Evaluating a Commercial Data Source for Use to Estimate Property Taxes

Zachary H. Seeskin[1,2]
NORC at the University of Chicago, 55 E. Monroe Street, 31st Floor, Chicago, IL 60603

Abstract
While commercial data sources offer promise to statistical agencies for use in production of official statistics, challenges can arise in their use as the data are not collected for statistical purposes. This paper evaluates 2008-2010 property tax data from CoreLogic, Inc. (CoreLogic), aggregated from county and township governments from across the country, for use to improve 2010 American Community Survey (ACS) estimates of property tax amounts for single-family homes. Particularly, the research evaluates the potential to use CoreLogic to reduce respondent burden and measurement error by using CoreLogic data directly for property tax estimates in place of survey responses. The research found that the coverage of the CoreLogic data varies between counties as does the correspondence between ACS and CoreLogic property taxes. Further, large differences between CoreLogic and ACS property taxes in certain counties seem to be due to conceptual differences between what is collected in the two data sources. The research examines three counties, Clark County, NV, Philadelphia County, PA and St. Louis County, MO, and compares how estimates would change using CoreLogic data in place of ACS responses.

Key Words: Commercial data, data quality, respondent burden, measurement error

1. Introduction

The use of administrative records and commercial data for producing official statistics is growing at statistical agencies in the U.S. and internationally. These data sources can be inexpensive 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. The change toward increased use of administrative records and commercial data represents a shift for statistical agencies relying more on "found" data (data sources taken as is) in addition to surveys and censuses where statistical agencies design the collection of the data using scientific principles (Groves 2011, Japec et al. 2015). Thus, careful evaluations are needed before using administrative records or commercial data for statistical products.

[1] Address correspondence to Seeskin-Zachary@norc.org.
[2] This research was supported through a U.S. Census Bureau Dissertation Fellowship during the author's doctoral studies at Northwestern University through contract YA-1323-15-SE-0097. All views expressed are solely those of the author.

This paper evaluates 2008-2010 commercial property tax data available from CoreLogic, Inc. (CoreLogic) for improvement of survey estimates of property tax amounts from the 2010 American Community Survey (ACS). CoreLogic aggregates property tax records from counties and townships across the country into one dataset. While data sources like CoreLogic offer potential opportunity for statistical products, because the data are "found" data, statistical agencies must proceed with caution in evaluating such data sources for statistical use. I focus on single-family homes, where the record linkage is less challenging than for multi-unit structures.

There are two goals for the research. First, 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. A major concern for the ACS is the respondent burden from the survey length and content. Thus the research considers the possibility of using CoreLogic alone to construct property tax estimates for geographic areas across the U.S. In addition, as the data reflect information from property tax records, the research studies what can be learned about survey response error in the ACS using CoreLogic. Separate research investigates the usefulness of CoreLogic data to mitigate the effects of survey nonresponse (Seeskin 2016).

The research finds that the quality of the CoreLogic data varies between 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, likely due to conceptual differences between what is collected in the two sources. In these counties, the values reported on property tax records may not reflect the property taxes actually paid. Thus, using CoreLogic nationwide in place of asking about property taxes on the ACS is not advised. 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.

Examining Clark County, NV, Philadelphia County, PA and St. Louis County, MO, I compare estimates that use and do not use the CoreLogic data. In St. Louis County, MO, where there is evidence that CoreLogic data may be a "gold standard," mean county property tax estimates using ACS responses are 2 to 3 percent lower than estimates using data from CoreLogic records. This indicates the effect of ACS response error on the ACS estimate in St. Louis County if the CoreLogic data can indeed be viewed as a "gold standard." In examples of counties where CoreLogic data may be less trustworthy, using CoreLogic records instead of ACS responses yields estimates that are about 7 percent higher in Clark County and 8 percent lower in Philadelphia County. Thus, using CoreLogic data directly in these counties would lead to very different estimates of county property taxes.

Section 2 discusses ACS housing statistics as well as previous research on statistical uses of administrative records and commercial data. Then, Section 3 provides an overview of the CoreLogic property tax data file and investigates the quality of the data. Section 4 compares ACS and CoreLogic estimates of property taxes, and 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

2.1 American Community Survey Housing Statistics

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. The housing statistics collected by the ACS are important for a number of purposes. For example, understanding the costs involved with home ownership helps provide measures of housing affordability. ACS property tax estimates are used for formula block grant funds, for mass transportation and metropolitan planning, for determining eligibility for housing assistance, for policy evaluation and to inform efforts to plan affordable housing (Census Bureau 2014, Ruggles 2015).

There are some weaknesses in using survey responses for estimates. One concern with surveys is error in respondents' reports. This kind of error is often referred to as *response error* or *measurement error*, where the respondent misreports the information requested for the survey. Past research has found measurement error to be a concern when studying home value. Kiel and Zabel (1999) compare 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. They found that survey responses tend to be higher than selling prices and that the difference is greater for recent buyers than for homeowners with longer tenure. Benitez-Silva et al. (2008) 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. In addition, they found the difference to be greater when homeowners purchased their homes during an economic boom.

While there has been extensive research on measurement error for home values, measurement error for property taxes has been less well-studied. The nature of the measurement error may be different as a home's value requires some subjective judgment while property tax amount is an objective concept reflecting the amount that households are billed annually toward property taxes. Some evidence comes from Murphy (2013) in discussion of a content reinterview survey of the 2012 ACS. For this study, respondents from the 2012 ACS were contacted soon after the original interview and asked some of the same questions. Disagreement in responses between the two surveys indicates a reason to be concerned about the accuracy of survey responses. Examining property taxes as a categorical variable with thirteen categories, Murphy found an aggregate gross difference rate of 6.4 percent for annual property tax amount, interpreted as a moderate level of inconsistency. This evidence suggests some need for concern about response error for ACS property tax estimates. One possible reason for the response error discussed is that some respondents pay some or all of their property taxes as part of their mortgage payment. Thus, it may be difficult for these respondents to calculate their annual property taxes.

2.2 Uses of Administrative Records and Commercial Data for Official Statistics

One development in federal statistics at agencies nationally and internationally is the increased use of administrative records and commercial data for statistical purposes. Data can either be used directly, in place of conducting a census or survey, or indirectly, to assist with conducting a census or survey. In many cases, uses of administrative records and commercial data can help to mitigate the weaknesses of survey data. Johnson, Massey and O'Hara (2014) provide an overview of uses of these data in the U.S. Administrative records

can be used to assist the construction of survey frames, for respondent contact, in data collection and processing and for statistical modeling postcollection. The present review focuses on uses of administrative records and commercial data in data collection and processing specifically.

