Statistical Journal of the IAOS 34 (2018) 543–551
DOI 10.3233/SJI-180452
IOS Press

Evaluating the utility of a commercial data source for estimating property tax amounts

Zachary H. Seeskin
NORC at the University of Chicago, 55 E. Monroe Street, 31st Floor, Chicago, IL 60603, USA
Tel.: +1 312 325 2506; E-mail: Seeskin-Zachary@norc.org

Abstract. This paper investigates the utility of a commercial property tax data from CoreLogic, Inc. (CoreLogic) aggregated from county and township governments from across the country, for use to improve American Community Survey (ACS) estimates of property tax amounts for single-family homes. Particularly, the research uses linkages of the CoreLogic file to the 2010 ACS to evaluate the use of CoreLogic data directly to replace survey responses for estimation of property tax amounts, potentially reducing measurement error and respondent burden. I find that the coverage of CoreLogic data varies among geographic areas across the U.S., as does the correspondence between ACS and CoreLogic property taxes. Large differences between CoreLogic and ACS property taxes in some instances may reflect conceptual differences between what is collected in the two data sources for certain counties. This research draws attentions to the challenges of using non-survey data sources that are aggregated from many state or local agencies with different practices for data collection and curation.

Keywords: Data quality, respondent burden, measurement error, linked data

1. Introduction

Administrative records and commercial data can be inexpensive data sources for official statistics and offer some strengths that mitigate weaknesses of censuses and surveys. In particular, surveys place burden on respondents, are subject to errors in responses and can have high levels of nonresponse. Administrative records and commercial data, when of sufficient quality, can be less prone to errors in recordkeeping and offer broad coverage of the population. In some cases, they can even eliminate the need for questions on surveys. Yet, quality can vary across different data sources, as administrative records and commercial data are not collected for statistical purposes. Thus, careful evaluations are needed before using administrative records or commercial data for statistical products.

This paper evaluates commercial property tax data available from CoreLogic, Inc. (CoreLogic) for improvement of survey estimates of property tax amounts from the American Community Survey (ACS), examining 2010 data. CoreLogic aggregates property tax records from counties and townships across the country into one dataset. I focus on single-family homes, for which the record linkage is less challenging than for multi-unit structures.

Specifically, I evaluate whether the CoreLogic data are of sufficient quality that the data can be used in place of asking a question about property taxes on the ACS or to substitute for ACS responses. A major concern for the ACS is the respondent burden from the survey length and content. Thus, I consider the possibility of using CoreLogic alone to construct property tax estimates for geographic areas across the U.S. Separate research investigates the utility of CoreLogic data to potentially mitigate the effects of survey nonresponse [1].

I find that the quality of the CoreLogic data varies among counties and townships across the country, both in the coverage of the CoreLogic data and in the correspondence between ACS and CoreLogic property tax values. In some counties, large differences are found between the ACS and CoreLogic records, possibly due to conceptual differences between what is collected in the two sources. This demonstrates the challenge of using data aggregated from many state and local agencies. These findings do not support the use of Core-

Logic nationwide in place of asking about property taxes on the ACS. Nonetheless, there may be counties where CoreLogic can be viewed as a "gold standard" for property tax amounts. Further research could work to identify these counties and townships and determine if the CoreLogic data should be used in place of survey responses.

Section 2 discusses background literature related to reporting error for housing statistics and the use of administrative records and commercial data to address survey reporting error. Section 3 provides an overview of the CoreLogic property tax data file, while Section 4 investigates the quality of the CoreLogic and compares ACS and CoreLogic reports of property taxes. Section 5 concludes by discussing the implications of the research both for using the CoreLogic data for ACS property tax estimates and more broadly for other uses of commercial data for federal statistical products.

2. Background

The American Community Survey (ACS) is one important source of housing statistics for the U.S. The large sample size of the ACS allows for producing estimates in geographic areas across the U.S., including census block groups for the ACS 5-year estimates. ACS property tax estimates are used for research on housing affordability, for determining formula block grant funds, for mass transportation and metropolitan planning, for determining eligibility for housing assistance, and to inform efforts to plan affordable housing [2,3].

Reporting error on surveys is a concern for estimates of property tax amounts. A content reinterview survey of the 2012 ACS was conducted and found a moderate level of disagreement [4], indicating the potential for reporting error. While property tax amount response error has not been thoroughly studied, past research has found reporting error to be a concern when studying the related topic of home value. A comparison of survey responses on the 1979–1991 American Housing Survey metropolitan samples to the sale prices of the homes that were sold in the twelve months before the survey interview and found that survey responses tended to be higher than selling prices [5]. A separate study compared survey-reported home values from the Health and Retirement Study to sales prices and also found that the survey responses were greater than sales prices [6].

