Can Administrative Housing Data Replace Survey Data?

Emily Molfino
Gizem Korkmaz
Sallie A. Keller
Aaron Schroeder
Stephanie Shipp
Daniel H. Weinberg
Virginia Tech

Abstract

This article examines the feasibility of using local administrative data sources for enhancing and supplementing federally collected survey data to describe housing in Arlington County, Virginia. Using real estate assessment data and the American Community Survey (ACS) from 2009 to 2013, we compare housing estimates for six characteristics: number of housing units, type of housing unit, year built, number of bedrooms, housing value, and real estate taxes paid. The findings show that housing administrative data can be repurposed to enhance and supplement the ACS, but limitations exist. We then discuss the challenges of repurposing housing administrative data for research.

Introduction

Many federal statistical agencies provide national and regional survey data. Limited sample sizes driven by limited budgets, however, prevent these agencies from providing timely data at a small geographic granularity. The challenges behind collecting survey data in an era of declining public cooperation have led federal statistical agencies to explore the potential benefits and drawbacks of using data that are external to their organizations and the federal system for supplementing official statistical products. For this article, the term *external data* is defined as data from local governments collected for administrative purposes designed to meet the operational needs of the locality where they are collected.

This article addresses the following question: "Can non-federally collected local government administrative data enhance, supplement, or replace federally collected survey data?" The extensive existing literature on housing economics and housing policy regularly depends on using survey data from sources such as the American Community Survey (ACS) and the American Housing Survey (AHS; Weinberg, 2015, 2014). When used for research, however, these surveys can be limiting due to their relatively small sample sizes in any particular geographic area and the timing of their administration (Bazuin and Fraser, 2013; Jarosz and Hofmockel, 2010; Weinberg, 2015). We address that question in this article through a detailed case study using housing data from one locality encompassing 5-year period. As such, although these results have limited generalizability, the article can serve as a guide to a better understanding of some alternatives to survey data when studying certain policy issues.

Recent studies have examined the use of administrative data by attempting to match these data to census records. Ruggles (2015) provides a review of administrative data sources that could be used to replace or improve specific questions in the ACS. The author considered potential data sources, including federal, state, and local administrative records for taxes and benefit programs, private sector records, and third party data aggregators. She found that several questions in the ACS could potentially be replaced with matched or imputed data, especially housing-related questions that collect information that overlaps with data found in local property tax records (for example, year built, number of acres, value of the unit, and property taxes paid). Moore (2015) investigated the feasibility of replacing or supplementing the year built question on the ACS with administrative records. The author found that the match between the year built category on the 2012 ACS and vendor data (2006–2010) is 78.3 percent when linked using Master Address File Identification Numbers, or MAFIDs, and 76.0 percent when linked using basic street addresses.

Building on this body of work, this article examines the feasibility of using local administrative data by comparing estimates based on local housing administrative data for six housing characteristics (number of housing units, housing unit type, year built, number of bedrooms, housing value, and real estate taxes paid) to the respective estimates from the ACS. The ACS is a large monthly survey carried out by the U.S. Census Bureau. It provides annual estimates on many topics for areas and population groups of 60,000 or more, and it provides estimates aggregated over a 5-year period for smaller jurisdictions, census tracts, census block groups, and population subgroups with populations less than 60,000.[1] The geographic region for this study is Arlington County, Virginia, an urban county that is part of the Washington, D.C. metropolitan area. The housing research reported in this article is part of a broader study sponsored by the Census Bureau focused on leveraging external non-federally collected data sources to enhance official statistics and products (Keller et al., 2016). Our results demonstrate that local housing administrative data can be repurposed to produce estimates similar to (or better than) ACS tabulations for three of the six comparisons of housing data: year built, housing value, and real estate taxes paid. Although the other three comparisons (number of housing units, housing type, and number of bedrooms) did not result in estimates that fully align with those of the ACS, the lack of alignment cannot be attributed to one source of data being better than the other. We will provide some explanations for the misalignment. These results show that local housing administrative data can replace or supplement the ACS estimates in one locality, but limitations and challenges exist that must be considered.

[1] In contrast to the ACS, the AHS, the key data set for much housing research, provides only biennial estimates for national, regional, and key metropolitan areas. For more information of alternate sources of housing data, see Weinberg (2015, 2014). For more information on the ACS, see https://www.census.gov/programs-surveys/acs/.

The remainder of this article is organized as follows. The Data and Methods section describes the data used, the data preparation required for this study, and the metrics used to compare the estimates. The Comparisons of Housing Characteristics section provides comparisons between the estimates that were obtained using real estate assessment data and the ACS for six selected housing characteristics. The Limitations and Opportunities of Using Local Administrative Data section presents insights about the potential of using housing administrative data that we discovered during our research.

Data and Methods

Real estate assessment data for Arlington County, Virginia (hereafter, Arlington) from 2009 to 2013 were used in this study. These data were acquired directly from the Real Estate Department of the Arlington County government and from CoreLogic, Inc., a commercial vendor. These data will be referred to as *AC assessment data* and *CL assessment data*, respectively.

Arlington's Department of Real Estate collects AC assessment data for administrative and tax purposes. CoreLogic acquires real estate assessment data across the country directly from the jurisdictions. It then repackages the data into common formats, and sells them for a variety of uses, such as studying market trends, property valuation, or housing policy. Recent research has looked into the potential of CoreLogic data specifically. Brummet (2014) matched the Census Bureau Master Address File (MAF) to housing units in three different sources of data: (1) the 2009 AHS, (2) commercial data obtained from CoreLogic for 2009, and (3) administrative records from 2010 and the 2011 Tenant Rental Assistance Certification System (TRACS) obtained from the U.S. Department of Housing and Urban Development (HUD). The author found that the commercial data from CoreLogic matched the MAF at a much lower rate (64 percent) than the other data sets (2011 HUD TRACS matched at 90 percent, for example), with the survey data matching at the highest rate of 92 percent. Kingkade (2013) matched occupied housing unit records in the 2009 ACS to CoreLogic to examine the relationship between the self-reported value obtained from the ACS householder and measures of value derived from administrative records. The author found a match rate of 80 percent for single-family owned homes.

CoreLogic's coverage for Virginia ranges from 99 to 112 counties and independent cities (out of 133 counties and independent cities in the state) from 2009 and 2013. The counties missing from this range are in rural areas of Virginia. CoreLogic standardizes the data across jurisdictions, even though assessment regulations and processes vary across jurisdiction lines. For instance, whereas each jurisdiction has its own land use codes, CoreLogic creates a standardized land use code and applies it across all jurisdictions. Unfortunately, the standardization process can mean that some level of detail is lost, as the process involves assumptions that allow for the data to fit CoreLogic's rule set. For instance, reports of zero bedrooms are assumed to be missing data, which implies that efficiency units are not properly identified. Such loss of detail should be taken into account when the CoreLogic data are used.

Provenance associated with commercial administrative data is frequently missing. CoreLogic's postprocessing assumptions are unknown, as the details about the algorithms are proprietary. For instance, CL assessment data contain information about the number of units within the buildings for individual condominiums and absent-owner properties, whereas the AC assessment data on which the CL assessment data are based do not. In another example, the CL assessment data for Arlington included condominiums with listed values as high as $1 billion. In these cases, we used information obtained from the AC assessment data to recategorize some CoreLogic parcels as multifamily parcels.

In this article, the authors will compare the estimates of housing characteristics based on the AC and CL assessment data sources to those in the 2009–2013 ACS, as obtained from American FactFinder (U.S. Census Bureau, 2014b).

Defining a Fitness-for-Use Metric

One main goal of this research is to understand the feasibility of repurposing administrative data to enhance and supplement the ACS for housing characteristics. Keller et al. (2016) provided more comprehensive information on the development of the data framework we used; it encapsulated a general approach for repurposing data, from data source discovery to analysis to inference. Within that framework, and more generally, *fitness for use* is a term that is used to define how well the data meet the needs of the user. Assessing fitness for use requires considering the modeling and analyses in which the data will be used. A fitness-for-use metric needs to be a function of the data coverage (representativeness) needs and the data quality needs of the model(s) used in the analyses (Dippo, 1997; Keller et al., 2016).

We used the following fitness-for-use metric to compare the tabulations created from the AC and CL real estate assessment data to the ACS tabulations—

$$fitness\ ratio = \frac{ACS\ estimate - external\ estimate}{90\%\ ACS\ margin\ of\ error}. \qquad (1)$$

The fitness ratio quantifies the degree of alignment between the two estimates. A negative fitness ratio means that the external estimate is higher than the ACS estimate, and a positive fitness ratio means that the external estimate is lower. A fitness ratio ≤ |1| means that the two estimates align within the ACS estimate's 90-percent margin of error, defined as the 90-percent confidence interval for ACS estimates.

The challenge in interpreting the comparisons comes from a lack of knowledge about which estimate might be right, because no obvious gold standard exists. As will be discussed, it is likely that the real estate assessment data may be more accurate than the ACS in some cases, because local governments have financial incentives to have accurate information on characteristics such as the housing value for property tax assessments and real estate taxes paid. In the case of year built, the real estate assessment data may better reflect reality because the local data do not rely on self-reporting as a survey does. For example, renters are much less likely than owner-occupiers to know a unit's year of construction, and the county government knows the year a new or renovated home is updated in their database through its building permit system. Finally, the real estate assessment data may also generate more accurate estimates than a survey because it is essentially a census of housing units, whereas the latter has sampling and nonsampling errors.

