CARRA Working Paper Series

Working Paper #2016-06

Evaluating the Use of Commercial Data to Improve Survey Estimates of Property Taxes

Zachary H. Seeskin

NORC at the University of Chicago

Paper Issued: August 2016

Address correspondence to Zachary H. Seeskin, NORC at the University of Chicago, 55 E. Monroe St. # 3000, Chicago, IL 60603. Email: Seeskin-Zachary@norc.org

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.

This paper, which has been subject to a limited review, is released to inform interested parties of research and to encourage discussion. The views expressed are those of the author and not necessarily those of the U.S. Census Bureau.

Evaluating the Use of Commercial Data to Improve Survey Estimates of Property Taxes

Zachary H. Seeskin, NORC at the University of Chicago

While commercial data sources offer promise to statistical agencies for use in production of official statistics, challenges can arise as the data are not collected for statistical purposes. This paper evaluates the use of 2008-2010 property tax data from CoreLogic, Inc. (CoreLogic), aggregated from county and township governments from around the country, 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, to study survey response error and to improve adjustments for survey nonresponse. The research found that the coverage of the CoreLogic data varies between counties as does the correspondence between ACS and CoreLogic property taxes. This geographic variation implies that different approaches toward using CoreLogic are needed in different areas of the country. 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 with different approaches using the CoreLogic data. Mean county property tax estimates are highly sensitive to whether ACS or CoreLogic data are used to construct estimates. Using CoreLogic data in imputation modeling for nonresponse adjustment of ACS estimates modestly improves the predictive power of imputation models, although estimates of county property taxes and property taxes by mortgage status are not very sensitive to the imputation method.

Key words: Commercial Data, Data Quality, Respondent Burden, Measurement Error, Nonresponse

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, Amy O’Hara, Michaela Patton, 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, Tommy Wright 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, Amy O’Hara, Danielle Sandler and others for their assistance at the RDC.

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, i.e., 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.

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 around 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 three 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 around 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. Finally, even when the commercial data are not a "gold standard," the data can be valuable for imputation and nonresponse adjustment if the commercial data improve the predictive power of imputation models. Therefore, I compare different methods either using or not using the CoreLogic data for imputation modeling.

The research finds that the quality of the CoreLogic data varies between counties and townships around 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 from different methods of using or not using the CoreLogic data, either directly or for imputation modeling. 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 6 to 7 percent higher in Clark County and 6 to 11 percent lower in Philadelphia County. Thus, using CoreLogic data directly in these counties would lead to very different estimates of county property taxes.

Another aspect of the investigation studies using CoreLogic data in modeling imputations. The research compares imputation methods of linear regression, predictive mean matching and recursive partitioning to the ACS's hot deck imputations. Models are compared that either include or do not include CoreLogic covariates. In all three counties, using CoreLogic data in imputation modeling increases R^2 of the linear regression models for imputation by about 0.04 to 0.09. While these increases are modest, they indicate that using the CoreLogic data improves the predictive power of imputation models and is therefore valuable for imputation modeling. However, the estimates of mean county property taxes and property taxes by mortgage status are not very sensitive to the imputation method or whether or not CoreLogic data are used. Still, evidence suggests that using CoreLogic for imputation modeling may be valuable for other estimates constructed with ACS property tax responses, as using predictive mean matching or recursive partitioning imputation with CoreLogic data decreases the average percentage differences between imputations and the CoreLogic values relative to the ACS's hot deck allocations. Further evaluations could examine the impact of different imputation approaches on other estimates of interest. In addition, a method of using CoreLogic to construct partially identified interval estimates of mean property taxes is presented. This approach relaxes the assumption of most imputation models that the survey responses are missing at random.

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. The methods for the various property tax estimates are presented in Section 4, followed by the results in Section 5. Section 6 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 2014a, Ruggles 2015).

There are some weaknesses in using survey responses for the estimates. One issue is that of nonresponse. Some respondents selected for the ACS sample either do not respond to the ACS (unit nonresponse) or do not complete all questions from the survey questionnaire or interview (item nonresponse). The ACS uses imputation models, logical edits and weighting to adjust estimates for both types of nonresponse. When the assumptions of these approaches are incorrect, ACS estimates can be biased. For the ACS, unit nonresponse is not as much of a concern as response to the ACS is mandated by law.[1] Item nonresponse is a concern for some questions on the ACS. The ACS distinguishes between two kinds of missing data for items: *assignments* where the missing data can be inferred logically from other survey responses and *allocations* where the missing data cannot be inferred logically and an imputation is used for that respondent.