While there are few examples of commercial data being used to address response error for government statistics, research has shown some compelling examples using administrative data. Much of the research in this area has pertained to program receipt. For example, the Census Bureau is using Social Security Administration data linked to the Survey of Income and Program Participation to correct responses about supplementary security income receipt and disability insurance receipt [7]. Medicaid records have been used to adjust Current Population Survey estimates of Medicaid for underreporting [8]. Other studies have examined linking the Current Population Survey with administrative records for the Supplemental Nutrition Assistance Program, Temporary Assistance for Needy Families, General Assistance and housing assistance to improve estimates of program receipt and poverty [9,10].

3. CoreLogic data

The CoreLogic, Inc. 2008–2010 property tax file (CoreLogic) aggregates property tax records from counties and townships across the U.S. While the majority of the records on the file are listed as from 2009, there also records from 2008 and 2010. The full file contains more than 169 million records and includes information on a rich set of housing characteristics, including property value, tax amount, physical and structural characteristics, mortgage, sales and ownership information and geography. The fields available can differ among counties and townships.

Using the geographic and address information from CoreLogic records, the Census Bureau linked the CoreLogic file to the Census Bureau's Master Address File (MAF), through which CoreLogic records are linked with records from the ACS and other Census Bureau products. The linkage procedure standardizes addresses and uses information on house number, street prefix, directional prefix, street name, street suffix, direction suffix, apartment number/description, and five-digit zip code to conduct the linkage [11]. A two-step probabilistic matching process is conducted.

Challenges in linking CoreLogic to the MAF are further described in [11]. Overall, 63.4 percent of records are linked to the MAF. In studying the linkage of CoreLogic to the 2009 American Housing Survey through the MAF, 79.0 percent of single-unit structures are successfully linked, compared with only 14.8 percent of multi-unit structures.

I examine single-family, owner-occupied records from the ACS and CoreLogic, due the greater availability of linked CoreLogic data for single-unit structures and because the ACS only asks owner-occupied households are asked about their property taxes. Certainly, future research could investigate the quality of CoreLogic information for renter-occupied units.

Previous research conducted by Census Bureau researchers has studied using CoreLogic data for estimates of home values and year that a structure is built. A study of how CoreLogic and 2009 ACS home values compare for single-family homes found that ACS home values tend to be higher than the values from CoreLogic [12]. The difference between ACS and CoreLogic home values tends to increase with the time since the last move, which suggests that recent movers better estimate the value of their homes. An evaluation of CoreLogic data for the year that a structure is built in the 2012 ACS finds that 56.7 percent of single-family, detached homes in the ACS can be linked to CoreLogic records with year built information available [13]. Further, agreement for the time range in which a structure was built between ACS and CoreLogic was found for 78.3 percent of the linked records with reported year built information.

4. Results

4.1. Availability of linked CoreLogic data

In the 2010 ACS file, there are 1,116,568 records for single family, owner-occupied households. Among these, 69.1 percent were linked to CoreLogic records with property tax information available. When property tax information was not available, it may have been due to one of a few reasons: that no corresponding record was available from CoreLogic, that the CoreLogic record was available but the linkage to the ACS was not successful, or that a CoreLogic record was linked but the record did not contain property tax information. I exclude linkages when the ACS report of the year the structure was built was 2009 or 2010, as the CoreLogic file may not have been updated to reflect the recently built structures.

The availability of CoreLogic property tax information varies across states, counties and townships. The match rates for states are presented in Table 1. In Nevada, 89.6 percent of single-family, owner-occupied households in the 2010 ACS are linked to CoreLogic property tax information, while linked CoreLogic tax information is not available in Montana, New Hampshire or Vermont.

