	0.558	0.553	0.823	0.814

Source: 2019 American Housing Survey combined with 2019 property tax records.

9. Conclusion

Several research initiatives are underway at the Census Bureau and other federal statistical agencies to preserve data quality and reduce respondent burden in era of rising nonresponse rates to household surveys. At the same time, growing availability of administrative records presents intriguing opportunities for supplementation or replacement of household survey items, as well as to create experimental data products that rely less heavily on survey responses and the timetables of survey processing.

In this paper, we explored the fitness-for-use of property tax records, which are kept by county assessor offices and aggregated by third-party vendors, to supplement or replace certain items in the AHS. We had two related aims. First, given that previous waves of the AHS already used one data source in its response-editing models, and the Census Bureau recently switched to working with another source, we aimed to assess the similarity of the two data sources. This could inform whether future AHS waves would be impacted by this switch. Second, we aimed to provide a broad overview of coverage and agreement rates between the AHS and property tax data, with attention to heterogeneity across survey items, geographies, and housing characteristics.

Our findings indicate that property data are reliable across vendors. Although the internal coding of parcel numbers varies considerably between the two sources, the housing-unit-level data on addresses and characteristics are similar. This suggests that AHS data quality should not be affected by transitions between the two vendors analyzed here, although future work is needed to understand if these results generalize to other vendors. In neither source do we find that a complete replacement of survey responses with property tax records would be feasible for any of the survey items studied. However, the high coverage and agreement rates for certain items, as well as more generally for single-family and owned structures, suggest the viability of partial-replacement designs.

How might such designs proceed? One consistent finding is that coverage and agreement rates do not exhibit a discernible pattern by state population or region. Researchers should be sensitive to heterogeneity across states and survey items in production processes and in data analyses. In some cases, a generalized process for all states or for all survey items may be warranted, while in others, a state-by-state or item-by-item process may be necessary. In addition to year built and lot size, the following four items had state-median unconditional agreement rates of greater than 50%, and should be the subjects of more specialized future research: property tax amount, number of units in building, garage type, and legal subdivision status.

In summary, our results suggest that there are high returns to continued study of the supplementation of the AHS and other housing surveys with selected items contained in property tax records. There appears to be real potential to implement partial-replacement designs that may reduce respondent burden without altering data quality. Data quality may even improve for items where respondents have a history of providing imprecise responses. To deliver on this potential, further fitness-for-use research is needed on an item-specific basis. Analysts will need to answer the following two questions, separately for each item: 1) Are there reliability concerns in how the AHS or property tax assessments capture a particular concept? 2) Do agreement rates meet quality thresholds needed for the particular use case? This approach is how the AHS began its imputation work with property data for the year built and lot size items (see Molfino 2021a, Molfino 2021b).

References

Agafţei, Mihaela, Fabrice Gras, Wim Kloek, Fernando Reis and Sorina Vâju. 2015. “Measuring Output Quality for Multisource Statistics in Official Statistics: Some Directions.” *Statistical Journal of the International Association of Official Statistics* 31: 203-211.

https://content.iospress.com/download/statistical-journal-of-the-iaos/sji902?id=statistical-journal-of-the-iaos%2Fsji902

Brummet, Quentin. 2015. "Matching Addresses between Household Surveys and Commercial Data." CARRA Working Paper Series WP-2015-04. https://www.census.gov/content/dam/Census/library/working-papers/2015/adrm/carra-wp-2015-04.pdf

Dillon, Michaela. 2019. "Use of Administrative Records to Replace or Enhance Questions about Housing Characteristics on the American Community Survey." American Community Survey Research and Evaluation Report Memorandum Series ACS18-RER-02. https://www.census.gov/content/dam/Census/library/working-papers/2019/acs/2019_Dillon_01.pdf

Eggleston, Jonathan and Ashley Westra. 2020. "Incorporating Administrative in Survey Weights for the Survey of Income and Program Participation." Social, Economic and Housing Statistics Division Working Paper Series SEHSD-WP-2020-07. https://www.census.gov/library/working-papers/2020/demo/SEHSD-WP2020-07.html.

Federal Committee on Statistical Methodology. 2020. "A Framework for Data Quality." FCSM 20-04. Federal Committee on Statistical Methodology. September 2020. https://nces.ed.gov/FCSM/pdf/FCSM.20.04_A_Framework_for_Data_Quality.pdf#:~:text=Te%20FCSM%20Data%20Quality%20Framework%20provides%20a%20common,on%20the%20quality%20of%20data%20products%20and%20outputs.