Some statistical agencies in other counties use administrative data registers as major parts of their statistical systems, including Denmark, Finland, Iceland, Norway and Sweden. Research in these countries has examined the strengths and weaknesses of administrative records for official statistics and has discussed possible data quality frameworks for assessing administrative records (Tønder 2008; Laitila, Wallgren and Wallgren 2011; Zhang 2012; Wallgren and Wallgren 2014). Administrative records can help to reduce cost, lower respondent burden and sometimes offer greater geographic and temporal detail. The challenges with using administrative records and third party data arise largely due to the fact that the data are not collected for statistical purposes. When using administrative records with survey responses, one must beware differences in concepts measured, population coverage and time of measurement as well as errors in record linkage.

This research in particular considers two ways in which the use of commercial data from CoreLogic could benefit estimates of ACS property taxes: to reduce respondent burden and to better assess ACS measurement error. The following discusses previous research on uses of administrative records and commercial data for these purposes. As will be seen, often both benefits are achieved from a single use of administrative records or commercial data. For example, removing a question from a survey interview and instead using administrative or commercial data to produce estimates may both reduce respondent burden and reduce measurement error.

2.3 Respondent Burden

One concern with surveys is the burden placed on respondents by the time and effort required to participate in the survey interview. For the ACS, this is a particular concern due to the length of the interview. The 2016 questionnaire includes 48 questions, many of which are multipart (Census Bureau 2016). Ruggles (2015) conducted a review of administrative record and commercial data sources that could be used in place of questions on the ACS. She proposed that if alternative data sources were of sufficient data quality, estimates for certain topics could be developed from the alternative data sources. In addition, the shorter length of the ACS interview could reduce respondent fatigue as well as response error to other questions on the ACS (Bradburn 1978). Using CoreLogic for property taxes and other housing topics was identified by Ruggles as a possible way of reducing respondent burden for the ACS.

In some other instances, statistical agencies have used alternative data sources to reduce respondent burden. For example, Donaldson and Streeter (2011) discuss how Geographic Information Systems can be used in place of survey questions on the American Housing Survey for estimates of distances of households from neighborhood amenities. The administrative registers of the Nordic countries mentioned previously are examples of large-scale efforts that have reduced respondent burden.

2.4 Response Error

Administrative records have also been useful to understand response error in estimates and in some cases to adjust estimates for response error. Much of the research in this area has pertained to program receipt. For example, the Census Bureau is using Social Security

Administration (SSA) data linked to the Survey of Income and Program Participation (SIPP) to correct responses about supplementary security income receipt and disability insurance receipt (Giefer et al. 2016). Medicaid records have been used to adjust Current Population Survey (CPS) estimates of Medicaid for underreporting (Davern et al. 2008). Other studies have examined linking the CPS with administrative records for food stamps, Temporary Assistance for Needy Families, Generalized Assistance and housing assistance to improve estimates of program receipt and of poverty (Meyer and Goerge 2011, Meyer and Mittag 2015). Another focus has been using administrative and commercial data in census and survey processes to correct move dates for respondents (Mulry, Nichols and Childs 2014). This study used the U.S. Postal Service's National Change of Address File to examine census error in reported move date, so that individuals are enumerated in the correct location based on where they actually lived on Census Day.

Some of the above mentioned research has assumed that the administrative records are a "gold standard," or that when linked values for a field are available from the administrative records that they reflect the true value. However, in some cases, there are good reasons to believe that both the data from administrative records and commercial data have error. This requires a more complex approach toward using administrative records and commercial data to study response error. Kapteyn and Ypma (2007) study the linkage of population censuses to longitudinal income registries in Sweden in developing improved estimates of earnings, pensions and taxes. They were concerned about incorrect linkages and thus do not view the registry data as a "gold standard." Their estimates account for theory regarding response error and linkage error. Abowd and Stinson (2013) extend Kapteyn and Ypma's work and provide a general framework for estimation from linked survey and administrative data when both sources have measurement error. Their approach involves placing Bayesian priors on the reliability of each data source and estimating the true value as a weighted average of all available measures. In addition, Herzog, Scheuren and Winkler (2007) provide an overview of methods to account for the uncertainty in record linkage in statistical estimation.

3. CoreLogic Data

3.1 Overview of CoreLogic Property Tax File

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: property value, tax amount, physical and structural characteristics, mortgage, sales and ownership information and geography. The fields available can differ between counties and townships.

Using the geographic and address information from CoreLogic records, the Census Bureau's Center for Administrative Records Research and Applications 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. Brummet (2014) documents the linkage procedure and some of the challenges in linking CoreLogic to the MAF. More than 18 percent of CoreLogic records are missing an address field (e.g., street name or zip code) needed to link the record to the MAF. 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, Brummet (2014) finds

that 79.0 percent of single-unit structures are successfully linked, compared with only 14.8 percent of multi-unit structures. Some of this difference is due to CoreLogic records reflecting the structure rather than the unit for multi-family structures.

I examine single-family, owner-occupied records from the ACS and CoreLogic, because only owner-occupied households are asked about their property taxes for the ACS. Thus, focusing on owner-occupied records allows the CoreLogic property tax values to be compared to the ACS property tax values. Nonetheless, future research could investigate the quality of CoreLogic information for renter-occupied units. Only single-family homes, both attached and detached, are studied due to the greater availability of linked CoreLogic records for single-family units than for multi-family structures.

Previous research conducted by Census Bureau researchers has studied using CoreLogic data for estimates of home values and year that a structure is built. Kingkade (2013) studies how CoreLogic and 2009 ACS home values compare for single-family homes and finds that ACS home values tend to be higher than the values from CoreLogic. 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. Moore (2015) evaluates the use of CoreLogic for the year that a structure is built in the 2012 ACS and finds that 56.7 percent of single-family, detached homes in the ACS can be linked to CoreLogic records with year built information available, with linkage rates varying across states. In the ACS, respondents report that that the year the structure was built falls within a certain range, often a decade. Using MAF linkage, Moore finds agreement for year built between ACS and CoreLogic for 78.3 percent of the linked records with reported year built information.

3.2 Comparing the CoreLogic and ACS Files

The present research focuses on the 2010 ACS single-year file after considering examining both the 2009 and 2010 files and finding a somewhat better correspondence between CoreLogic and 2010 ACS property taxes than for the 2009 ACS. 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.[3]

The availability of CoreLogic property tax information varies across states, counties and townships. The match rates for states are presented in Table 1 and for large counties in Table 2. Three counties that will be the focus of later analyses (Clark County, NV, Philadelphia County, PA and St. Louis County, MO) are shown in bold. 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. Among large counties, many have 90 percent or more of the 2010 ACS records studied linked to CoreLogic property tax information, while Miami-Dade County, FL and Shelby County, TN have no linked CoreLogic tax information available.