The availability of linked CoreLogic tax information also varies by household characteristics, as presented in Table 2. Notably, 78.5 percent of ACS households in urban areas are linked to CoreLogic tax information, compared with only 53.0 percent of ACS households in rural areas. Households of higher socioeconomic status are also better represented among linked CoreLogic records than are households of lower socioeconomic status, a finding similar to that found in other studies of administrative record linkage to surveys [14]. Of households not in poverty, 69.6 percent have linked CoreLogic information compared with only 60.7 percent of households in poverty. When the householder is a college graduate, 73.7 percent of households have CoreLogic information compared with only 62.5 percent of households where the householder did not graduate high school. In Table 3, which compares characteristics for ACS records with and without linked CoreLogic property tax information, the median household income for records with CoreLogic information is almost $68,000 while the median household income for records without CoreLogic information is about $56,000. These findings demonstrate a strong association between the availability of CoreLogic data and household socioeconomic status and education. Further, these relationships hold within counties, as shown by estimating a multivariate logistic regression model among ACS records with CoreLogic property tax data using county fixed effects to condition on geographic differences [1].

4.2. Correspondence of ACS and CoreLogic property taxes

In order to evaluate the CoreLogic data, I compare responses for property taxes in CoreLogic and the 2010 ACS. A major challenge in interpreting the comparisons is that both data sources may be prone to errors. The ACS suffers from respondent error, and CoreLogic data are only as accurate as the property tax records aggregated from counties and townships by CoreLogic. Nonetheless, comparing property taxes from the two data sources can help with evaluating CoreLogic's usefulness and help better understand errors in ACS responses.

Across the U.S., there is an overall Pearson correlation of 0.724 between ACS and CoreLogic property taxes when both are reported and available. Since ACS and CoreLogic records are linked, considering the percentage difference between ACS and CoreLogic property taxes is useful. The percentage difference is de-

Table 1
Percentage of ACS records linked to CoreLogic property tax information by state

State	Match rate (%)	Number of records	State	Match rate (%)	Number of records
Nevada	89.6	6,673	Utah	67.2	9,916
California	87.7	90,958	Minnesota	66.5	39,358
Maryland	87.2	19,719	New York	65.2	53,141
New Jersey	87.0	28,908	New Mexico	62.8	v6,575
Rhode Island	86.9	3,166	Kentucky	62.4	16,845
Ohio	83.7	48,811	Wyoming	62.3	2,221
Connecticut	79.6	12,927	Michigan	61.8	52,827
Massachusetts	79.4	20,213	District of Columbia	61.2	1,221
Oregon	78.1	13,191	Oklahoma	59.0	17,068
Virginia	78.0	27,383	Mississippi	57.9	9,592
Illinois	77.7	48,943	Missouri	57.6	26,795
Texas	76.6	74,408	Alabama	57.2	18,422
Georgia	75.7	28,659	Iowa	56.2	19,884
Washington	75.1	23,262	Maine	53.6	7,929
Delaware	75.0	3,747	Nebraska	49.1	11,182
Louisiana	75.0	15,164	Alaska	44.8	2,750
Wisconsin	74.9	39,081	West Virginia	42.1	7,782
Arizona	74.0	17,742	Hawaii	35.0	3,538
North Carolina	73.9	31,382	South Dakota	32.6	4,876
South Carolina	73.8	14,452	North Dakota	23.3	4,875
Pennsylvania	73.5	64,331	Kansas	8.6	14,489
Colorado	73.1	18,340	Tennessee	1.8	22,516
Indiana	72.1	27,681	Montana	0.0	5,080
Florida	69.6	51,019	New Hampshire	0.0	6,059
Idaho	69.3	6,138	Vermont	0.0	4,498
Arkansas	67.8	10,831			
			United States	69.1	1,116,568

Source: 2010 ACS single-family, owner-occupied households linked to 2008–2010 CoreLogic data.

Table 2
ACS match rates with CoreLogic property tax information by household characteristics

Group	Match rate (%)	Number of records
Education level of householder		
No high school diploma	62.5	99,846
High school diploma or G.E.D.	64.7	292,649
Some college	69.6	334,973
College graduate	73.7	389,100
Poverty status		
In poverty	60.7	65,328
Not in poverty	69.6	1,051,240
Urbanicitys		
Urban	78.5	705,697
Rural	53.0	410,871
Overall	**69.1**	**1,116,568**

Source: 2010 ACS single-family, owner-occupied households.

Table 3
ACS characteristics for records with and without linked CoreLogic property tax information

Group	Records with matches	Records without matches
Median household income ($)	67,865	56,005
Median home value ($)	189,000	150,000
Median property taxes paid ($)	2,100	1,500
Number of records	771,582	344,986

Source: 2010 ACS single-family, owner-occupied households.

fined to be

$$100 \times \left(\frac{ACS - CoreLogic}{CoreLogic} \right),$$

where *ACS* and *CoreLogic* are the respective property tax measures from the two sources.