Table 1 presents information on the tax amount allocation rates by household characteristics for single-family, owner-occupied households.[2] The overall item allocation rate is 13.0 percent. Households in poverty and with householders with less education are less likely to respond to the ACS property tax question. In addition, the ACS median household income is $65,699 for respondents to the ACS property tax question and $52,143 for nonrespondents. This evidence indicates that response to the property tax question is strongly associated with respondent socioeconomic status and education. Any method to adjust estimates for nonresponse must consider these patterns.

Table 1. Allocation Rates for ACS Property Tax Question by Household Characteristics

Group	**Nonresponse Rate (%)**	**Number of Records**
Overall	**13.0**	**1,116,568**
Race of Householder		
White	12.0	921,548
Black	21.9	71,403
Hispanic	19.8	44,819
Asian	11.2	34,167
Other Race	15.9	44,631
Poverty Status		
In Poverty	21.9	65,328
Not in Poverty	12.5	1,051,240
Education Level of Householder		
No High School Diploma	20.1	99,846
High School Diploma or G.E.D.	15.8	292,946
Some College	12.8	334,973
College Graduate	9.3	389,100

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

[1] The estimated unit nonresponse rate for the 2010 ACS was 2.5% (Census Bureau 2016b).

[2] The numbers presented Sections 2, 3 and 4 are not adjusted for the survey weights unless noted.

Another 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. The research 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 reason 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 with 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 three ways in which the use of commercial data from CoreLogic could benefit estimates of ACS property taxes: to reduce respondent burden, to better assess ACS measurement error and to improve adjustments for survey nonresponse. The following discusses previous research on uses of administrative records and commercial data for these three purposes. As will be seen, sometimes multiple 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 2016a). Ruggles (2015) conducted a review of administrative records and commercial data sources that could be used in place of questions on the ACS. If alternative data sources were of sufficient data quality, estimates for certain topics could be developed from the alternative data sources. The shorter length of the ACS interview could reduce respondent fatigue and reduce 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.

2.5.Nonresponse

Administrative records and commercial data can also be valuable in adjusting estimates for survey nonresponse. Many surveys use imputation for nonrespondents in producing estimates for population characteristics. Imputations are filled-in data values for the survey records with missing information. After making the imputations, the data are treated as complete, and traditional survey techniques are used to develop estimates for the relevant population. Two approaches to using administrative records and commercial data for imputation include using the data directly for the imputations and for modeling the imputations. In the context of using these data in modeling for nonresponse adjustment, administrative records and commercial data are often referred to as auxiliary data.

Different methods of using administrative data for imputation are discussed and contrasted in Zanutto and Zaslavsky (2002), National Research Council (2009) and National Research Council (2013). Using direct substitution requires that the data used for imputation is accurate and measures the same concept as the survey does. Using administrative records in modeling depends on the validity of the model used for imputation. An advantage of modeling is that it does not require the administrative records to measure the concept of interest accurately as long administrative records improve the goodness of fit of the imputation models. Zanutto and Zaslavsky argue, "Where the administrative data can be regarded as an imperfect measure of the values elicited on the survey, [using administrative records as covariates] corrects for systematic differences between the two data systems."

There are several examples of research on the use of administrative records for imputation and nonresponse adjustment. The Census Bureau is considering the use of administrative records and commercial data to impute for nonresponding households for the 2020 Census (Mule and Keller 2014; Keller 2016; Morris, Keller and Clark 2016). This research typically studied the direct use of administrative and commercial data for imputations. Bee, Gathright and Meyer (2015) use IRS 1040 records for direct imputation to correct for bias due to unit nonresponse in CPS estimates of income, self-employment status, marital status, number of children and social security receipt. For most variables, they only find small changes to estimates by using the IRS 1040 records. Benedetto, Motro and Stinson (2016) have found administrative data valuable in modeling imputations for program receipt for SIPP, including receipt of food stamps and disability. They use a sequential regression multiple imputation approach.

The literature has found mixed results in determining whether commercial data sources are valuable for imputation and nonresponse adjustment. Peytchev and Raghunathan (2013) use

Experian commercial purchase data in imputation for survey estimates of health and tobacco use. While the match rates of the Experian data to the survey were only 52-54 percent for demographic variables and 30 percent for tobacco product variables, they found that using Experian data substantially changed estimates of self-ratings of health. Estimates for tobacco use did not change as much. West and Little (2013) use commercial data with socioeconomic variables for imputation in survey estimates of household income and housing unit area. They account for measurement error for variables in both datasets and find that the approach increases estimates for both variables, suggesting downward bias from using imputation approaches that do not use the commercial data in modeling. West et al. (2015) study the use of marketing demographic data, voter files and credit bureau data for nonresponse propensity models to adjust National Survey of Family Growth estimates of marital status, number of children and pregnancy. They find concerns about the quality and accuracy of the commercial data due to both high missing data rates and lack of agreement with the survey data. They also find that using the commercial data for nonresponse adjustment did not change estimates substantially.