[3] In addition, large discrepancies were found between the CoreLogic and ACS information when the ACS reported the year the structure was built as 2009 or 2010. Due to these discrepancies, the research does not use CoreLogic linkages when the structure was built in 2009 or 2010.

Table 1: Match Rates with 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	6,575
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.

The availability of linked CoreLogic tax information also varies by household characteristics. Table 3 shows that 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 (Bond et al. 2014). 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 4, 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.

To understand which of these characteristics have the strongest association with availability of linked CoreLogic tax information and to adjust estimates for geographic variation, multivariate logistic regression models were estimated to model the probability

that a record has linked CoreLogic tax information available. Logistic regression is useful for modeling binary dependent variables as it models the log odds of the dependent variable as a linear function of the independent variables. If p is a record's probability of having CoreLogic tax information available, and X are independent variables, then logistic regressions estimate

$$\ln\left(\frac{p}{1-p}\right) = \beta' X, \tag{1}$$

where $\ln\left(\frac{p}{1-p}\right)$ is the log odds ratio and β are estimated coefficients for the independent variables.

Table 2: ACS Match Rates with CoreLogic Property Tax Information by County

County	Match Rate (%)	Number of Records	County	Match Rate (%)	Number of Records
Saint Louis Cty, MO	**95.6**	**4,274**	Los Angeles Cty, CA	88.1	19,912
Clark Cty, NV	**94.2**	**4,650**	Salt Lake Cty, UT	88.1	3,326
Sacramento Cty, CA	93.4	4,005	Allegheny Cty, PA	88.0	5,789
Orange Cty, FL	93.4	2,881	Mecklenburg Cty, NC	87.7	2,726
Dallas Cty, TX	93.1	5,927	Franklin Cty, OH	87.5	3,728
Wake Cty, NC	92.9	2,944	Milwaukee Cty, WI	87.4	2,798
Fairfax Cty, VA	92.0	3,442	Cook Cty, IL	84.0	12,007
Alameda Cty, CA	91.9	3,865	Oakland Cty, MI	83.0	5,342
Harris Cty, TX	91.6	9,812	Suffolk Cty, NY	82.6	5,704
Hillsborough Cty, FL	91.3	3,320	Nassau Cty, NY	82.1	5,394
Montgomery Cty, MD	91.3	3,351	Fulton Cty, GA	81.5	2,286
Contra Costa Cty, CA	91.1	3,074	Maricopa Cty, AZ	80.8	10,533
Pima Cty, AZ	90.4	2,972	Hennepin Cty, MN	79.9	4,544
Orange Cty, CA	90.2	7,937	Middlesex Cty, MA	79.6	4,364
Philadelphia Cty, PA	**90.2**	**3,815**	Palm Beach Cty, FL	77.2	3,855
Cuyahoga Cty, OH	90.2	5,243	Westchester Cty, NY	76.9	2,464
Santa Clara Cty, CA	90.1	4,699	King Cty, WA	71.6	5,970
Wayne Cty, MI	89.9	6,576	Broward Cty, FL	60.6	3,915
Riverside Cty, CA	89.8	5,721	Bronx Cty, NY	47.8	500
San Diego Cty, CA	89.1	7,299	Kings Cty, NY	43.3	1,620
Tarrant Cty, TX	88.7	5,473	Queens Cty, NY	42.8	2,786
Fresno Cty, CA	88.7	2,066	Honolulu Cty, HI	35.6	2,115
Travis Cty, TX	88.7	2,827	New York Cty, NY	15.7	89
San Bernardino Cty, CA	88.5	4,242	Miami-Dade Cty, FL	0.0	4,482
Bexar Cty, TX	88.4	4,805	Shelby Cty, TN	0.0	2,723

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

Logistic regression models were fit using iteratively reweighted least squares. Odds ratio estimates are presented for the independent variables in Table 5. These can be interpreted as the multiplicative effect of the independent variable on the odds ratio. Two models are presented, one with a set of indicator variables for counties, and one without. Estimates from the model without county indicators can be interpreted as overall effects across the

U.S., while estimates from the model with county indicators model represent the effects of characteristics within counties.

Table 3: 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
Urbanicity		
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 4: 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.

Overall, the odds ratio estimates from the models with and without county indicators are very similar, indicating that the association of the presented demographic characteristics with CoreLogic availability is similar whether investigating patterns within a county or across the country. The Nagelkerke R^2 (Nagelkerke 1991) increases from 0.115 in the model without county indicators to 0.171 in the model with county indicators, indicating that counties account for a modest amount of the variation in the availability of CoreLogic tax information across the country.

The urbanicity of households has a particularly strong association with availability of CoreLogic tax information. Adjusting for other variables, the odds of availability of CoreLogic homes in rural areas is one-third that of homes in urban areas. Socioeconomic characteristics are also associated with CoreLogic availability. Households in poverty have an odds of CoreLogic availability of about 20 percent less than that of households not in poverty. Holding all other variables constant, the odds of CoreLogic availability is about 2 percent higher with each $10,000 increase in household income and about 8 percent higher with each $1,000 increase in property taxes, although the odds decrease by 1 percent with each $10,000 increase in home value.

Table 5: Odds Ratio Estimates from Logistic Regression Models of Probability of ACS Record Having Linked CoreLogic Property Tax Information Available

Group	Model Without County Indicators	Model With County Indicators
No High School Diploma	**0.824**	**0.833**
[95% Confidence Interval]	*[0.811, 0.838]*	*[0.818, 0.847]*
High School Diploma or G.E.D.	**0.924**	**0.931**
[95% Confidence Interval]	*[0.913, 0.935]*	*[0.920, 0.942]*
Some College	**1.011**	**1.010**
[95% Confidence Interval]	*[0.999, 1.021]*	*[0.999, 1.022]*
In Poverty	**0.793**	**0.790**
[95% Confidence Interval]	*[0.779, 0.807]*	*[0.776, 0.805]*
Rural	**0.327**	**0.337**
[95% Confidence Interval]	*[0.324, 0.330]*	*[0.334, 0.340]*
Household Income ($10,000s)	**1.021**	**1.017**
[95% Confidence Interval]	*[1.014, 1.027]*	*[1.010, 1.023]*
Home Value ($10,000s)	**0.991**	**0.989**
[95% Confidence Interval]	*[0.989, 0.992]*	*[0.988, 0.991]*
Property Taxes Paid ($1,000s)	**1.078**	**1.078**
[95% Confidence Interval]	*[1.076, 1.080]*	*[1.076, 1.081]*
AIC	**1285500**	**1237351**
Nagelkerke R^2	**0.115**	**0.171**
Number of Records	**1,116,568**	**1,116,568**

Source: 2010 ACS single-family, owner-occupied households. Models also include householder race, householder age, year home built, year moved, number of bedrooms and home insurance amounts. AIC for Intercept Only model is 1380692. Survey weights not used for estimation.