Table 4 presents quantiles of the percentage difference for linked records by different household characteristics. Overall, the median percentage difference is 0.0 percent. The 5^{th} and 95^{th} percentiles and the interquartile range, the difference between the 75^{th} and 25^{th} percentiles, are presented to study the spread of the percentage difference by characteristic. While for most household characteristics, the median percentage difference is near 0.0 percent, the interquartile range varies. The interquartile range tends to be greater for households with characteristics associated with greater response error, such as low socioeconomic status [15]. The interquartile range is 16.6 percent for households who respond to the survey questionnaire, but 29.1 percent for CATI and 28.4 percent for CAPI. The interquartile range is 28.6 percent when the householder does not have a high school diploma, but 15.7 percent when the householder is a college graduate. House-

Table 4
Distribution of percentage difference of ACS property taxes from CoreLogic property taxes by household characteristics

Household Characteristic	Percentiles for % diff. of ACS from CoreLogic 5^{th}	25^{th}	50^{th}	75^{th}	95^{th}	Interquartile range	Number of records
Response mode							
Questionnaire	−56.7	−8.9	0.0	7.7	83.6	16.6	555,296
CATI	−65.5	−16.3	−0.8	12.7	109.0	29.1	75,693
CAPI	−58.8	−15.4	−0.7	12.9	103.7	28.4	45,853
Race of householder							
White	−54.5	−9.2	0.0	8.2	84.6	17.3	559,601
Black	−89.3	−21.4	−0.3	13.7	145.2	35.2	40,824
Hispanic	−72.2	−20.2	−1.8	9.4	88.0	29.6	28,271
Asian	−51.5	−7.8	0.0	6.4	64.2	14.1	24,415
Other race	−66.9	−12.9	−0.1	11.0	108.2	23.9	23,731
Education level of householder							
No high school diploma	−87.5	−17.0	−0.1	11.6	130.9	28.6	50,125
High school diploma or G.E.D.	−66.2	−11.0	0.0	10.0	108.5	21.0	160,840
Some college	−56.2	−10.0	0.0	9.2	87.9	19.1	204,305
College graduate	−50.0	−8.9	0.0	6.8	69.1	15.7	261,572
Year moved							
1989 or earlier	−67.5	−10.0	0.0	8.8	101.9	18.7	209,359
1990–1999	−53.7	−9.9	0.0	7.6	79.5	17.4	166,764
2000–2004	−51.5	−9.6	0.0	7.9	75.0	17.4	137,929
2005–2010	−56.7	−11.0	−0.1	9.6	92.8	20.6	162,790
Poverty status							
In poverty	−88.1	−18.1	−0.1	12.6	136.0	30.6	31,241
Not in poverty	−56.6	−9.8	0.0	8.3	86.3	18.1	645,601
Overall	−58.2	−10.1	0.0	8.5	88.4	18.5	676,842

Source: 2010 ACS single-family, owner-occupied households linked to 2008–2010 CoreLogic records.

holds in poverty have an interquartile range of 30.6 percent, while the interquartile range for households not in poverty is 18.1 percent.

Interestingly, the interquartile range does not vary as much by the year moved, suggesting that survey recall of property tax amounts differs from patterns for home values found in research [5]. However, while the interquartile range is not as sensitive to the year moved, the 5^{th} and 95^{th} percentiles are somewhat sensitive. For households where the respondent has not moved since 1989 or earlier, the 5^{th} percentile for the percentage difference is -67.5 percent and the 95^{th} percentile is 101.9 percent, which are both greater in magnitude than the 5^{th} (-58.2 percent) and 95^{th} (88.4 percent) percentiles of the percentage difference for households overall.

While comparisons by household characteristics may reflect patterns in ACS response error, comparing ACS and CoreLogic property taxes by geographic area can possibly help with understanding errors in the CoreLogic data. As the property tax data are maintained by different authorities for each county and township, it is not surprising that CoreLogic's quality and accuracy vary by county. Some patterns emerge by examining statistics for the percentage difference by large county in Table 5 and Fig. 1.

Across counties, the distributions of the percentage difference between ACS and CoreLogic property taxes can differ greatly. Many counties have a median percentage difference near 0.0 percent. However, there are some geographic areas for which ACS and CoreLogic property taxes disagree. For example, four counties in Texas have median percentage differences less than -10.0%, indicating that CoreLogic property taxes tend to be greater than those of the ACS. By contrast, in Fulton County, GA and Nassau County, NY, CoreLogic property taxes tend to be less than the ACS reports. This variation across geographies may reflect differences among local property tax authority practices and the extent to which property tax records reflect the amount that households are actually billed.