In contrast, ACS data might be more accurate than assessment data in certain circumstances and are the only source of certain housing characteristics, such as the components of rent and monthly owner costs (for example, utility costs). Another way ACS data may be more accurate is in determining the number of housing units in a building. A large complex of independent buildings on one parcel is likely to have a much larger unit count in the assessment data (perhaps in the hundreds) than the count reported by the respondent living in a building on that parcel with (say) eight apartments. The latter is a more accurate description of the respondent's living conditions.

Data Preparation

The data preparation process began with identifying the potential differences in the data sources among the AC assessment data, the CL assessment data, and the ACS data. This step is important, as real estate assessment data are collected for the purpose of property tax assessments, which is a distinctly different purpose than for ACS data collection. Appendix A includes a detailed data quality checklist for property data, which was developed through this research. The data preparation process lays the foundation for assessing if the real estate assessment data can be used in research analogous to how the ACS data are used.

The AC assessment data contain the collection of all parcels in the county. A *parcel* is a defined piece of real estate (for example, lot) that is identified for taxation purposes in the jurisdiction (State of Virginia, 1996). Parcels may or may not contain housing units. For example, some parcels may represent vacant lots or parking spaces. Some parcels contain multiple housing units, as in apartment complexes. The CL assessment data, as provided by CoreLogic, are a subset of all parcels in the jurisdiction and are supposed to be restricted to nonvacant, non-parking-lot residential parcels. In contrast, the ACS data for Arlington contain a sample selected from all residential housing units, not parcels, in the county. A *housing unit* is a house, apartment, mobile home, group of rooms, or a single room that is intended for occupation as separate living quarters (U.S. Census Bureau, 2014a).

To compare the data, nonresidential parcels and vacant parcels were identified in the AC assessment data and the CL assessment data.[2] These parcels were removed if found, as they are not within the scope of the ACS, leaving roughly 60,000 residential parcels (see exhibit 1). For the AC assessment data, we identified parking lots and vacant parcels by pinpointing parcels with either

Exhibit 1

Estimate of Housing Units by Data Source, Arlington County, 2009–2013

Data Source	2009	2010	2011	2012	2013	2009–2013
ACS	103,813	105,490	106,720	107,734	109,689	106,740
MOE	± 872	± 619	± 417	± 537	± 504	± 191
AC weighted	100,991	101,867	102,299	102,299	103,987	102,33
AC parcels	60,261	60,203	60,465	60,688	60,966	NA
CL weighted	97,589	98,690	83,640	79,521	79,804	87,849
CL parcels	59,593	59,959	60,077	60,220	60,343	NA

AC = Arlington County. ACS = American Community Survey. CL = CoreLogic, Inc. MOE = the 90-percent ACS margin of error. NA = not applicable.

Sources: 2009–2013 ACS 5-year estimates; AC real estate assessments, 2009–2013; CL real estate assessments, 2009–2013

[2] All editing of the assessment data is by the authors.

no total value or no improvement (building) value but with a land value.[3] Common areas and additional vacant parcels are identified using a combination of land use codes. Land use codes are classification schemes created and managed by local jurisdictions that describe the class of property permitted on that parcel (State of Virginia, 1996). A small number of parcels that are partially in Arlington and partially in another jurisdiction were removed from the AC assessment data, because the assessed value pertains solely to the Arlington portion, and the ACS response will apply to the whole property. Hotels were removed from the CL assessment data using Arlington's land use code as reported in the CL assessment data. Moreover, we identified and removed a parking spot and a common area in the CL assessment data.

Another key difference between the real estate assessment data and ACS data is that whereas the real estate assessment data pertain to each *parcel* in the jurisdiction, the ACS collects information for each *housing unit* in its sample. For example, multifamily buildings are typically one parcel, but the ACS treats each apartment within the multifamily building as a separate unit. The real estate assessment data indicate the number of units on each parcel, allowing for the data to be reweighted to create comparable statistics to ACS tabulations. In the AC assessment data, some parcels had varying unit counts across the years. The most notable one was a building identified as having 842 units in 2012 and 266,412 units in 2013. As this high (and obviously incorrect) unit count would bias estimates after weighting, we replaced the 2013 counts for the AC assessment data with 2012 numbers in such cases because the 2012 number more closely matched the 2009-through-2011 data.

The AC assessment data also included data for multiple-dwelling parcels, which could have included, for example, a parcel with a primary housing unit and a registered additional housing unit, such as a basement. In these cases, the data have to be restructured so that they include one observation per dwelling. In contrast, the CL assessment data do not require unit count editing and the data provided are for the primary dwelling on the parcel. However, CoreLogic does not conduct longitudinal editing on its variables (including multifamily unit counts) and as a result errors go undiscovered.[4] Longitudinal editing would not be difficult to do, as multiple years of data are accessible. We chose not to do so, primarily to provide a contrast between what a local jurisdiction could do with its own data and what a federal agency would find difficult to do one by one for the thousands of jurisdictions in the country.

The AC assessment data do not have direct information on the number of units in condominium buildings. For tax purposes, Arlington creates a different parcel number (ID) for each condominium. It is assumed by the authors that missing unit counts are single-family housing units. This assumption results in condominiums being classified as single-family attached residences, and some adjustment is needed to correct this misclassification for condominiums. The parcel IDs associated with condominiums in the same structure have the same GIS code. The number of condominium units in a structure was imputed by aggregating the number of condominiums based on their GIS codes. Lastly, duplexes do not have unit counts; however, a specific duplex land use code allowed for the appropriate categorization of these properties.

[3] Removing parcels with no land value does not eliminate condominiums, as Virginia law states that condominiums must have a land value.

[4] For example, the multifamily complex listed by Arlington County as having 266,412 units in 2013 had its unit count reset by CoreLogic through truncation to 6,412.

Within the AC assessment data are two variables in relation to the age of the building: property-year-built and dwelling-year-built. In the property-year-built data, all properties are included and the unit of observation is the unique parcel number. In the dwelling-year-built data, only single-family properties are included and the unit of observation is the dwelling. Thus, a parcel with two dwellings could have two different dwelling-year-built values if the second dwelling is built later than the original dwelling. To create a year built variable from the assessment data that has only one value for each dwelling and multifamily property, we added the property-year-built data for multifamily properties to the dwelling-year-built data. The CL assessment data have only one year-built value, as the unit of observation is the parcel and not the dwelling. The year built value pertains to the primary dwelling on the parcel.

To place the housing units into census geographic boundaries (tracts and block groups), we needed geographic coordinates (latitudes and longitudes). See appendix B for a description of how we accomplished this step.

A Micro-Level Comparison of Arlington County and CoreLogic Assessment Data

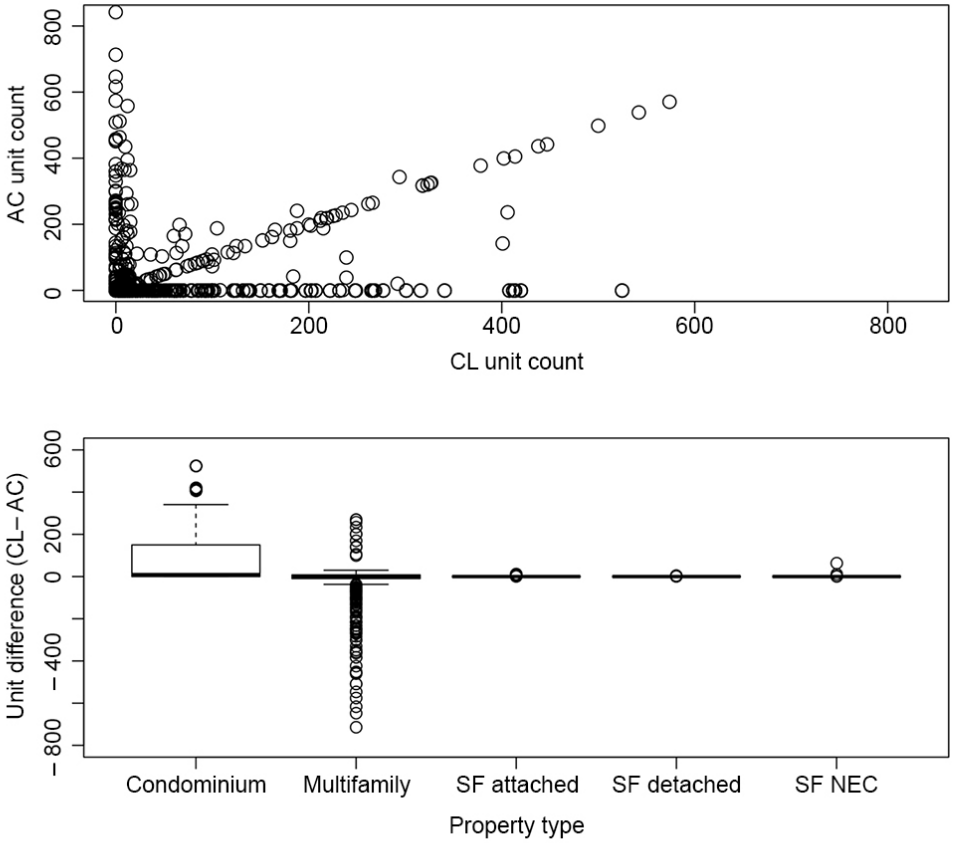
Once the preliminary data preparation steps were accomplished, we compared the individual elements of the AC assessment data and CL assessment data, as matched by parcel ID. As exhibit 1 indicates, the number of parcels in each database is different for each year from 2009 to 2013 (we will discuss the number of housing units in the next section), though the difference is no more than 1 percent (always higher in the AC assessment data).