Keller, Sallie, Gizem Korkmaz, Mark Orr, Aaron Schroeder, and Stephanie Shipp. 2017. "The Evolution of Data Quality: Understanding the Transdisciplinary Origins of Data Quality Concepts and Approaches." *Annual Review of Statistics and Its Application* 4.1: 85-108. https://sites.nationalacademies.org/cs/groups/depssite/documents/webpage/deps_191048.pdf

Molfino, Emily. 2021. "Imputing Lot Size with Property Tax Data." US Department of Housing and Urban Development Office of Policy Development and Research. https://www.census.gov/content/dam/Census/programs-surveys/ahs/working-papers/Imputing-Lot-Size-Property-Tax-Data.pdf

Molfino, Emily. 2021. "Imputing Year Built with Property Tax Data." US Department of Housing and Urban Development Office of Policy Development and Research. https://www.census.gov/content/dam/Census/programs-surveys/ahs/working-papers/Imputing-Year-Built-Property-Tax-Data.pdf

Molfino, Emily, Gizem Korkmaz, Sallie A. Keller, Aaron Schroeder, Stephanie Shipp, and Daniel H. Weinberg. 2017. "Can Administrative Housing Data Replace Survey Data?" *Cityscape* 19.1: 265-292. https://www.huduser.gov/portal/periodicals/cityscpe/vol19num1/ch16.pdf

Rothbaum, Jonathan, Jonathan Eggleston, Adam Bee, Mark Klee, and Brian Mendez-Smith. 2021. "Addressing Nonresponse Bias in the American Community Survey during the Pandemic using Administrative Data." American Community Survey Research and Evaluation Report Memorandum Series ACS21-RER-05. https://www.census.gov/library/working-papers/2021/acs/2021_Rothbaum_01.html

U.S. Census Bureau. 2020. "Agility in Action 3.0: A Snapshot of Enhancements to the ACS." https://www.census.gov/programs-surveys/acs/operations-and-administration/agility-in-action/agility-in-action-3.html

U.S. Department of Housing and Urban Development, Office of Policy Development and Research (HUD). 2013. HUD Research Roadmap, FY 2014–FY 2018. http://www.huduser.org/portal/pdf/Research_Roadmap.pdf.

Weinberg, Daniel H. 2015. "Data Sources for US Housing Research, Part 2: Private Sources, Administrative Records, and Future Directions." *Cityscape* 17.1: 191-206. https://www.huduser.gov/portal/periodicals/cityscpe/vol17num1/ch14.pdf

Appendix: State Coverage Rates Table

Table A1 demonstrates that coverage of AHS sample units in property tax records varies by state and source. S1 has higher coverages in 32 of the states—coverage of S1 is on average 5.6% higher for these states. In the 19 states where S1 coverage is lower, coverage is on average 2.5% lower.

Table A1: Coverage of AHS Sample in Property Tax Data by Source

State	Observations in Analysis Sample	S1 Coverage	S2 Coverage	S1 minus S2 Coverage
AL	820	65.0%	69.5%	-4.5%
AK	80	71.1%	69.7%	1.3%
AZ	2,370	86.3%	87.5%	-1.2%
AK	500	66.7%	72.5%	-5.8%
CA	8,780	82.6%	81.1%	1.6%
CO	2,250	83.2%	77.3%	5.9%
CT	420	78.8%	79.5%	-0.7%
DE	210	74.5%	74.0%	0.5%
DC	270	77.0%	78.5%	-1.5%
FL	4,640	79.0%	75.7%	3.3%
GA	2,840	81.1%	71.4%	9.7%
HI	170	74.6%	39.3%	35.3%
ID	80	85.5%	81.6%	3.9%
IL	2,020	79.5%	78.3%	1.3%
IN	1,220	75.0%	77.1%	-2.1%
IA	360	84.3%	86.8%	-2.5%
KS	1,000	79.8%	77.6%	2.2%
LA	920	75.7%	72.1%	3.6%
LA	2,800	75.2%	70.4%	4.8%
ME	150	81.1%	83.8%	-2.7%
MD	1,470	76.0%	75.7%	0.3%
MA	1,800	77.9%	77.7%	0.2%
MI	3,250	79.8%	79.4%	0.4%
MN	710	83.8%	84.5%	-0.7%
MS	680	75.0%	72.2%	2.8%
MO	1,700	76.2%	74.1%	2.1%
MT	300	83.2%	79.2%	4.0%
NE	340	71.9%	80.3%	-8.4%
NV	340	84.5%	86.8%	-2.3%
NH	300	84.8%	87.5%	-2.7%
NJ	1,190	78.4%	75.2%	3.2%
NM	220	70.0%	55.6%	14.3%
NY	2,440	67.2%	67.9%	-0.6%

NC	3,420	75.8%	72.5%	3.3%
ND	50	73.6%	73.6%	0.0%
OH	3,910	81.0%	72.8%	8.2%
OK	350	69.3%	76.2%	-7.0%
OR	2,260	79.5%	70.8%	8.7%
PA	3,320	83.0%	77.9%	5.1%
RI	140	80.9%	78.7%	2.2%
SC	420	77.1%	80.0%	-2.8%
SD	120	41.0%	34.4%	6.6%
TN	2,230	73.7%	71.1%	2.6%
TX	5,800	73.5%	74.5%	-1.0%
UT	270	75.2%	75.6%	-0.4%
VT	130	72.3%	68.5%	3.8%
VA	1,950	78.0%	75.1%	2.9%
WA	3,040	77.9%	78.1%	-0.2%
WV	220	35.4%	29.6%	5.8%
WI	3,220	76.7%	66.8%	9.9%
WY	70	75.0%	54.4%	20.6%
National	77,560	78.1%	75.4%	2.7%

Source: 2019 American Housing Survey.
Note: observation counts have been rounded for disclosure avoidance purposes.