Examining the interquartile range as a measure of spread of the percentage difference can help with assessing the accuracy of CoreLogic property taxes compared with ACS numbers. Among the smallest interquartile ranges are those of Milwaukee County, WI (5.9 percent) and Wake County, NC (6.8 percent). On the other hand, Dallas County, TX has an interquartile range of 42.6 percent and Harris County, TX has an interquartile range of 79.0 percent. The spread of the percentage difference distribution for a county being much smaller than the distribution for the U.S., as

Table 5
Distribution of percentage difference of ACS property taxes from CoreLogic property taxes by county

	Percentiles for % difference of ACS from CoreLogic					Interquartile	ACS-CoreLogic	Number of
State	5^{th}	25^{th}	50^{th}	75^{th}	95^{th}	range	correlation	records
Fulton Cty, GA	−49.9	−5.4	4.8	25.2	144.3	30.6	0.91	1,577
Nassau Cty, NY	−59.9	−4.4	4.5	20.9	73.3	25.4	0.86	3,969
Allegheny Cty, PA	−83.9	−16.0	2.4	15.7	70.2	31.7	0.83	4,371
Middlesex Cty, MA	−47.2	−5.8	1.9	6.8	23.8	12.5	0.92	3,156
Suffolk Cty, NY	−25.4	−2.5	0.8	9.5	47.4	12.0	0.86	4,346
Salt Lake Cty, U	−32.3	−4.1	0.6	8.9	43.8	13.0	0.73	2,586
King Cty, WA	−49.6	−8.5	0.6	6.8	40.6	15.3	0.90	3,856
Fairfax Cty, VA	−62.9	−8.7	0.5	5.3	57.1	14.0	0.76	2,744
Broward Cty, FL	−41.2	−4.3	0.2	10.0	54.0	14.4	0.91	2,106
Oakland Cty, MI	−39.3	−7.6	0.0	8.7	45.4	16.3	0.87	3,931
Montgomery Cty, MD	−49.6	−8.5	0.0	8.7	53.5	17.2	0.78	2,759
Saint Louis Cty, MO	−46.4	−4.9	0.0	4.5	26.2	9.4	0.92	3,592
Cuyahoga Cty, OH	−60.2	−10.1	0.0	4.4	36.7	14.5	0.91	4,115
Santa Clara Cty, CA	−37.6	−5.8	0.0	3.4	62.4	9.2	0.84	3,903
Mecklenburg Cty, NC	−45.8	−5.3	0.0	2.8	37.6	8.0	0.84	2,055
Contra Costa Cty, CA	−47.3	−7.4	0.0	4.3	50.9	11.7	0.84	2,486
Wayne Cty, MI	−52.2	−9.7	0.0	9.8	56.7	19.5	0.71	4,593
Milwaukee Cty, WI	−29.5	−4.2	0.0	1.8	20.1	5.9	0.89	2,268
Wake Cty, NC	−36.1	−5.0	0.0	1.8	31.3	6.8	0.85	2,452
Sacramento Cty, CA	−52.8	−8.0	0.0	5.6	63.7	13.6	0.75	3,262
Riverside Cty, CA	−50.4	−9.6	0.0	7.2	68.5	16.8	0.71	4,391
Los Angeles Cty, CA	−50.0	−8.4	0.0	3.8	52.6	12.2	0.85	15,368
Orange Cty, CA	−44.5	−6.7	0.0	4.1	43.5	10.8	0.88	6,422
Fresno Cty, CA	−51.2	−8.7	−0.1	6.9	80.3	15.6	0.47	1,492
San Bernardino Cty, CA	−50.4	−9.2	−0.1	6.9	75.8	16.1	0.81	3,182
Alameda Cty, CA	−43.4	−8.0	−0.1	2.9	46.7	10.8	0.90	3,235
Philadelphia Cty, PA	−33.8	−3.5	−0.1	6.7	79.7	10.2	0.82	2,925
San Diego Cty, CA	−49.6	−7.3	−0.1	2.8	60.1	10.2	0.86	5,780
Hillsborough Cty, FL	−42.3	−5.5	−0.2	11.1	75.1	16.6	0.93	2,639
Pima Cty, AZs	−53.4	−11.8	−0.8	2.9	40.8	14.7	0.82	2,267
Hennepin Cty, MN	−47.5	−9.0	−1.0	1.8	22.7	10.8	0.93	3,326
Franklin Cty, OH	−52.1	−12.1	−1.3	0.6	30.6	12.7	0.76	2,885
Queens Cty, NY	−51.4	−13.6	−1.5	5.4	97.8	19.1	0.66	1,026
Kings Cty, NY	−56.2	−15.3	−1.8	5.1	116.0	20.4	0.60	516
Honolulu Cty, HI	−68.7	−14.3	−2.1	7.7	78.0	22.0	0.34	613
Maricopa Cty, AZ	−51.9	−15.4	−2.3	1.9	42.7	17.3	0.84	6,944
Cook Cty, IL	−46.3	−12.1	−2.3	10.0	86.6	22.1	0.91	9,043
Orange Cty, FL	−42.1	−8.7	−2.6	5.8	54.5	14.5	0.85	2,373
Palm Beach Cty, FL	−40.3	−8.6	−2.7	4.7	44.3	13.3	0.82	2,624
Clark Cty, NV	−62.3	−22.5	−3.7	2.9	55.16	25.4	0.81	3,514
Bexar Cty, TX	−86.8	−36.8	−6.3	4.5	134.9	41.3	0.89	3,632
Westchester Cty, NY	−66.6	−24.1	−7.9	3.6	136.2	27.7	0.44	1,752
Travis Cty, TX	−70.2	−25.0	−10.9	−5.6	15.4	19.4	0.90	2,246
Tarrant Cty, TX	−83.8	−34.2	−11.4	−2.3	21.6	31.9	0.93	4,316
Harris Cty, TX	−98.1	−44.6	−12.7	34.3	108.6	79.0	0.76	7,396
Dallas Cty, TX	−98.5	−50.4	−20.2	−7.8	27.0	42.6	0.79	4,480