We found a high match rate between the two data sources for 2013; 60,286 parcels appeared in both databases, 57 appeared solely in CL assessment data, and 680 appeared solely in AC assessment data. Of those that were in only the CL assessment data, 6 were condominiums, 46 were single-family attached properties, and 5 were single-family not elsewhere classified properties. For the 57 parcels in only the CL assessment data, the mean year built was 1946, and 17 parcels did not have a year built listed. Of those parcels that were in only the AC assessment data, 400 were condominiums, 204 were multifamily properties, 13 were single-family attached properties, and 63 were single-family detached properties. For the 680 parcels in only the AC assessment data, the mean year built was 1980, and 372 parcels did not have a year built listed. These data were matched by parcel. Thus, parcels with multiple dwellings were matched only to the primary dwelling.

For the 60,286 parcels that matched, we compared the number of units in both data sets, as these data were integral in this study to reweight the administrative data to the housing unit. When we compared the unit counts for each parcel individually and by property type, shown in exhibit 2, we observed that the difference in unit counts between the two data sources were primarily due to condominium and multifamily properties. For this comparison, we removed three observations that had more than 1,000 units in the AC assessment data. In the CL assessment data, these units are listed as having zero units.

Exhibit 2

Number-of-Unit Comparison for CoreLogic, Inc., and Arlington County Real Estate Assessment Data, 2013



AC = Arlington County. CL = CoreLogic, Inc. NEC = not elsewhere classified. SF = single-family.
Notes: Unit is the parcel. Three outliers with differences of greater than 1,000 were removed.
Sources: AC real estate assessments, 2013; CL real estate assessments, 2013

Comparisons of Housing Characteristics

In this section, we provide the results of comparisons between the estimates based on the real estate assessment data and the on six housing characteristics: number of housing units, type of housing unit, year built, number of bedrooms, housing value, and real estate taxes paid.

Number of Housing Units

Exhibit 1, in the previous section, presents comparisons of the number of housing units in Arlington for the 2009-to-2013 period. Housing unit counts from the AC assessment data are below the 90-percent margin of error for the ACS estimates. Real estate assessment data theoretically contain a census of housing units in the jurisdiction, yet reconciling differences of the unit of observation between the administrative data and ACS was not straightforward. As noted, the number of housing units for the assessment data is obtained by weighting the parcels by the number of units in each parcel. This process can be problematic due to missing and varying unit counts. The number of missing unit counts ranges from 206 parcels in 2009 to 463 in 2013 in the AC assessment data. Associating a large fraction of these missing data with multifamily dwellings could result in underestimation of the total number of housing units. In the CL assessment data, the number of units in a building for multifamily properties is complete for 2009–2013, yet the number of units varies across the years (with a range between 1 to 826 units per parcel from year to year) for about 30 percent of the properties. This variance is especially apparent when comparing the earlier years to the later years. Exhibit 1 shows that, from 2009 to 2013, the number of housing units was lower in the CL assessment data than in the ACS and AC assessment in every year, but it was substantially lower for 2011 and later. The latter is due to the unexplained difference among number of units data within the original data across the years.

Overall, the lower estimate of the number of housing units seems to be driven by discrepancies in unit counts of the multifamily structures, but the cause of the difference between housing unit counts in CoreLogic and in the real estate assessment data remains to be determined. Due to this significant difference, creating 5-year estimates using CL assessment data for housing characteristics that need to be weighted by number of units is problematic.

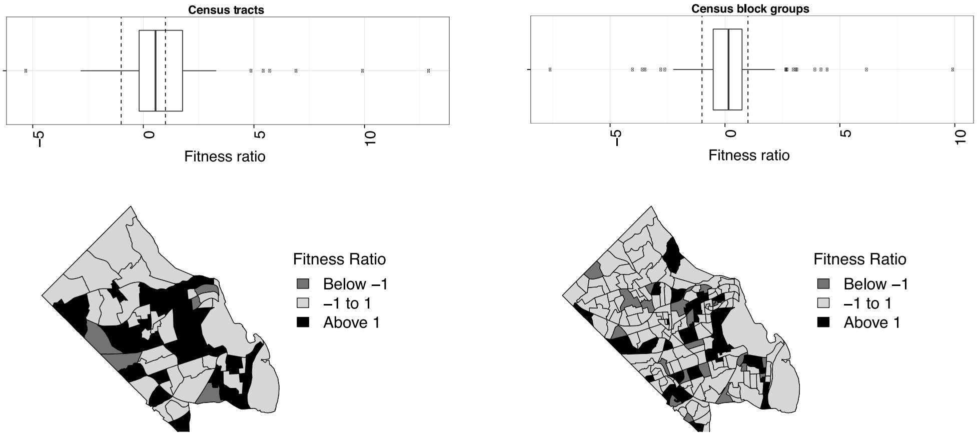
The differences in housing unit counts are also examined across census tracts and block groups. There are 59 census tracts containing 181 block groups in Arlington. Exhibit 3 provides boxplots and geographic distributions of the fitness ratios across census tracts and block groups, using the AC assessment data.[5] Although the estimates do not agree at the county level, the estimates obtained from the AC assessment data often align (|fitness ratio| ≤ 1) with ACS estimates at lower levels of geography.

We conjecture that estimates based on real estate assessment data that do not align with ACS estimates are aligned with housing unit density within Arlington. AC assessment data appear to generate lower estimates for multifamily buildings due to either different or missing unit counts for multifamily buildings. We addressed this discrepancy by exploring whether the unit counts that are significantly lower compared to the ACS estimate, (fitness ratios ≤ 1), are observed in areas with high multifamily property density at the census tract level as measured by ACS. Exhibit 4 visually compares ACS estimates of the number of housing units in multifamily properties to the geographic distributions of number of housing units' fitness ratios by census tract. The largest differences appear to be in geographic areas with high housing unit density; however, this finding does not explain all the divergence. The actual composition of the housing units by type is explored further and discussed next.

[5] Boxplots are graphical displays of several descriptive statistics. The bottom of the box, middle line, and top of the box are the 25th, 50th (median), and 75th percentiles, respectively. The lines coming from the bottom and top of the box are called *whiskers* and stop at the smallest and largest data value within 1.5 times the interquartile range (IQR), wherein IQR is the distance between the 75th and 25th percentiles and is equivalent to the width of the box. Data values outside the box are called outliers and are plotted at their actual values.

Exhibit 3

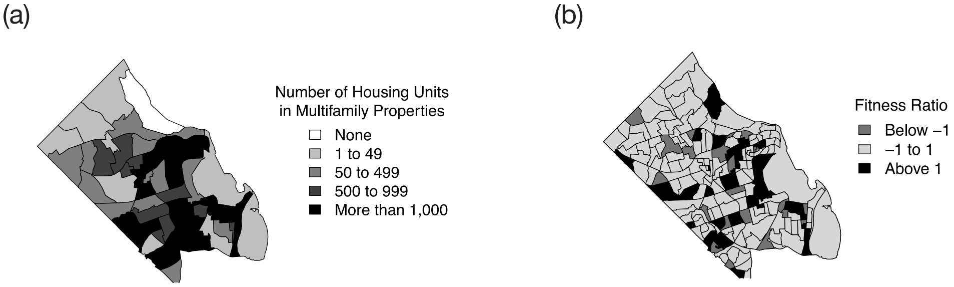
Fitness Ratios for Number of Housing Units by Census Tract and Block Group, Arlington County, 2009–2013



Notes: Boxplots compare the distribution of the fitness ratios at the census tract and block group levels, their relation to the 90-percent American Community Survey (ACS) margins of error, and geographic distributions. Estimates falling outside the dashed reference lines or lighter gray color of 1 were not within the ± 90-percent ACS margins of error.
Sources: Arlington County real estate assessments 2009–2013; 2009–2013 ACS 5-year estimates

Exhibit 4

Comparison of the Fitness Ratios for Housing Units to Density of Multifamily Properties by Census Tract, Arlington County, 2009–2013



Note: Panel (a) is the geographic distribution of the number of multifamily properties and panel (b) is the geographic distribution of housing unit count fitness ratios.
Sources: panels (a) and (b)—2009–2013 ACS 5-year estimates; panel (b)—Arlington County real estate assessment data, 2009–2013

Type of Housing Unit

The second set of comparisons involves the distributions associated with the units in a structure. The ACS poses the following question to survey respondents: “Which best describes this building?” The list of possible responses corresponds with the categories given in exhibit 5. For the AC assessment data, we used the numbers of units in the building and county land use codes to place housing units into the appropriate category. Although the unit counts differ across the years for the AC assessment data, the ACS tabulates group number of units into bins, which absorbs some of the difference.

Exhibit 5

Distribution of the Number of Housing Units in Structures: Comparison of ACS Estimates With Arlington County Data, 2009–2013

	ACS		Direct Estimate From Arlington County		
Units in Structure	Estimate (%)	90% MOE (%)	Estimate (%)	Difference (%)	Within 90% MOE?
1—detached	26.91	5.47	26.73	0.17	Yes
1—attached	9.42	5.17	5.78	3.64	Yes
2	0.97	2.53	0.53	0.44	Yes
3 or 4	3.68	3.09	0.07	3.61	No
5 to 9	5.46	4.24	0.93	4.54	No
10 to 19	7.72	5.62	1.83	5.89	No
20 to 49	5.19	4.59	5.08	0.10	Yes
50 or more	40.27	7.31	58.78	– 18.51	No

ACS = American Community Survey. MOE = margin of error.