3.3. 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 tax records provided by counties and townships to 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. A major difference in the distributions of ACS and CoreLogic property taxes is that ACS taxes are often reported as multiples of 500 or 1,000, while CoreLogic taxes are not. Histograms of the two distributions are presented in Figures 1 and 2. Other research has found that in some instances survey respondents tend to report round numbers for continuous variables (Pudney 2008, Manski and Molinari 2010). Aside from this bunching, the distributions overall appear to be similar.

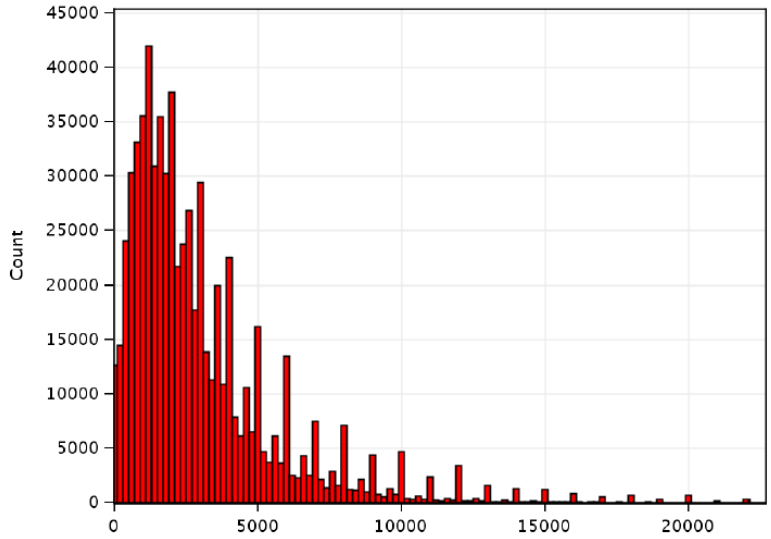

Figure 1: Histogram of 2010 ACS property taxes ($) from single-family, owner-occupied households linked to CoreLogic. 676,842 records.

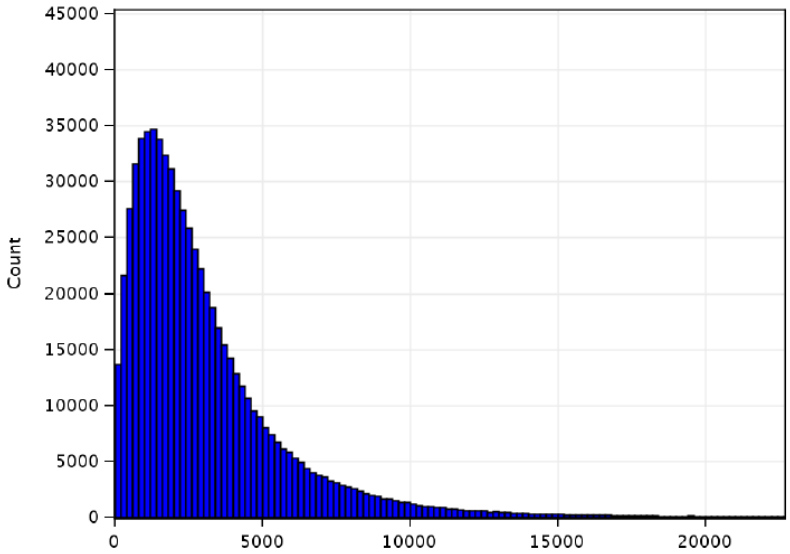

Figure 2: Histogram of CoreLogic property taxes ($) from records linked to 2010 ACS single-family, owner-occupied households. 676,842 records.

Since ACS and CoreLogic records are linked, considering the percentage difference between ACS and CoreLogic property taxes is useful. The percentage difference is defined 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 6 presents quantiles of the percentage difference for linked records by different household characteristics. Overall, the median percentage difference is 0.0 percent. The 5th and 95th percentiles and the interquartile range, the difference between the 75th and 25th 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 (Cahalan 1968). 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. Households in poverty have an interquartile range of 30.6 percent, while the interquartile range for households not in poverty is 18.1 percent.

Table 6: Distribution of Percentage Difference of ACS Property Taxes from CoreLogic Property Taxes by Household Characteristics

Household Characteristic	Percentiles for % Diff. of ACS from CoreLogic					Interquartile Range	Number of Records
	5th	25th	50th	75th	95th		
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.

Interestingly, the interquartile range does not vary as much by the year moved, indicating that survey recall of property taxes differs from patterns for home values found in research (Kiel and Zabel 1999). However, while the interquartile range is not as sensitive to the year moved, the 5th and 95th percentiles are somewhat sensitive. For households where the respondent has not moved since 1989 or earlier, the 5th percentile for the percentage difference is -67.5 percent and the 95th percentile is 101.9 percent, which are both greater in magnitude than the 5th (-58.2 percent) and 95th (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 is 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 boxplots of the percentage difference by state in Figure 3 and by large county in Figure 4. In addition,

Tables 7 and 8 provide the distributions of the percentage difference in addition to the correlation between ACS and CoreLogic records by state and large county respectively.