Source: 2010 ACS single-family, owner-occupied households linked to 2008–2010 CoreLogic records in select large counties.

for Milwaukee County and Wake County, may provide a reason to have more confidence in those counties' CoreLogic data.

5. Discussion

The findings of this paper illustrate some of the major challenges with using commercial data for official statistics. As the CoreLogic property tax data are aggregated from counties and townships around the country, the quality of the data varies across geographic areas and is subject to the practices of each local property tax authority. The amounts recorded on property tax records may not reflect the property taxes that are actually billed. For example, in several counties, particularly in Texas, large differences between the CoreLogic and ACS property tax amounts indicate

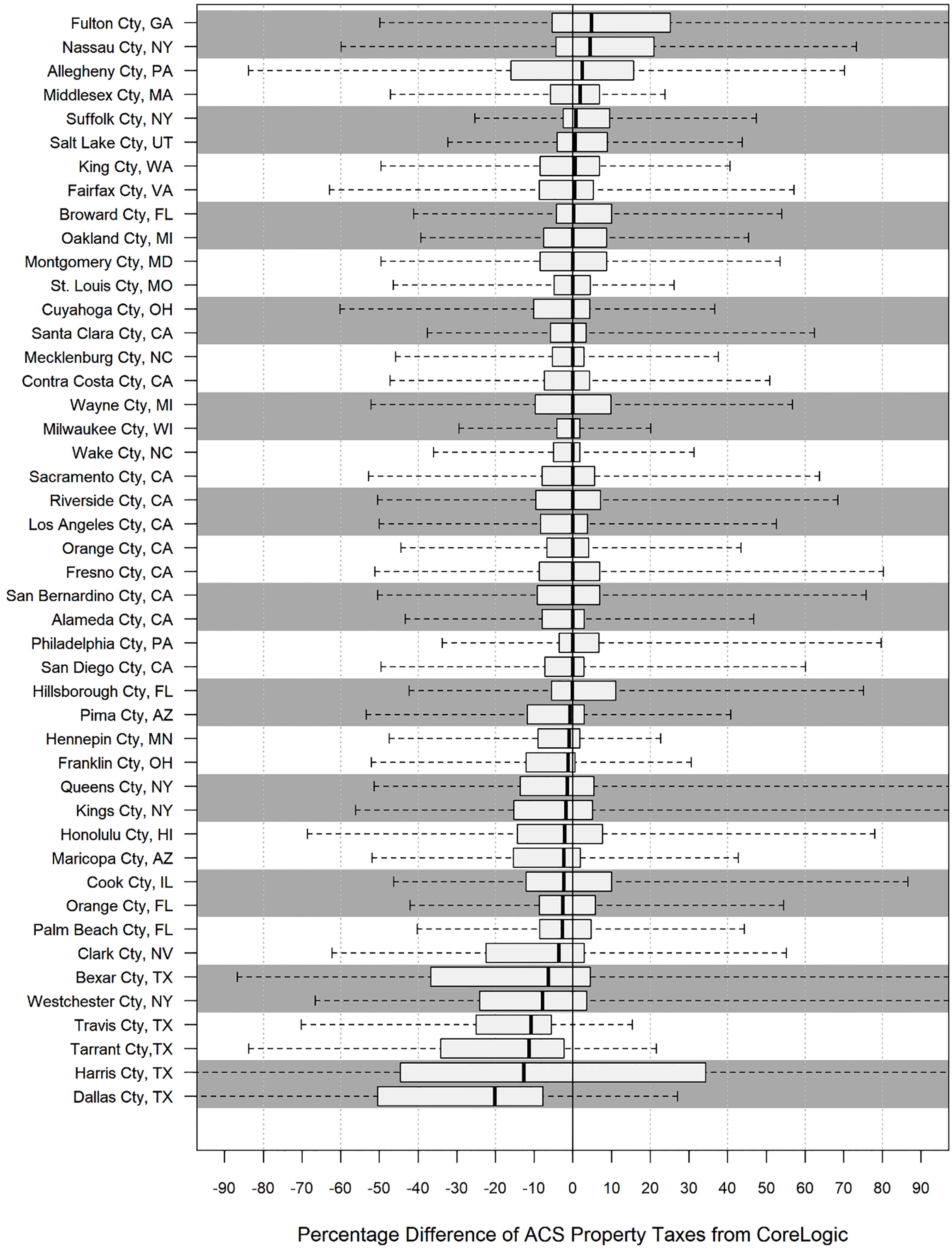

Fig. 1. Boxplots of percentage difference of ACS property taxes from linked CoreLogic property taxes by select counties. Whiskers indicate 5$^{\text{th}}$ and 95$^{\text{th}}$ percentiles.

that the CoreLogic data may reflect a different concept than that measured by the ACS. Even in counties CoreLogic for which property taxes are comparable to ACS reports as measured by the median percentage difference, the spread of the percentage difference can be large, which provides doubt for using CoreLogic data directly for estimates.

Further research can help improve understanding of the challenges with such data sources as CoreLogic and inform future use of commercial data to improve survey-based estimates. First, if counties and townships can be identified where the CoreLogic data is a "gold standard", then the Census Bureau should consider using CoreLogic data instead of survey responses in these counties. Further work would be needed to identify these counties. Obtaining a third independent data source with property tax information, if one can be found, is one possible way to verify the property tax data. It may also be helpful to