Sources: Arlington County real estate assessment data, 2009–2013; 2009–2013 ACS 5-year estimates, Table B25024

Exhibit 5 compares the distribution of units in structure from the 5-year estimates based on ACS to the AC assessment data. The estimates obtained from AC assessment data that fall outside the 90-percent ACS margins of error correspond to all multifamily buildings of three or more units, except for 20- to 49-unit buildings. The alignment of the estimates for single-family attached housing units is likely due to the restructuring of the AC assessment data that resulted in each single-family detached unit being on its own parcel and appearing an unweighted observation in the data. Neither source is unambiguously better for these estimates.

Year Built

In relation to the age of the property, the ACS asks respondents: “About when was this building first built?” Exhibit 6 reflects the categories of year-groupings among which respondents choose. However, many residents, especially those living in apartments, may not know the exact or even the approximate answer to this question. Information on year built within local housing administrative data is based on official assessments and permitting data, and refers to the first

Exhibit 6

Distribution by Year Built of Housing Units: Comparison of ACS Estimates With Arlington County Data, 2009–2013

	ACS		Direct Estimate From Arlington County		
Year Built	Estimate (%)	90% MOE (%)	Estimate (%)	Difference (%)	Within 90% MOE?
2000 to 2009	15.83	0.71	14.49	1.34	No
1990 to 1999	9.15	0.68	8.73	0.43	Yes
1980 to 1989	11.26	0.60	11.16	0.10	Yes
1970 to 1979	10.67	0.72	6.48	4.19	No
1960 to 1969	11.64	0.73	14.91	– 3.27	No
1950 to 1959	15.71	0.78	15.62	0.09	Yes
1940 to 1949	15.56	0.68	16.59	– 1.03	No
1939 or earlier	9.00	0.47	6.05	2.94	No

ACS = American Community Survey. MOE = margin of error.

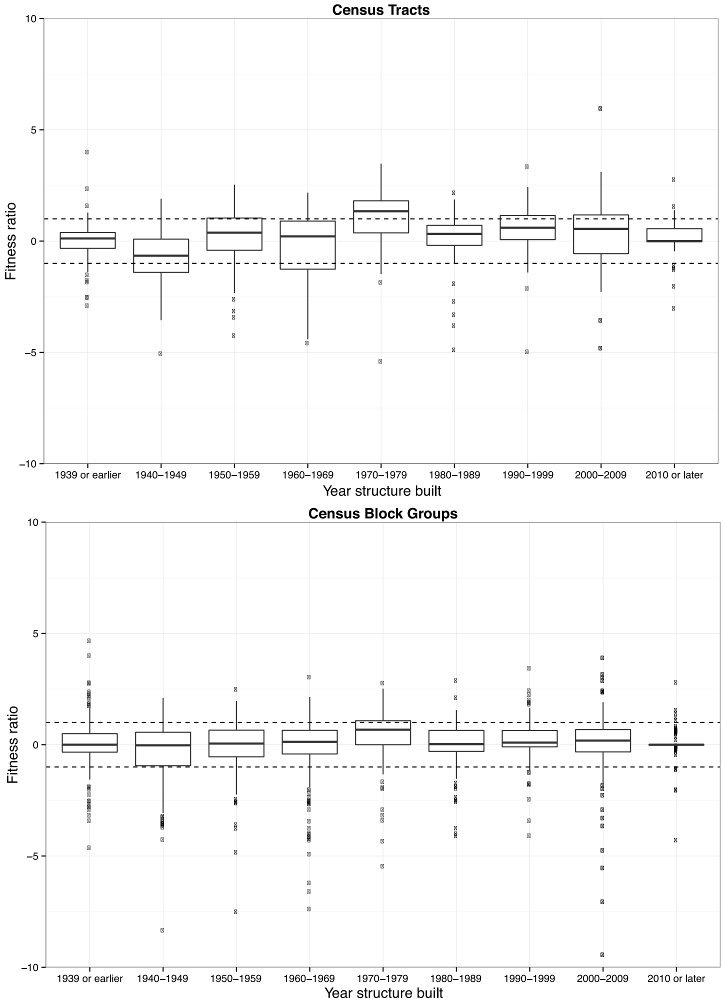
Sources: Arlington County real estate assessment data, 2009–2013; 2009–2013 ACS 5-year estimates, Table B25034

year a property has an improvement (usually a building) in the assessment records. The evidence provided in this section suggests that "year built" may be more accurately captured using real estate assessment data sources than by the ACS.

Exhibit 6 compares the 5-year estimates of year built based on the ACS to the estimates obtained from AC assessment data. At the county level, AC assessment data-based estimates fall outside the 90-percent ACS margins of error for both old and new structures. Exhibit 7 presents the fitness

Exhibit 7

Fitness Ratios for Number of Housing Units by Year Built Across Census Tracts and Census Block Groups for Arlington County, 2009–2013



Notes: Boxplots compare the distribution of the fitness ratios at the census tract and block group level. Estimates falling outside the dashed reference lines of 1 were not within the ±90-percent American Community Survey (ACS) margins of error. For presentation purposes, one extreme lower outlier (18.67 for 2000–2009) was removed from the census tracts boxplots and three extreme lower outliers (-35.79 for 1990–1999, 18.67 for 2000–2009, and 17.21 for 2000–2009) were removed from the real estate taxes boxplot.

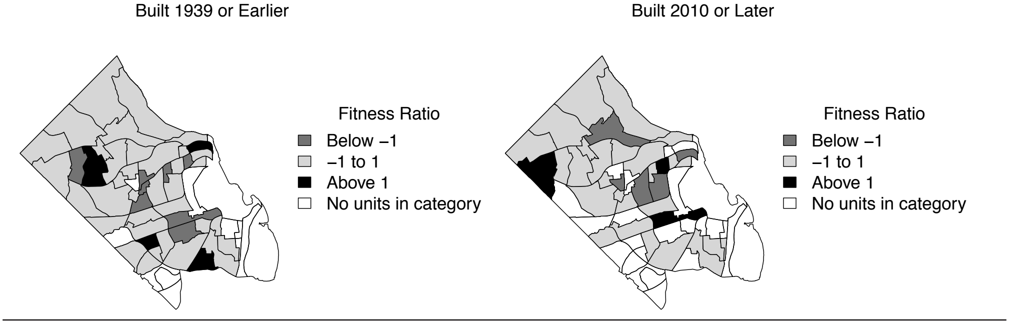
Sources: Arlington County real estate assessments, 2009–2013; 2009–2013 ACS 5-year estimates ratio distributions across census tracts and block groups for the comparison of the AC assessment data to ACS's 5-year estimates of housing units by year built. Most of the fitness ratios fall within the ±1 range across the years in both cases. Estimates using CL assessment data are not presented due to problems associated with weighting across all 5 years.

Exhibit 8 illustrates the geographic distribution of the fitness ratios by census tract, comparing the oldest (pre-1939) and the youngest (post-2010) structures in Arlington. The ACS estimates a lower count of both younger and older housing units in the center of the county when compared to the AC assessment data. Central Arlington is an area with high growth of both residential and commercial properties, with many renter-occupied housing units, and that follows the Metro (subway line). ACS data show a lower volume of young structures in central and southern Arlington, a region of lower income as compared to the AC assessment data. Those census tracts have a high volume of renter-occupied housing units according to the ACS. This comparison may be another indication of respondent misreporting.

When we compare the median year built of housing units in Arlington across the data sources, we observe differences depending on the timeframe. The 5-year estimate for median year built was 1961 based on AC assessment data, whereas the median year built in the ACS was 1968 (±2). The AC assessment data estimate of the median year built remained consistent at 1961 for the years between 2010 and 2013 and 1960 for 2009. However, the ACS single-year estimates steadily increased from 1965 (±2) in 2009 to 1973 (±2) in 2013. This increase is too high to be credible in a county of this size and provides additional evidence that the year built data from the AC assessment data may be more accurate than those data from the ACS.

Exhibit 8

Fitness Ratios for Number of Oldest and Youngest Housing Units by Census Tract, Arlington County, 2009–2013



Sources: Arlington County real estate assessments 2009–2013; 2009–2013 American Community Survey 5-year estimates

Number of Bedrooms

In the ACS, the question regarding bedroom counts is posed as, "How many of these rooms are bedrooms? Count as bedrooms those rooms you would list if this house, apartment, or mobile home were for sale or rent. If this is an efficiency/studio apartment, print '0.'" A set of rules within the Virginia State Building Code determines whether a room is classified as a bedroom in the real estate assessment data (State of Virginia, 1996). An occupant responding to ACS may not be reporting a bedroom consistent with Virginia state codes. For example, in Virginia, a bedroom must have at least one operable emergency escape and rescue opening, such as a window. However, a respondent may count a den that is being used as a bedroom as such even though it may not have an escape route. In the AC assessment data, this den would not appear as a bedroom and thus the datum would show one fewer bedroom. The distribution of the housing units by the number of bedrooms based on the ACS data compared to the AC assessment data is presented in exhibit 9.

The AC assessment data estimate for number of bedroom units did not align with the ACS estimate for housing units with lower numbers of bedrooms. This discrepancy could be due to the fact that the AC assessment data do not include information on the characteristics of individual apartments, which are often zero-, one-, and two-bedroom units. In addition, the definition of a bedroom based on the state code may exclude bedrooms that are counted in the ACS data. As noted previously, the CL assessment data have no zero-bedroom units by design of the standardization, but did have 6,938 units with missing bedroom data in 2013; Arlington had 7,955 units with zero bedrooms that same year. This incompleteness of data within real estate assessment data leads to misalignment with ACS estimates.