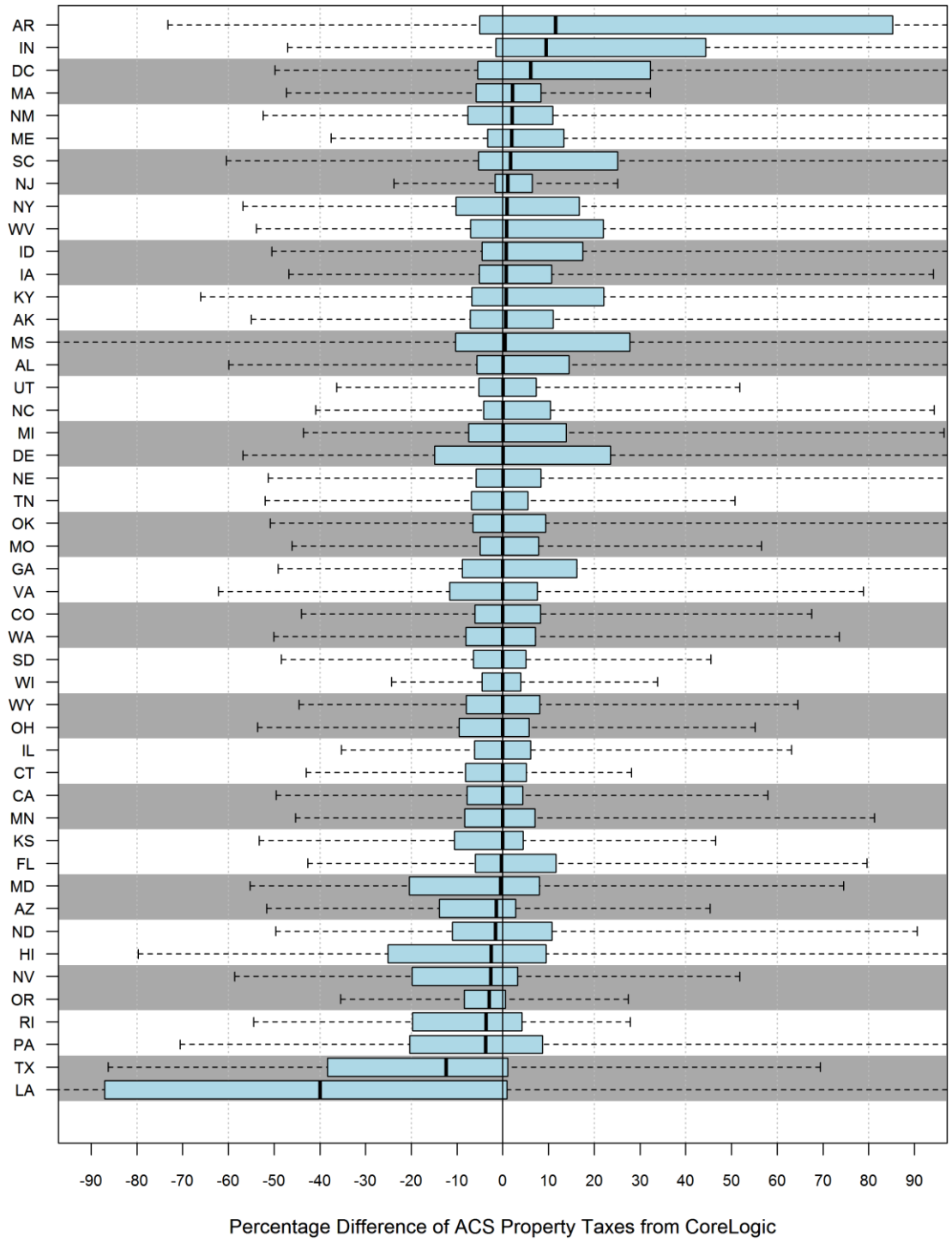

Figure 3: Boxplots of percentage difference of ACS property taxes from linked CoreLogic property taxes by state. 676,842 records. Whiskers indicate 5th and 95th percentiles.

Table 7: Distribution of Percentage Difference of ACS Property Taxes from CoreLogic Property Taxes by State

State	Percentiles for % Difference of ACS from CoreLogic					Interquartile Range	ACS-CoreLogic Correlation	Number of Records
	5th	25th	50th	75th	95th			
AR	-73.2	-5.0	11.6	85.3	844.6	90.3	0.76	6,099
IN	-47.0	-1.6	9.4	44.3	224.8	45.9	0.83	17,604
DC	-49.8	-5.5	6.1	32.2	179.2	37.7	0.84	631
MA	-47.3	-5.8	2.1	8.3	32.2	14.1	0.87	14,459
NM	-52.4	-7.6	2.1	10.9	110.3	18.5	0.69	3,484
ME	-37.5	-3.3	2.0	13.4	154.9	16.7	0.10	3,990
SC	-60.4	-5.3	1.7	25.1	198.0	30.4	0.75	9,212
NJ	-23.8	-1.7	1.1	6.4	25.1	8.1	0.90	23,343
NY	-56.8	-10.2	0.9	16.7	135.7	26.9	0.64	30,587
WV	-53.8	-7.0	0.8	22.0	196.4	29.0	0.52	2,847
ID	-50.5	-4.5	0.7	17.5	111.2	22.0	0.82	3,653
IA	-46.8	-5.1	0.7	10.6	94.2	15.8	0.86	10,003
KY	-66.0	-6.8	0.7	22.1	193.9	28.9	0.58	8,972
AK	-55.0	-7.1	0.7	11.0	103.8	18.1	0.77	1,079
MS	-100.0	-10.3	0.4	27.8	273.9	38.1	0.50	4,697
AL	-59.9	-5.7	0.1	14.5	132.2	20.1	0.86	8,990
UT	-36.3	-5.2	0.0	7.3	51.8	12.5	0.70	5,935
NC	-40.9	-4.2	0.0	10.4	94.3	14.6	0.83	20,116
MI	-43.6	-7.5	0.0	13.8	96.5	21.3	0.81	27,976
DE	-56.8	-14.9	0.0	23.6	153.2	38.5	0.75	2,414
NE	-51.3	-5.8	0.0	8.3	100.3	14.1	0.85	4,873
MO	-46.1	-4.9	0.0	7.9	56.5	12.8	0.90	13,390
OK	-50.9	-6.6	0.0	9.4	101.4	15.9	0.67	8,531
TN	-51.9	-6.9	0.0	5.5	50.8	12.3	0.83	355
GA	-49.1	-8.8	0.0	16.2	149.9	25.0	0.88	18,298
VA	-62.2	-11.6	0.0	7.6	78.8	19.2	0.84	17,840
CO	-44.0	-6.1	0.0	8.2	67.5	14.3	0.83	11,795
WA	-50.0	-8.1	0.0	7.1	73.6	15.2	0.85	15,456
SD	-48.4	-6.5	0.0	5.1	45.5	11.5	0.82	1,456
WI	-24.3	-4.5	0.0	4.0	33.9	8.5	0.91	27,202
WY	-44.5	-8.0	0.0	8.1	64.5	16.1	0.77	1,230
OH	-53.6	-9.5	0.0	5.7	55.2	15.2	0.85	35,705
IL	-35.3	-6.1	0.0	6.1	63.1	12.3	0.92	34,488
CT	-43.0	-8.2	0.0	5.1	28.2	13.3	0.85	9,328
CA	-49.5	-7.8	0.0	4.4	57.9	12.2	0.86	70,139
MN	-45.4	-8.3	0.0	7.0	81.3	15.4	0.89	23,711
KS	-53.3	-10.6	-0.1	4.5	46.5	15.1	0.87	1,069
FL	-42.6	-6.0	-0.3	11.7	79.6	17.6	0.85	31,008
MD	-55.2	-20.4	-0.5	8.0	74.5	28.5	0.77	15,023
AZ	-51.6	-13.9	-1.4	2.8	45.3	16.7	0.84	10,884
ND	-49.6	-11.0	-1.6	10.8	90.6	21.8	0.90	1,017
HI	-79.7	-25.1	-2.6	9.4	136.6	34.5	0.51	1,006
NV	-58.6	-19.8	-2.7	3.2	51.8	23.1	0.83	4,899
OR	-35.5	-8.4	-3.0	0.5	27.4	9.0	0.90	9,413
RI	-54.4	-19.7	-3.7	4.2	27.8	23.9	0.82	2,459
PA	-70.6	-20.4	-3.8	8.7	156.3	29.0	0.67	41,865
TX	-86.3	-38.4	-12.4	1.0	69.4	39.4	0.81	48,647
LA	-100.0	-87.0	-40.0	0.9	159.6	87.9	0.76	9,664
US	**-58.2**	**-10.1**	**0.0**	**8.5**	**89.9**	**18.6**	**0.72**	**676,842**