hold discussions with local property tax authorities to better understand the data. In addition, even when commercial data sources do not constitute a gold standard, the data may still be valuable to improve estimates by providing auxiliary information, such as for imputation modeling or for small area estimation [1] finds that using CoreLogic data modestly improves the predictive power of imputation models for property taxes, while having minimal impact on estimates. Further research can improve understanding of the value of commercial data as auxiliary information to support surveys, including accounting for when data are aggregated from many jurisdictions.

There are some limitations of the methods of this research conclusions for using CoreLogic. First, the research focused on single-family homes and does not consider other kinds of structures. Prior studies have documented the difficulties of using CoreLogic for multi-unit structures in surveys. Future research can study using CoreLogic for ACS multi-unit structure property taxes, although additional challenges would likely emerge. Second, the research does not use a "gold standard" measure of property taxes to verify the CoreLogic records. Without a "gold standard" measure, assessing the accuracy of the CoreLogic data is limited to comparing CoreLogic records to the ACS, which is subject to response error.

Commercial data offer great promise for official statistics and can mitigate some weaknesses of surveys. However, the research demonstrates the set of challenges that can emerge when data are collected and maintained by many authorities throughout the country. As new approaches toward federal statistical products are considered in the future, careful evaluations of these data sources will continue to be needed.

Acknowledgments

This research was supported through a U.S. Census Bureau Dissertation Fellowship during my doctoral studies at Northwestern University through contract YA-1323-15-SE-0097. All views expressed are solely my own.