Exhibit 10 illustrates the fitness ratios for zero-bedroom housing units by census tract. This exhibit reveals that most of the discrepancies between the ACS and the real estate assessment data correspond to North Arlington, an area that consists primarily of high-valued single-family detached homes, and are also observed in some tracts in South Arlington, in a residential neighborhood near the Pentagon.

The boxplots in exhibit 10 illustrate the distribution of the values of housing units by the number of bedrooms they have from the AC assessment data. The distribution of values for zero-bedroom units compared to the other units is wider than expected, which explains the discrepancies in the estimates of one-bedroom housing units in exhibit 9. This distribution implies that some of the zero-bedroom units in the AC assessment data correspond to one- and two-bedroom units in the ACS. As expected, the estimates for one- and two-bedroom housing units in exhibit 9 using the AC assessment data are lower than the corresponding ACS estimates. This finding is not surprising because the number of housing units in multifamily buildings are underestimated due to missing units counts, as mentioned previously. The counts for housing units with four and with five or more bedrooms are not affected, as housing units of these sizes tend to be single-family units, which the AC assessment data capture well.

Exhibit 9

Distribution of Number of Bedrooms: Comparison of ACS Estimates With Arlington County Data, 2009–2013

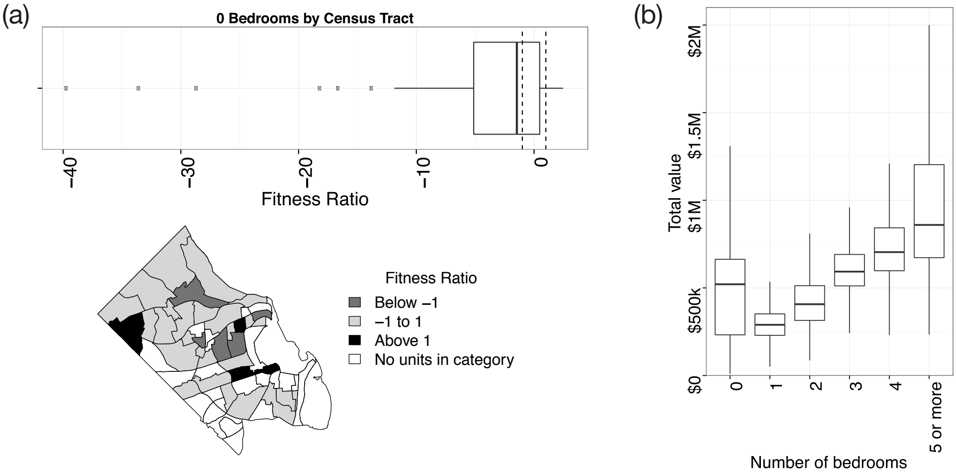
	ACS		Direct Estimate From Arlington County		
Number of Bedrooms	Estimate (%)	90% MOE (%)	Estimate (%)	Difference (%)	Within 90% MOE?
0	4.29	0.49	13.17	– 8.88	No
1	33.79	0.91	18.14	15.64	No
2	29.64	1.00	28.52	1.12	No
3	18.14	0.73	26.03	– 7.89	No
4	9.54	0.51	9.62	– 0.08	Yes
5 or more	4.61	0.42	4.48	0.13	Yes

ACS = American Community Survey. MOE = margin of error.

Sources: Arlington County real estate assessment data, 2009–2013; 2009–2013 ACS 5-year estimates, Table B25041

Exhibit 10

Distributions of Housing Value by Number of Bedrooms, Arlington County, 2009–2013



Notes: Panels provide the fitness ratio distributions of number of housing units by house values and number of bedrooms. Panel (a) is the boxplot and geographic distribution of the fitness ratios for "0-bedroom homes" by census tract from the Arlington County (AC) assessment data. Estimates falling outside the dashed reference lines of 1 were not within the ±90-percent American Community Survey (ACS) margins of error. Panel (b) is a collection of boxplots displaying the housing "total value" ($) distributions by number of bedrooms in the unit found within the AC assessment data.
Sources: panels (a) and (b)—AC real estate assessments 2009–2013; panel (a)—2009–2013 ACS 5-year estimates

Housing Value and Real Estate Taxes Paid

ACS respondents who own their housing unit are asked, "About how much do you think this house and lot, apartment, or mobile home (and lot, if owned) would sell for if it were for sale?" and "What are the annual real estate taxes on this property?" These questions could be difficult to answer, especially the former, if the owners are not recent homebuyers. There is a debate in the literature as to whether self-reported housing values are accurate. Early work did not show bias in self-reported value (Kain and Quigley, 1979), however subsequent research has found that homeowners overstate the values of their homes on the order of 5 to 16 percent (Benítez-Silva et al., 2010; DiPasquale and Somerville, 1995; Goodman and Ittner, 1992; Ihlanfeldt and Martinez-Vazquez, 1986; Kiel and Zabel, 1999). Housing value in real estate assessment data is based on annual assessments conducted by the jurisdiction, whether in person or through analytical valuation approaches.

One difficulty in this study is the potential difference in the populations between the ACS and AC assessment data. The ACS estimates housing value only for owner-occupied units. The AC assessment data estimates are based on single-family attached and detached housing units, including condominiums. AC assessment data do not have an indicator to determine if a housing unit is owner or renter occupied. Multifamily properties are excluded, as they are solely renter-occupied units. Given the lack of an indicator for owner-occupancy, the housing unit counts in the AC assessment data are expected to be larger than in the ACS data because the AC assessment data include both owner- and renter-occupied units. The magnitude of this difference is proportional to the number of single-family detached, single-family attached, and condominium units that are renter-occupied in that geographic area.

Exhibit 11 shows the comparison of the distribution of housing units by value of the AC assessment data with the ACS estimates. For all but a small fraction (2 percent) of units, the ACS and AC assessment estimates do not match; the mismatches are particularly for all units in categories of $90,000 or more in value, but not all the mismatches are in the same direction. For example, the ACS estimates that 33 percent of owner-occupied units in Arlington have a value between $500,000 and $749,999 compared with the AC assessment estimate of 35 percent. For housing valued at $750,000 to $999,999, the estimates are reversed; the ACS estimates 18 percent of owner-occupied units are in this category, and the AC assessment estimates only 9 percent.

Exhibit 12 shows the comparison of the distributions of real estate taxes paid. The match between the estimates of taxes from the ACS and the AC assessment data is much better than the match between estimates for housing value—a match for four out of six categories, including the one category with the most units (taxes paid of $3,000 or more).

Exhibit 11

Distribution of Housing Units by Value: Comparison of ACS Estimates With Arlington County Data, 2009–2013

	ACS		Direct Estimate From Arlington County		
Value of Housing Unit	**Estimate (%)**	**90% MOE (%)**	**Estimate (%)**	**Difference (%)**	**Within 90% MOE?**
Less than $10,000	0.25	0.12	0.00	0.25	No
$10,000 to $14,999	0.05	0.05	0.00	0.04	Yes
$15,000 to $19,999	0.00	0.07	0.00	0.00	Yes
$20,000 to $24,999	0.16	0.13	0.00	0.16	No
$25,000 to $29,999	0.11	0.11	0.00	0.11	Yes
$30,000 to $34,999	0.11	0.11	0.00	0.11	No
$35,000 to $39,999	0.07	0.08	0.00	0.07	Yes
$40,000 to $49,999	0.16	0.13	0.00	0.16	No
$50,000 to $59,999	0.14	0.08	0.01	0.13	No
$60,000 to $69,999	0.26	0.17	0.08	0.18	No
$70,000 to $79,999	0.19	0.15	0.11	0.08	Yes
$80,000 to $89,999	0.12	0.13	0.23	– 0.10	Yes
$90,000 to $99,999	0.07	0.06	0.28	– 0.21	No
$100,000 to $124,999	0.66	0.25	1.65	– 0.98	No
$125,000 to $149,999	0.65	0.28	1.52	– 0.88	No
$150,000 to $174,999	0.80	0.25	1.35	– 0.55	No
$175,000 to $199,999	1.11	0.33	1.51	– 0.40	No
$200,000 to $249,999	3.33	0.58	4.94	– 1.61	No
$250,000 to $299,999	5.54	0.91	6.97	– 1.43	No
$300,000 to $399,999	13.30	1.17	16.67	– 3.38	No
$400,000 to $499,999	11.88	1.04	14.05	– 2.17	No
$500,000 to $749,999	32.64	1.69	35.41	– 2.78	No
$750,000 to $999,999	18.45	1.25	8.74	9.70	No
$1,000,000 or more	9.96	0.92	5.20	4.76	No

ACS = American Community Survey. MOE = margin of error.
Sources: Arlington County real estate assessment data, 2009–2013; 2009–2013 ACS 5-year estimates, Table B25075

Exhibit 12

Distribution of Real Estate Taxes Paid: Comparison of ACS Estimates With Arlington County Data, 2009–2013

Real Estate Taxes Paid	ACS		Direct Estimate From Arlington County		
	Estimate (%)	90% MOE (%)	Estimate (%)	Difference (%)	Within 90% MOE?
No real estate taxes paid	381	363	655	– 274	Yes
Less than $800	1,076	775	53	1,023	No
$800 to $1,499	1,145	793	1,790	– 645	Yes
$1,500 to $1,999	1,912	1,113	1,777	135	Yes
$2,000 to $2,999	4,683	1,463	6,444	– 1,761	Yes
$3,000 or more	32,113	3,986	33,088	– 975	Yes