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

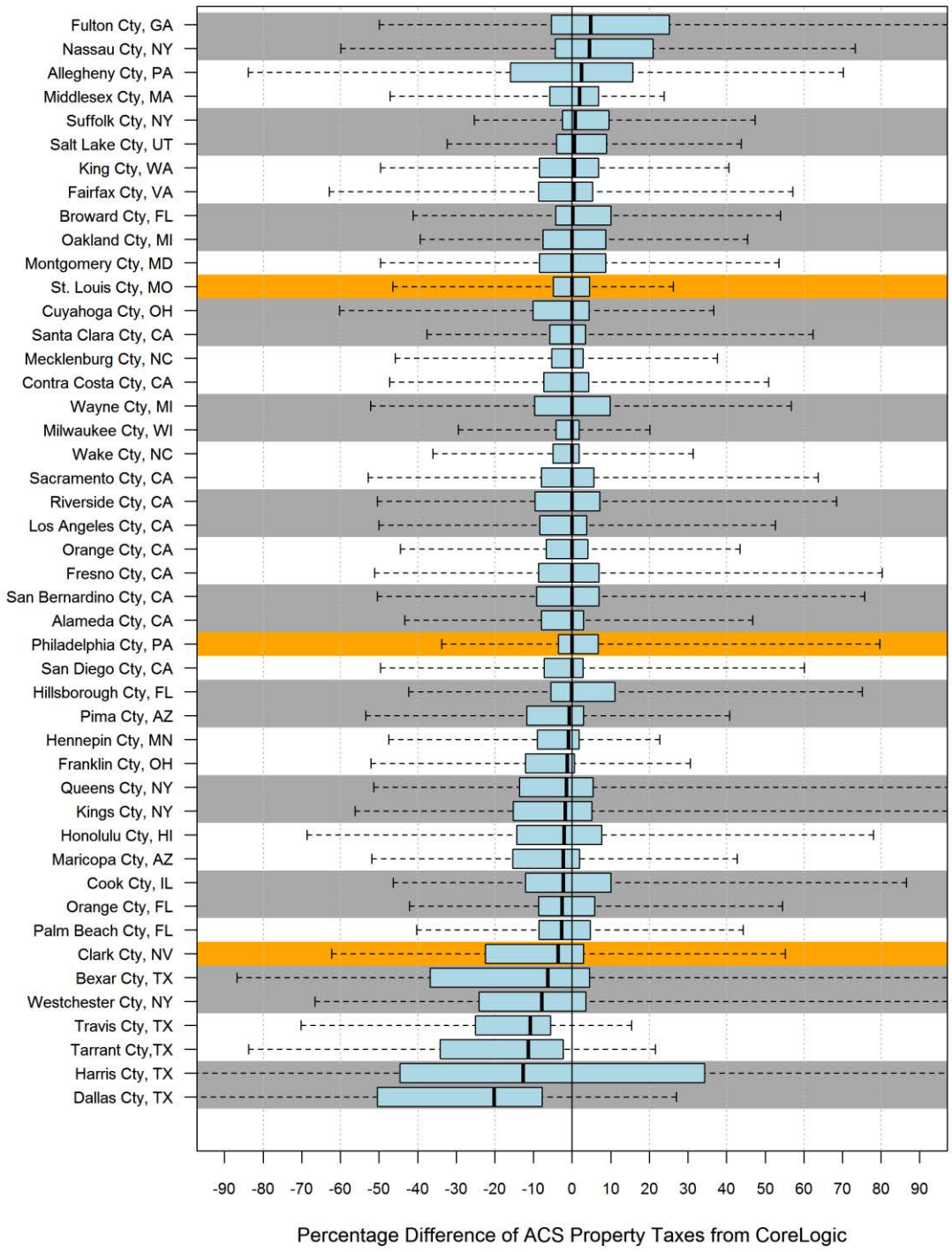

Figure 4: Boxplots of percentage difference of ACS property taxes from linked CoreLogic property taxes by select counties. Whiskers indicate 5th and 95th Percentiles.

Table 8: Distribution of Percentage Difference of ACS Property Taxes from CoreLogic Property Taxes by County

State	Percentiles for % Difference of ACS from CoreLogic					Interquartile Range	ACS-CoreLogic Correlation	Number of Records
	5th	25th	50th	75th	95th			
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, UT	-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,	-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, AZ	-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	4480

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

Across these geographic areas, the distribution of the percentage difference between ACS and CoreLogic property taxes can differ greatly. Many states and counties have a median percentage difference near 0.0 percent. However, there are some geographic areas with very different distributions for ACS and CoreLogic property taxes. In Arkansas and Indiana, ACS property taxes tend to be greater than those from CoreLogic with median percentage differences of 11.6 and 9.4 percent, respectively. In Texas and Louisiana, ACS property taxes tend to be less than CoreLogic property taxes with median percentage differences of -12.4 and -40.0 percent, respectively. 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 in different states. 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. When the spread of the percentage difference distribution for a county is much less than the distribution for the U.S., as for Milwaukee County and Wake County, it may provide a reason to have more confidence in CoreLogic data from those counties.

Further, the Pearson correlation between taxes for the ACS and CoreLogic records can help to assess the quality of CoreLogic information. Even when the two distributions differ, if the correlation is high, then tax information from either source may be useful in modeling the values for the other source. Twelve of the counties in Table 8 have correlations of 0.90 or greater, but there are also counties with low correlations including Westchester County, NY (0.44) and Honolulu County, HI (0.34).

3.4 Comparisons in Clark, Philadelphia and St. Louis Counties

This section analyzes three counties that are the focus of the remainder of this article: Clark County, NV, Philadelphia County, PA and St. Louis County, MO. All three counties have linked CoreLogic property tax information available for more than 90 percent of households in the ACS. While not arguing that the CoreLogic data are a "gold standard" for property tax amounts, some evidence was found to trust the St. Louis CoreLogic data. The St. Louis data include a tax code area field for every household. Information from St. Louis County indicates that this tax code area mostly determines a property's tax rate for owner-occupied households.[4] Further, this research found that the tax code area determined 98.9 percent of the variation in tax rates for St. Louis County CoreLogic records linked to the 2010 ACS.[5] This information combined with the smaller differences between ACS and CoreLogic taxes in St. Louis relative to other counties provides evidence that the St. Louis CoreLogic data are possibly a "gold standard." By comparing analyses for St. Louis to other counties, comparisons can be made between counties with different levels of quality of the CoreLogic data.