There are many individuals who I want to thank for their support and advice for this research. I thank Tommy Wright and Amy O'Hara for making this research possible and for connecting me with researchers at the Census Bureau. Bruce Spencer, my doctoral advisor, has provided incredible guidance. This research has benefited from conversations with and comments from many individuals, including Trent Alexander, Stephen Ash, Aileen Bennett, Quentin Brummet, Shawn Bucholtz, Bob Callis, George Carter, Tamara Cole, Art Cresce, Diane Cronkite, Craig Cruse, Denise Flanagan Doyle, Larry Hedges, Howard Hogan, Don Jang, Andrew Keller, Ward Kingkade, Arend Kuyper, Tom Louis, Chris Mazur, Chuck Manski, Carla Medalia, Bonnie Moore, Darcy Steeg Morris, Tom Mule, Mary Mulry, Michaela Patton, Steven Pedlow, Tom Petkunas, David Raglin, Michael Ratcliffe, Jerry Reiter, Kristine Roinestad, Joe Schafer, Jacob Schauer, David Sheppard, Eric Slud, Matt Streeter, Lars Vilhuber, Adeline Wilcox, Ellen Wilson, Bill Winkler, and Alan Zaslavsky.

Data analyses were conducted at the Chicago Census Research Data Center. I thank Trent Alexander, Stephanie Bailey, Quentin Brummet, Frank Limehouse, Joey Morales, Amy O'Hara, Danielle Sandler and others for their assistance at the Research Data Center.

References

[1] Seeskin ZH. Evaluating the use of commercial to improve survey estimates of property taxes. CARRA Working Paper #2016-06. Washington, D.C.: U.S. Census Bureau; 2016 Aug 1.

[2] U.S. Census Bureau. ACS questions and current federal uses. Washington, D.C.; 2014 Oct 1. Available from: https://www.census.gov/programs-surveys/acs/operations-and-administration/2014-content-review/federal-uses.html.

[3] Ruggles P. Review of administrative data sources relevant to the American Community Survey. Washington, D.C.: U.S. Census Bureau; 2015 Jan 31. Available from: https://www.census.gov/content/dam/Census/library/working-papers/2015/acs/2015_Ruggles_01.pdf.

[4] Murphy P. American Community Survey 2012 Content Reinterview Survey. Washington, D.C.: U.S. Census Bureau; 2014 Jan 6. Available from: http://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014_Murphy_01.pdf.

[5] Kiel KA, Zabel JE. The accuracy of owner-provided house values: The 1978–1991 American Housing Survey. Real Estate Economics. 1999 Jun 1; 27(2): 263-98.

[6] Benítez-Silva H, Eren S, Heiland F, Jiménez-Martín S. How well do individuals predict the selling prices of their homes? Working Paper, Levy Economics Institute. 2008; 571.

[7] Giefer K, Williams A, Benedetto G, Motro J. Program confusion in the 2014 SIPP: Using administrative records to correct false positive SSI reports. FCSM 2015 Proceedings. 2015.

[8] Davern M, Call KT, Ziegenfuss J, Davidson G, Beebe TJ, Blewett L. Validating health insurance coverage survey estimates: A comparison of self-reported coverage and administrative data records. Public Opinion Quarterly. 2008 Jan 1; 72(2): 241-59.

[9] Meyer BD, Goerge R. Errors in survey reporting and imputation and their effects on estimates of food stamp program participation. CES Working Paper #2011-14. Washington, D.C.: U.S. Census Bureau; 2011 Apr 1.

[10] Meyer BD, Mittag N. Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness and holes in the safety net. National Bureau of Economic Research; 2015 Oct 22.

[11] Brummet QO. Comparison of survey, federal, and commercial address quality. CARRA Working Paper #2014-06. Washington, D.C.: U.S. Census Bureau; 2014 Jun 30.

[12] Kingkade WW. 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, Montreal, Canada; 2013.

[13] Moore B. Preliminary research for replacing or supplementing the year built question on the American Community Survey with administrative records. Washington, D.C.: U.S. Census Bureau; 2015 Nov 24. Available from: https://www.census.gov/library/working-papers/2015/acs/2015_Moore_02.html.

[14] Bond B, Brown JD, Luque A, O'Hara A. The nature of the bias when studying only linkable person records: Evidence from the American Community Survey. CARRA Working Paper #2014-08. Washington, D.C.: U.S. Census Bureau; 2014 Apr 22.

[15] Cahalan D. Correlates of respondent accuracy in the Denver validity survey. Public Opinion Quarterly. 1968 Jan 1; 32(4): 607-21.