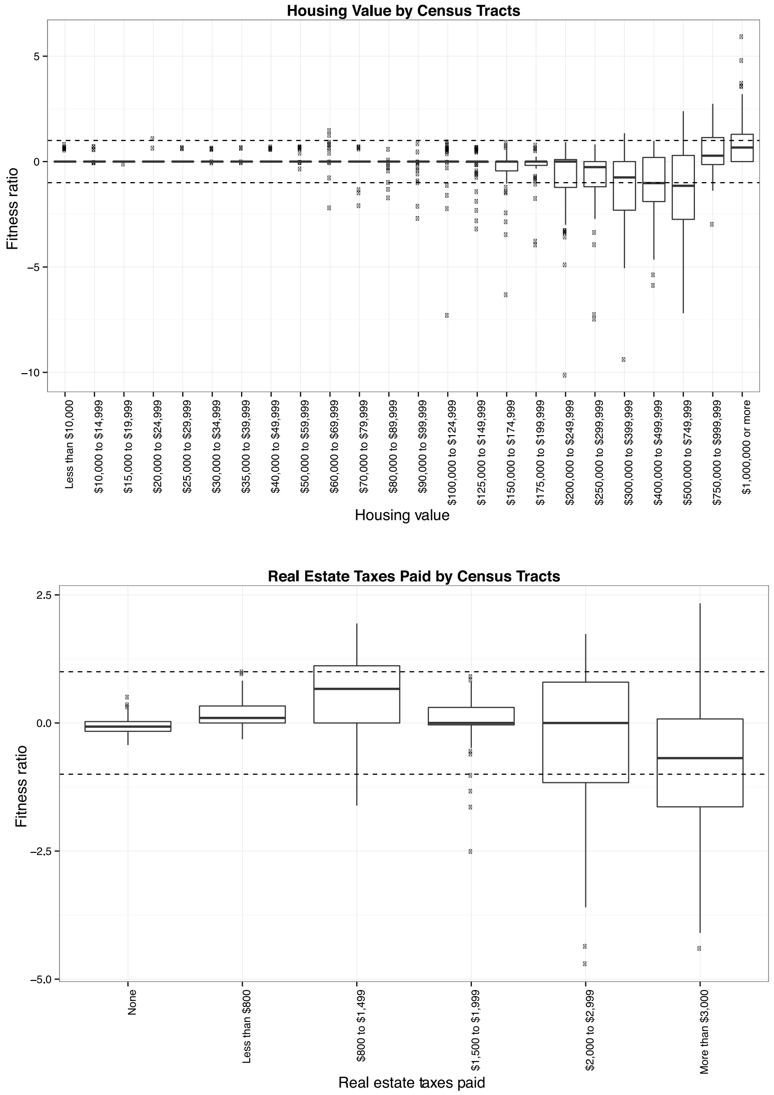
ACS = American Community Survey. MOE = margin of error.
Sources: Arlington County real estate assessment data, 2009–2013; 2009–2013 ACS 5-year estimates, Table B25102

The census tract fitness ratio distributions in exhibit 13 that use AC assessment data show larger counts of higher-priced housing units in the $200,000-to-$749,999 range and housing units that paid more than $2,000 in real estate taxes. Exhibit 14 presents the geographic detail for this pattern based on the fitness ratios for housing value between $500,000 and $749,999. The tracts where the AC assessment data estimates more units in this range (fitness ratios < -1) are mostly in North Arlington, an area with a high number of larger single-family homes. This misalignment could be the result of a high volume of renters in these homes, leading to ACS respondents who underestimate the value of their rental, though this result is unlikely. The two highest housing value categories did not follow these same trends and were also in price ranges where rentals may be less likely.

Similar to the AC assessment data, CL assessment data aligned with the ACS in areas where weighting does not affect final estimates: housing value and real estate taxes paid. Exhibit 15 shows the fitness ratios for these comparisons across census tracts. AC assessment data included an indicator on whether the property has an absent owner, which CoreLogic imputed using a proprietary algorithm. Such units were excluded in these calculations. The benefits of this absent-owner indicator are seen when comparing the housing value boxplots from exhibits 13 to 15. That is, a greater proportion of CL assessment data estimates fell within the range of fitness ratios < -1. This benefit was not observed in regards to the real estate taxes paid.

Exhibit 13

Fitness Ratios for Number of Housing Units for Housing Value and Real Estate Taxes Paid Across Census Tracts for Arlington County, 2009–2013

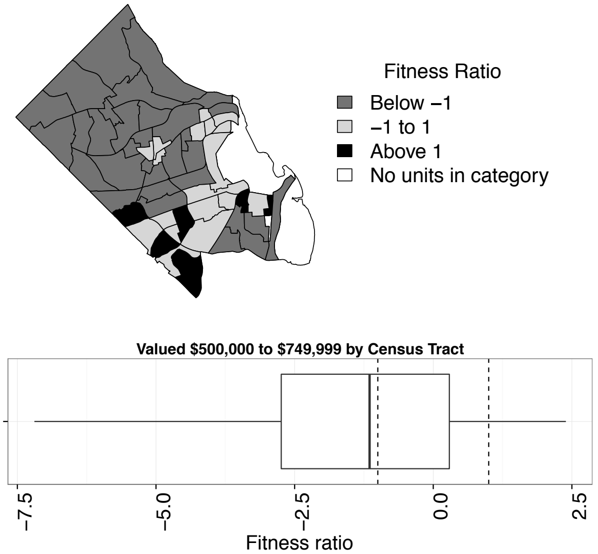


Notes: Boxplots display the distributions of the fitness ratios at the census tract level. Estimates falling outside the dashed reference lines of 1 were not within the ±90-percent American Community Survey (ACS) margins of error. For presentation purposes, three extreme lower outliers (-15.35 for $200,000 to $249,999, -32.19 for $100,000 to $124,999, and -35.12 for $150,000 to $174,999) were removed from the value boxplots and one extreme lower outlier (-14.98 for $800 to $1,499) was removed from the real estate taxes boxplot.

Sources: Arlington County real estate assessments 2009–2013; 2009–2013 ACS 5-year estimates

Exhibit 14

Fitness Ratios for Number of Housing Units Valued at $500,000 to $749,999 by Census Tracts for Arlington County, 2009–2013

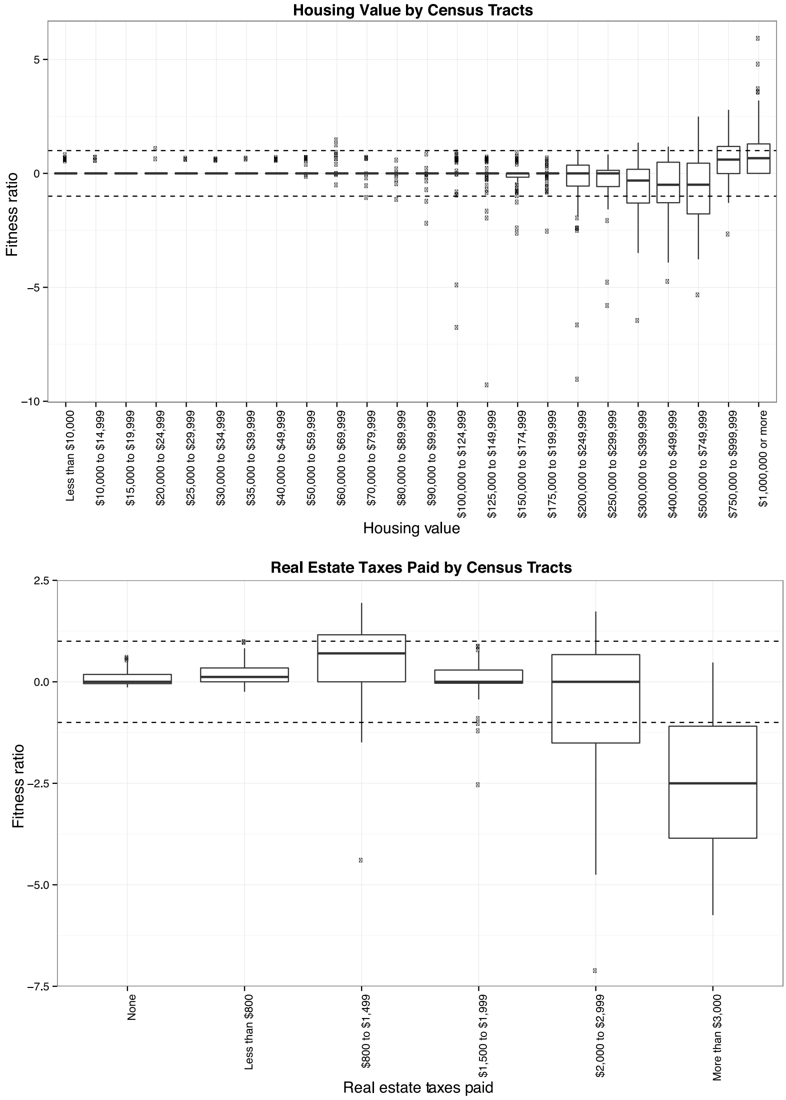


Notes: Boxplots compare the distribution of the fitness ratios for housing units valued between $500,000 to $749,999 and their relation to the ±90-percent American Community Survey (ACS) margins of error. Estimates falling outside the dashed reference lines or light gray color of 1 were not within the ±90-percent ACS margins of error.