Table 8 shows that the correlation between ACS and CoreLogic somewhat differs between the three counties (St. Louis 0.92, Philadelphia 0.82, Clark 0.81) as do the distributions of percentage differences between ACS and CoreLogic. Among these three counties, the percentage differences of ACS from CoreLogic are greatest in Clark County. Notably, the

[4] <https://revenue.stlouisco.com/Collection/YourTaxRates.aspx>. Accessed April 17, 2016.

[5] The tax rate was calculated as the ratio of CoreLogic property taxes to the CoreLogic assessed property value.

median percentage difference in Clark County is -3.7 percent.The mean absolute percentage differences are presented for the three in Table 9. Clark County (27.0 percent) and Philadelphia County (26.1 percent) have much greater mean absolute percentage differences than does St. Louis County (13.6 percent).

Table 9: Mean Absolute Percentage Difference of ACS Property Taxes from CoreLogic Property Taxes by County

Clark Cty, NV	Philadelphia Cty, PA	St. Louis Cty, MO
27.0	26.1	13.6

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

4. Results

This section compares county estimates of mean property taxes using either ACS and CoreLogic information. I investigate the use of CoreLogic to address reponse error in estimates. Estimates are compared for Clark, Philadelphia and St. Louis counties.

When presenting the ACS estimates, I use the ACS's allocations for the nonrespondents, which are imputed using a hot deck approach (Stiller and Dalzell 1998). When CoreLogic estimates are presented, I substitute the ACS responses or allocations for the missing CoreLogic data. All estimates are estimated using the survey weights, and confidence intervals are estimated using the ACS's replicate weights with jackknife replication using the R *survey* package (Lumley 2010).

Results can be found in Table 10. There are large differences in estimates depending on whether the ACS responses or the CoreLogic records are primarily used to construct estimates. In St. Louis County, the estimate primarily using the CoreLogic records is 2.6 percent larger than the ACS estimate. Viewing the St. Louis CoreLogic data as a "gold standard" for household property taxes, this difference can be interpreted as the impact of response error on CoreLogic estimates. In other words, if respondents accurately reported their property taxes on the ACS, then the St. Louis County estimate would be 2.6 percent larger. In Clark County and Philadelphia County, which may not have "gold standard" property tax data in CoreLogic, there are even larger differences between the ACS and CoreLogic-based estimates. The estimates primarily based on CoreLogic records are 6.9 percent higher than the ACS estimate in Clark County and 8.3 percent lower in Philadelphia County.

Table 10: Estimates of Mean Property Tax Amounts ($) for Single-Family, Owner-Occupied Homes with Various Imputation Methods for Three Counties

Estimates (Standard Errors)	Clark Cty, NV	Philadelphia Cty, PA	St. Louis Cty, MO
ACS	**2160 (28)**	**1526 (28)**	**2788 (39)**
[95% Confidence Interval]	*[2105, 2216]*	*[1471, 1581]*	*[2711, 2864]*
CoreLogic – ACS Substitutions	**2309 (24)**	**1399 (21)**	**2860 (33)**
[95% Confidence Interval]	*[2262, 2356]*	*[1357, 1441]*	*[2794, 2925]*
Number of Records	4,650	3,815	4,274

Source: 2010 ACS single-family, owner-occupied households and linked CoreLogic records in three counties.

In addition, estimates are presented for mean property taxes by mortgage status in each of the three counties in Tables 11 and 12. Mostly, the same pattern emerges as for the overall

mean property tax estimates for the county. There are sometimes large differences between the ACS- and CoreLogic-based estimates. In fact, for Clark County, the direction of the comparison of mean property taxes by mortgage status changes. For the ACS-based estimates, mean property taxes for households without a mortgage are higher than for households with a mortgage, but for the CoreLogic estimates, they are either about equal or lower. However, this change in the difference between households with and without a mortgage is not statistically significant.

Table 11: Estimates of Mean Property Tax Amounts ($) for Single-Family, Owner-Occupied Homes with a Mortgage Using Different Imputation Methods for Three Counties

Estimates (Standard Errors)	Clark Cty, NV	Philadelphia Cty, PA	St. Louis Cty, MO
ACS	**2156 (31)**	**1529 (26)**	**2774 (49)**
[95% Confidence Interval]	*[2094, 2217]*	*[1477, 1582]*	*[2676, 2872]*
CoreLogic	**2318 (29)**	**1456 (22)**	**2859 (43)**
[95% Confidence Interval]	*[2261, 2375]*	*[1412, 1500]*	*[2774, 2943]*
Number of Records	3,709	2,201	2,945

Source: 2010 ACS single-family, owner-occupied households with a mortgage and linked CoreLogic records in three counties.

Table 12: Estimates of Mean Property Tax Amounts ($) for Single-Family, Owner-Occupied Homes Not Mortgaged Using Different Imputation Methods for Three Counties

Estimates (Standard Errors)	Clark Cty, NV	Philadelphia Cty, PA	St. Louis Cty, MO
ACS	**2182 (79)**	**1344 (41)**	**2822 (69)**
[95% Confidence Interval]	*[2024, 2340]*	*[1263, 1424]*	*[2684, 2960]*
CoreLogic	**2269 (72)**	**1316 (41)**	**2862 (72)**
[95% Confidence Interval]	*[2125, 2414]*	*[1235, 1396]*	*[2717, 3006]*
Number of Records	941	1,614	1,329

Source: 2010 ACS single-family, owner-occupied households not mortgaged and linked CoreLogic records in three counties.

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 vary across geographic areas and are 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 Harris County, TX and Fulton County, GA, large differences between the CoreLogic and ACS property tax amounts indicate that the CoreLogic data reflect a different concept than that measured by the ACS. Even in Clark County, NV and Philadelphia County, PA, where the distribution of the percentage difference between ACS and CoreLogic taxes appears reasonable, using CoreLogic data instead of ACS data would lead to large changes in estimates of mean property taxes. In these two counties, it seems that CoreLogic is not a "gold standard" for all records throughout the county.

On the other hand, CoreLogic is possibly a "gold standard" in St. Louis County, MO. Using CoreLogic data instead of ACS data increases mean property tax estimates by about 2.6 percent, indicating that response error has a substantial effect on the estimates. 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, including to verify if St. Louis County's data is a "gold standard." 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.

There are some limitations of the methods of this research and the research's implications for using CoreLogic. First, the research focused on single-family homes and does not consider other kinds of structures. Previous research has 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, but 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, and "found" data more generally, 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 local authorities throughout the country. As new approaches toward federal statistical products are considered in the future, careful evaluations of "found" data will continue to be needed.

Acknowledgments

There are many individuals who I want to thank for their support and advice for this research. The responsibility for any and all errors is my own. I thank Tommy Wright and Amy O'Hara for making this research possible and for connecting me with researchers at the Census Bureau. I thank my doctoral advisor, Bruce Spencer, and my dissertation committee members, Larry Hedges and Chuck Manski, for their incredible guidance and mentorship. My 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, Howard Hogan, Andrew Keller, Ward Kingkade, Arend Kuyper, Tom Louis, Chris Mazur, 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 (RDC). I thank Trent Alexander, Quentin Brummet, Frank Limehouse, Joey Morales, Danielle Sandler and others for their assistance at the RDC.

References

Abowd, J. M. & Stinson, M. H. (2013). Estimating measurement error in annual job earnings: a comparison of survey and administrative data. *Review of Economics and Statistics*, *95*(5), 1451–1467.

Benitez-Silva, H., Eren, S., Heiland, F. & Jimenez-Martin, S. (2008). How well do individuals predict the selling prices of their homes? Working Paper, Levy Economics Institute, No. 571.

Bond, B., Brown, J. D., Luque, A. & O'Hara, A. (2014). 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.

Bradburn, N. (1978). Respondent burden. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 35-40. Alexandria, VA: American Statistical Association.

Brummet, Q. O. (2014). Comparison of survey, federal, and commercial address quality. CARRA Working Paper #2014-06. Washington, D.C.: U.S. Census Bureau.

Cahalan, D. (1968). Correlates of respondent accuracy in the Denver validity survey. *Public Opinion Quarterly*, *32*(4), 607–621.

Census Bureau (2014). ACS questions and current federal uses. <https://www.census.gov/programs-surveys/acs/operations-and-administration/2014-content-review/federal-uses.html>. Accessed May 3, 2016.

Census Bureau (2016). 2016 ACS form & instructions. <https://www.census.gov/programs-surveys/acs/about/forms-and-instructions/2016-form.html>. Accessed April 11, 2016.

Davern, M., Call, K. T., Ziegenfuss, J., Davidson, G., Beebe, T. J. & Blewett, L. (2008). Validating health insurance coverage survey estimates: A comparison of self-reported coverage and administrative data records. *Public Opinion Quarterly*, *72*(2), 241–259.

Donaldson, K. & Streeter, M. (2011). Measured versus reported distances in the American Housing Survey. SEHSD Working Paper #2011-30. Washington, D.C.: U.S. Census Bureau.

Giefer, K., Williams, A., Benedetto, G. & Motro, J. (2016). Program confusion in the 2014 SIPP: Using administrative records to correct false positive SSI reports. Proceedings of the 2015 Federal Committee on Statistical Methodology Research Conference.

Groves, R. M. (2011). Three eras of survey research. *Public Opinion Quarterly*, *75*(5), 861-871.

Herzog, T. N., Scheuren, F. J. & Winkler, W. E. (2007). *Data quality and record linkage techniques*. Springer Science & Business Media.

Japec, L., Kreuter, F., Berg, M., Biemer, P., Decker, P., Lampe, C., Lane, J., O'Neil, C. & Usher, A. (2015). Big data in survey research. AAPOR Task Force. *Public Opinion Quarterly*, *79*(4), 839–880.

Johnson, D. S., Massey, C. & O'Hara, A. (2014). The opportunities and challenges of using administrative data linkages to evaluate mobility. *The ANNALS of the American Academy of Political and Social Science*, *657*(1), 247–264.

Kapteyn, A. & Ypma, J. Y. (2007). Measurement error and misclassification: A comparison of survey and administrative data. *Journal of Labor Economics*, *25*(3), 513–551.

Kiel, K. A. & Zabel, J. E. (1999). The accuracy of owner-provided house values: The 1978–1991 American Housing Survey. *Real Estate Economics*, *27*(2), 263–298.

Kingkade, W. 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, Montreal, Canada.

Laitila, T., Wallgren, A. & Wallgren, B. (2011). Quality assessment of administrative data. *Research and Development–methodology Reports from Statistics Sweden*, *2*, 2011.

Lumley, T. (2010). Missing data. *Complex Surveys: A Guide to Analysis Using R*, 185–201.

Manski, C. F. & Molinari, F. (2010). Rounding probabilistic expectations in surveys. *Journal of Business and Economic Statistics, 28(2),* 219-231.

Meyer, B. D. & Goerge, R. (2011). 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.

Meyer, B. D. & Mittag, N. (2015). Using linked survey and administrative data to better measure income: Implications for poverty, program effectiveness and holes in the safety net. NBER Working Paper No. 21676. Cambridge, MA.

Moore, B. (2015). 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. <https://www.census.gov/library/working-papers/2015/acs/2015_Moore_02.html>. Accessed May 3, 2016.

Mulry, M. H., Nichols, E. M. & Childs, J. H. (2016). A case study of error in survey reports of move month using the U.S. Postal Service change of address records. *Survey Methods: Insights from the Field*. <http://surveyinsights.org/?p=7944>. Accessed September 23, 2016.

Murphy, P. (2013). American Community Survey 2012 Content Reinterview Survey. Washington, D.C.: U.S. Census Bureau. <http://www.census.gov/content/dam/Census/library/working-papers/2014/acs/2014_Murphy_01.pdf>. Accessed May 3, 2016.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika*, *78*(3), 691–692.

Pudney, S. (2008). *Heaping and leaping: Survey response behaviour and the dynamics of self-reported consumption expenditure*. Institute for Social and Economic Research Working Paper #2008-09. University of Essex, U.K.

Ruggles, P. (2015). Review of administrative data sources relevant to the American Community Survey. <https://www.census.gov/content/dam/Census/library/working-papers/2015/acs/2015_Ruggles_01.pdf>. Accessed May 3, 2016.

Seeskin, Z. H. (2016). Evaluating the use of commercial to improve survey estimates of property taxes. CARRA Working Paper #2016-06. Washington, D.C.: U.S. Census Bureau.

Stiller, J. & Dalzell, D. R. (1998). Hot-deck imputation with SAS arrays and macros for large surveys. Proceedings of SAS Community SUGI 23, Nashville, TN.

Tønder, J.-K. (2008). The register-based statistical system. Paper presented at the IAOS Conference "Reshaping Official Statistics."

Wallgren, A. & Wallgren, B. (2014). *Register-based statistics: Statistical methods for administrative data*. New York: John Wiley and Sons.

Zhang, L.-C. (2012). Topics of statistical theory for register-based statistics and data integration. *Statistica Neerlandica*, *66*(1), 41–63.