Sources: Arlington County real estate assessments; 2009–2013 ACS 5-year estimates

Exhibit 15

Fitness Ratios for Number of Housing Units for Housing Value and Real Estate Taxes Paid by Census Tracts Using CoreLogic for Arlington County, 2009–2013



Notes: Boxplots display the distributions of the fitness ratios at the census tract level. Estimates falling outside the dashed reference lines of 1 were not within the ±90-percent American Community Survey (ACS) margins of error.
Sources: CoreLogic, Inc., real estate assessments, 2009–2013; 2009–2013 ACS 5-year estimates

Limitations and Opportunities of Using Local Administrative Data

Some challenges in the use of real estate assessment data are associated with the characteristics of administrative data. The time and effort it takes to acquire and prepare data can be substantial and varies by jurisdiction and data source. Data aggregators like CoreLogic may help to standardize the data. Such data come with their own set of limitations, however, as detail can potentially be lost through the standardization process. Without knowledge of the intricacies of the real estate assessment data on which they are based, it can be difficult to understand these limitations, which must be evaluated based on how the data will be used. One particular technological advance that would assist federal statistical agencies in their use of data from aggregators is to implement automated longitudinal editing to identify and correct errors. Doing so requires multiple years of data and can be done by data aggregators or federal statistical agencies. Other methods include "borrowing strength" from nearby geographical areas to identify possible issues and even correct those errors through modeling. For example, it is unlikely for changes in house values in northwestern Arlington to be very different from house prices trajectories in neighboring parts of Fairfax County, Virginia. *Dasymetric mapping*, a method of mapping that uses areal symbols to spatially classify volumetric data, may further support our findings by improving alignment between real estate assessment data and ACS estimates of the total number of housing units and number of units (see Leyk et al., 2013).

Another approach to improving administrative data is to reconcile differences using other sources of housing data. For example, multiple listing service (MLS) data on real estate transactions is another source for many of these same data. MLS data include information on the physical characteristics of a house and the transaction at the time the house was sold on the open market. To the extent that these data are available to survey organizations, it might be possible to model such items as housing value using hedonic indexes. More research on accuracy of other potential external data sources and the feasibility of incorporating such data to enhance and supplement federal statistics is needed.[6]

In principle, both the AC and the CL assessment data can be linked to confidential, internal ACS data for Arlington to examine the relationship of respondent-reported values to the assessment data. For instance, the CL assessment data do not include zero-bedroom units, and zero-bedroom units in the AC assessment data seemingly include housing units with one or more bedrooms based on their value—in fact, the highest-value housing unit in Arlington is listed as having zero bedrooms. Examining how zero-bedroom units appear in both data sets can provide insight into how external real estate data can be edited to better capture these housing units. Although such an undertaking is beyond the scope of this article, it would need to be at a larger scale than one urban county so the results could be generalizable. Replication of the work described in this article, along with additional attempts to compare microdata from the ACS and perhaps the AHS with administrative assessment data, could be used to move further along the path to productive use of administrative data in the federal sphere. One possibility is for the Census Bureau to start with addresses known to be in the ACS and the AHS and acquire CoreLogic data for those addresses.[7]

[6] See Keller et al. (2016) for an indepth inventory of housing data, additional comparisons of some of these other data sources, and an investigation of hedonic models using administrative data.

[7] To avoid revealing the addresses that appear in the survey, any request made to the vendor would have to include a large number of addresses not in the surveys, or the vendor's programmer would have to become a Special Sworn Status employee of the Census Bureau.

Conclusions

This article examined the potential of using local sources of administrative data to enhance, supplement, or replace federally collected survey data about housing units. We examined real estate assessment data for Arlington County, Virginia, obtained directly from the county and from a commercial data aggregator, and compared estimates derived from the assessment data at various geographic levels to corresponding estimates from the ACS. Our findings demonstrate that some real estate assessment data can be repurposed for statistical purposes. Three of the six real estate assessment data-based estimates examined—year built, house value, real estate taxes paid—align well with ACS estimates and may indeed be better. These data could potentially replace the corresponding questions in the survey. Real estate assessment data provide greater granularity, which opens the opportunity to conduct indepth research on local housing and neighborhood characteristics that is not possible with the ACS alone.

Whereas ACS housing estimates for small areas are available only as 5-year estimates, real estate assessment data are available annually at the parcel level, which means estimates can be produced each year at the block-group or even lower geographic levels. This level of detail opens up a wide range of possibilities for local government officials. They could use ACS data to benchmark their own administrative data-based estimates, and then use their own data for more frequent updates and for comparisons to nearby jurisdictions. One particular example is housing value. Using their own data, local government officials could track changes annually for geographic levels as small as a block, thereby providing an early peek at changes under way in the local housing market.

Appendix A. Data Quality Checks for Property Data

The steps outlined in this appendix constitute a guide for assessing the quality of property data. They are written in general terms; further steps might need to be taken depending on the jurisdiction/data source. Assessing quality of property data requires judgment and creativity. These steps are based on review of property data for 5 years from 2009 to 2013. Data for later or earlier years may have additional rules and hence require new checks.

I. **Unique identifier:** Each observation in the property data should come with a unique identifier. Ideally, this unique identifier will be a parcel number (ID), which will allow for quick linkage to other data sources.

  A. The parcel IDs should be unique for each year of data and specific county or political jurisdiction.

    i. If the parcel IDs are not unique, these parcels might be multi-dwelling properties, defined as separate housing units on the same parcel. Then one needs to ask whether the data flag or somehow indicate multi-dwelling properties?

  B. If the parcel IDs are formatted (for example, with dashes), they should be in a consistent format.

II. **Location:** Property data often come with geographic information.

A. Geographic coordinates (latitudes and longitudes) allow for properties to be placed geospatially with high levels of granularity.

i. One must know the coordinate reference systems, which define a specific map projection. Map projections allow for the representation of the globe on a two-dimensional surface. Without this information, it would be difficult to spatially join additional geographic information that might not be present in the current data (for example, census tracts).

B. Another type of geographic information in property data includes census tract codes. Census tracts are identified by up to a four-digit integer code that may have an optional two-digit suffix. Additional prefixes may be attached to identify the state and county by Federal Information Property Standard (FIPS) code. Within a property data set, these census tract codes should be in the same format.

i. If the census tract code is not consistent, there might be a pattern (for example, some might be tract codes and others block group codes; some might include the state FIPS code and others not).

ii. These census tract codes should represent the appropriate jurisdiction (for example, Washington, D.C. properties should have Washington, D.C. tracts). One can confirm the jurisdictions by crosschecking with the Census Tract Reference Maps[8] for that jurisdiction or breaking down the ID to FIPS (state and county) codes.

C. Property data might come with a corresponding shapefile that allows for the creation of maps and other data visualizations. It should be confirmed that the unique identifier for the shapefile is the parcel ID. If that is not the case, another way to merge property data onto the shapefile needs to be found beyond address as addresses often have different formats.

i. Condominiums can be treated in different ways in a shapefile. Condominium complexes can be represented by one polygon or they can be represented through multiple stacking polygons (one for each condominium within the complex). Each implies a different type of join when merging property data to the shapefile: a one-to-one relationship or a one-to-many relationship.

III. **Housing type classification:** Land use codes classify the type of housing located on a parcel (for example, single-family detached). Depending on how housing type is classified in the data, it might be preferable to recode land use into broader categories (single-family not elsewhere classified [NEC], single-family detached, single-family attached, condominium, and apartment) to aid in the data preparation process. One should confirm that parcels have been properly classified and that parcels not of interest are flagged (for example, hotels). One way is to plot the distribution of total value by property type/land use code.

[8] Census Tract Reference Maps can be found at: https://www.census.gov/geo/maps-data/maps/2010tract.html.

A. If the maximum value by property type is high, there might be an error in property type/classification (for example, If a condo is listed at $800 million, the property might be a condo complex or other multifamily building and not an individual condo). Checking other fields for these parcels might provide more evidence of misclassification.

B. Low values by property type (for example, a single-family unit listed as $2,000) are also of concern, as these parcels might be vacant land, common areas, multi-jurisdiction properties, or parking lots.

   i. Vacant land might have no improvement value, as no building is on the property. Parking lots might not have a land value. These rules depend on rules in the jurisdiction. Removing such properties temporarily and reevaluating the value distributions might aid in deciding which need to be removed from the analysis.

   ii. If only the improvement value is low, it may indicate vacant land or a property under construction, especially if other data (for example, the number of bedrooms) are missing.

C. Some property data that include multiple jurisdictions may also include standardized land use codes as land use codes are created by each individual jurisdiction. Confirming consistency in classification can be done by creating a cross-tab of standardized land use codes to the original land use codes to check for consistency (for example, the standardized land use code for condominiums is used for properties listed as townhouses or lowrise apartments in the original land use code) or through selecting a random sample and checking the standardized land use code to the property type listed on an online property search tool (search by way of parcel number). This method can be done if the data are for small area studies and/or if county land use codes are known.

IV. **Individual housing characteristics:** Each variable of interest needs to be individually examined with respect to consistency, validity, completeness, and uniqueness. In doing so, it is important to subset by land use code. For ease, it might be preferable to recode land use into broader property type categories and subset by single-family NEC, single-family detached, single-family attached, condominium, and apartment. Although the distribution might be valid overall, there could be patterns further down that are problematic (for example, are all missing bedroom counts for properties classified as multifamily properties? Do condo buildings go up to only a certain number of stories?).

A. *Variability:* Is there variability within each field in the data (for example, all properties listed as being owner occupied)? Is there variability by property type (for example, all condominiums listed as having paid $999 in property taxes)?

B. *Range:* The minimum and maximum values in each field should make sense. Often, this characteristic is a judgment call as there is no universal rule (for example, what is the maximum possible bedroom count or lot size?). Subsetting by property type will help pinpoint potential errors (for example, a condominium valued at $1 billion might actually be a multifamily building).

C. *Missing data and zeros:* Missing data can represent true missing data or zeros if zeros are not transferred from the original data. One way to check is to subset by a different variable (for example, subsetting housing value by number of bedrooms, which might show housing values that are low and have NA for number of bedrooms may actually be studios/efficiencies).

D. *Definitions:* Depending on the research purpose, how a housing characteristic is defined will affect if the data capture the same housing characteristic of interest (for example, heating can mean heating type or heating fuel). The definition should be consistent both across jurisdictions and within the field (for example, data on heating having both baseboard and electric baseboard, which are not mutually exclusive).

E. *Multiple years of data:* If the data are for a longitudinal study, locating large unexplained changes within the same parcel across the years will aid in data preparation. Missing values can be assigned from a previous or later year. Is there a pattern of missing data across the years (for example, missing values of year built occurring when a new building is under construction and thus the year built is updated when completed)?

Appendix B. Assigning Geographic Coordinates

The Arlington County real estate assessment (AC assessment) data do not include geographic coordinates. To overcome this issue, we created a master address list that contains all the unique addresses in the AC assessment data on which geographic information can be attached. Creating geographic coordinates by a parcel's address ensures that a parcel has the same coordinates and thus census boundary information across years. Occasionally parcels change addresses across the years. For example, 191 parcels in the AC assessment data have at least one different address during the period from 2009 to 2013. These addresses are recoded to match the majority or most complete address. Although the CoreLogic, Inc., real estate assessment (CL assessment) data do include longitudinally consistent geographic coordinates, to ensure geographic consistency with the AC assessment data, we recoded the CL assessment data coordinates. For completeness, we added addresses in CL assessment data that are not in the AC assessment data to the master address list.

We then used this master address list to geocode the addresses using the Google Maps API based on address. Google Maps API is a service provided by Google, Inc., to convert an address into geographic coordinates that can then be plotted on a map (Google, 2015).[9] Using Google Maps API geographic coordinates from the integrated master address file, we placed the addresses within the appropriate census tract and block group. The three main types of geocoding outputs are rooftop, approximate, interpolated.

[9] Access to the API to code large numbers of addresses can be obtained for a relatively low cost.

- *Rooftop* is the most accurate as it measures the exact location. The goal was to have all outputs be rooftop.
- *Approximate* is when Google Maps does not have either the specific address or a close match to the address, so it will output the centroid of the smallest known administrative area, for example, the center of Arlington County.
- *Interpolated* is when Google Maps does not have the specific street number but has close street numbers so it outputs the center between two points, for example, the ends of a street block.

To ensure that all addresses were appropriately placed within the county, we edited addresses to match the structure of Google Maps data (for example, with unit numbers removed and state information added). We then merged this master address list back into both the AC assessment data and CL assessment data by address with attached census tract and block group information. Due to inherent spatial errors that occur while geolocating (see Cayo and Talbot, 2003), 56 parcels that are on the border of Arlington had coordinates outside the county lines. Although off only by mere feet, the geocoded point is still technically outside the county polygon. The census tract and block group were manually edited for these data.

Acknowledgments

This article is based on a report prepared for the U.S. Census Bureau, under the supervision of Rick Knowles, MITRE Corporation, and Ronald Prevost, Census Bureau. We specifically acknowledge the contributions of our co-authors on that report: Dave Higdon, Mark Orr, Bianica Pires, and Kathryn Ziemer. We also acknowledge Andrea Morris, Andrew D'huyvetter, Elizabeth Hardy, Chester Carr, and Clay Gottschall from Arlington County for their help throughout this project.

Authors

Emily Molfino is a political scientist and postdoctoral associate at the Social and Decision Analytics Laboratory in the Biocomplexity Institute of Virginia Tech.

Gizem Korkmaz is a research assistant professor at the Social and Decision Analytics Laboratory in the Biocomplexity Institute of Virginia Tech.

Sallie A. Keller is a professor of statistics and Director of the Social and Decision Analytics Laboratory in the Biocomplexity Institute of Virginia Tech.

Aaron Schroeder is an information architect and data scientist at the Social and Decision Analytics Laboratory in the Biocomplexity Institute of Virginia Tech.

Stephanie Shipp is the Deputy Director of the Social and Decision Analytics Laboratory in the Biocomplexity Institute of Virginia Tech.

Daniel H. Weinberg is Principal at DHW Consulting and was a visiting scholar at the Social and Decision Analytics Laboratory in the Biocomplexity Institute of Virginia Tech.

References

Bazuin, Joshua Theodore, and James Curtis Fraser. 2013. “How the ACS Gets It Wrong: The Story of the American Community Survey and a Small, Inner City Neighborhood,” *Applied Geography* 45: 292–302.

Benítez-Silva, Hugo, Selçuk Eren, Frank Heiland, and Sergi Jiménez-Martín. 2010. “Using the Health and Retirement Study To Analyze Housing Decisions, Housing Values, and Housing Prices,” *Cityscape* 12 (2): 149–158.

Brummet, Quentin. 2014. Comparison of Survey, Federal, and Commercial Address Data Quality. Center for Administrative Records Research and Applications. Working paper 2014-06. Washington, DC: U.S. Census Bureau.

Cayo, Michael R., and Thomas O. Talbot. 2003. “Positional Error in Automated Geocoding of Residential Addresses,” *International Journal of Health Geographics* 2 (1): 1–12.

DiPasquale, Denise, and C. Tsuriel Somerville. 1995. “Do House Price Indices Based on Transacting Units Represent the Entire Stock? Evidence From the American Housing Survey,” *Journal of Housing Economics* 4 (3): 195–229.

Dippo, Cathryn S. 1997. “Survey Measurement and Process Improvement: Concepts and Integration.” In *Survey Measurement and Process Quality*, edited by Lars Lyberg, Paul Biemer, Martin Collins, Edith De Leeuw, Cathryn Dippo, Norbert Schwartz, and Dennis Trewin. New York: John Wiley & Sons: 455–474.

Goodman, John L., and John B. Ittner. 1992. “The Accuracy of Home Owners’ Estimates of House Value,” *Journal of Housing Economics* 2 (4): 339–357.

Google. 2015. “Web Services: Geocoding API.” https://developers.google.com/maps/documentation/geocoding/intro.

Ihlanfeldt, Keith R., and Jorge Martinez-Vazquez. 1986. “Alternative Value Estimates of Owner-Occupied Housing: Evidence on Sample Selection Bias and Systematic Errors,” *Journal of Urban Economics* 20 (3): 356–369.

Jarosz, Beth, and John Hofmockel. 2010. “Research Note: What Counts as a House? Comparing 2010 Census Counts and Administrative Records,” *Population Research and Policy Review* 32 (5): 753–765.

Kain, John F., and John M. Quigley. 1979. “Note on Owner’s Estimate of Housing Value,” *Journal of the American Statistical Association* 67 (340): 803–806.

Keller, Sallie, Stephanie Shipp, Mark Orr, Dave Higdon, Gizem Korkmaz, Aaron Schroeder, Emily Molfino, Bianica Pires, Kathryn Ziemer, and Daniel Weinberg. 2016. *Leveraging External Data Sources To Enhance Official Statistics and Products*. Report prepared by the Social and Decision Analytics Laboratory, Biocomplexity Institute of Virginia Tech. Washington, DC: U.S. Census Bureau.

Kiel, Katherine A., and Jeffrey E. Zabel. 1999. "The Accuracy of Owner-Provided House Values: The 1978–1991 American Housing Survey," *Real Estate Economics* 27 (2): 263–298.

Kingkade, W. 2013. "Self-Assessed Housing Values in the American Community Survey: An Exploratory Evaluation Using Linked Real Estate Records." Paper presented at the 2013 Joint Statistical Meetings, Montréal, Québec, Canada, August 3–8.

Leyk, Stefan, Barbara P. Buttenfield, Nicholas N. Nagle, and Alexander K. Stum. 2013. "Establishing Relationships Between Parcel Data and Land Cover for Demographic Small Area Estimation," *Cartography and Geographic Information Science* 40 (4): 305–315.

Moore, Bonnie. 2015. *Preliminary Research for Replacing or Supplementing the Year Built Question on the American Community Survey With Administrative Records*. Report prepared for the U.S. Census Bureau. Washington, DC: U.S. Census Bureau.

Ruggles, Patricia. 2015. *Review of Administrative Data Sources Relevant to the American Community Survey*. Report prepared for the U.S. Census Bureau. Washington, DC: U.S. Census Bureau.

State of Virginia. 1996. "Virginia State Building Code." http://www.dhcd.virginia.gov/index.php/va-building-codes/building-and-fire-codes/regulations/virginia-state-building-codes-and-regulations-1996-present.html.

U.S. Census Bureau. 2014a. "American Community Survey—Design and Methodology." Version 2.0. http://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/acs_design_methodology_report_2014.pdf.

———. 2014b. "American FactFinder 2009–2013." http://factfinder2.census.gov.

Weinberg, Daniel H. 2015. "Data Sources for U.S. Housing Research, Part 2: Private Sources, Administrative Records, and Future Directions," *Cityscape* 17 (1): 191–205.

———. 2014. "Data Sources for U.S. Housing Research, Part 1: Public Sector Data Sources," *Cityscape* 16 (3): 131–147.