Most of the research models nonresponse by assuming that the data are *missing at random,* or that the distributions of the observed and missing responses are the same conditional on covariates from either the survey alone or from both the survey and auxiliary data sources together. Another possible approach presented in Manski (2007) is to obtain partially identified interval estimates that relax the assumption that the data are missing at random. This approach recognizes that a range of values are possible for the missing data. Instead of imposing strong assumptions and obtaining *point-identified* estimates of a single value for the population mean of a variable, an interval estimate called a *partially identified* estimate can be obtained for the population characteristic.[3] Partial identification in the context of survey nonresponse is discussed by Manski (2015) and Manski (2016). Hokayem, Bollinger and Ziliak (2014) determine a partially identified Current Population Survey estimate of the poverty rate due to survey nonresponse.

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 township 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 CoreLogic records 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

[3] This interval estimate, representing uncertainty about the missing ACS responses, should not be confused with a confidence interval that estimates the uncertainty in the estimate of the population parameter due to random sampling. Estimates presented in Section 5 demonstrate partially identified interval estimates with confidence intervals.

multi-unit structures. Some of this difference is due to CoreLogic records reflecting the structure rather than the unit for multi-family structures. Also, while recognizing the challenges with record linkage, the estimates presented in this research are not adjusted for linkage uncertainty (Herzog, Scheuren and Winkler 2007).[4]

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 year built information using MAF linkage, 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

[4] In some instances, multiple CoreLogic records are available and linked to a single ACS record. For the purposes of this research, only one CoreLogic record is studied for each ACS record. When multiple CoreLogic records are available, one record is chosen based upon completeness of the record, recency, additional geographic information and similarity of the physical characteristics to those of the ACS record.

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.[5]

The availability of CoreLogic property tax information varies across states, counties and townships. The match rates for states are presented in Table 2 and for large counties in

Table 3. 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

[5] 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.

records studied linked to CoreLogic property tax information, while Miami-Dade County, FL and Shelby County, TN have no linked CoreLogic tax information available.

Table 2. ACS 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.

The availability of linked CoreLogic tax information also varies by household characteristics. Table 4 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 5, which compares characteristics for ACS records with and without linked CoreLogic property tax information, I find that 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. A good model of the propensity of a record to have a CoreLogic match could be also be useful in developing inverse probability weighted estimates of household characteristics using CoreLogic variables alone (Seaman and White 2013). 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 3. 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.

Table 4. 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 5. 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.

Logistic regression models were fit using iteratively reweighted least squares. Odds ratio estimates are presented for the independent variables in Table 6. 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. In addition to the variables presented, the model also includes householder race, householder age, year home built, year moved, number of bedrooms and home insurance amount as independent variables.

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 some 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 32% to 34 percent 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 79 percent 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 6. 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 CoreLogic and ACS 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. The intraclass correlation rounds to 1.000, indicating that the vast majority of the variation in property tax reports is between respondents rather than between ACS and CoreLogic reported values within a respondent.

The overall distributions of ACS and CoreLogic property taxes across the U.S. are similar between the two sources. Quantiles of the distributions are presented in Table 7. At the median, ACS property taxes are $2,200 compared with $2,302 for CoreLogic. At the 5th and 95th percentiles, the ACS distribution is slightly more extreme than the CoreLogic distribution. ACS property taxes are $350 at the 5th percentile, compared with $388 for CoreLogic, and $8,890 at the 95th percentile, compared with $8,707. The similarity between the ACS and CoreLogic property tax distributions contrasts with the differences between the two distributions for home values. As documented in Kingkade (2013), ACS home values are much greater than CoreLogic home values in part due to errors in respondent reporting of home value, and in part due to local property tax authorities determining an assessed value for taxation purposes rather than a market value.

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.

Table 7. Distribution of Corresponding ACS and CoreLogic Variables when Linked CoreLogic Property Tax Information is Available

Variable	Distribution 5th	 25th	 50th	 75th	 95th	Number of Records
Property Taxes						
ACS	350	1,200	2,200	4,000	8,890	676,842
CoreLogic	388	1,236	2,302	3,978	8,707	
Home Value						
ACS	60,000	120,000	200,000	320,000	700,000	641,672
CoreLogic	17,900	76,466	142,534	246,072	538,985	

Source: 2010 ACS single-family, owner-occupied households linked to 2008-2010 CoreLogic 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 8 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 has 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.

Figure 1. Histogram of 2010 ACS Reported Property Taxes ($) from Records Linked to CoreLogic

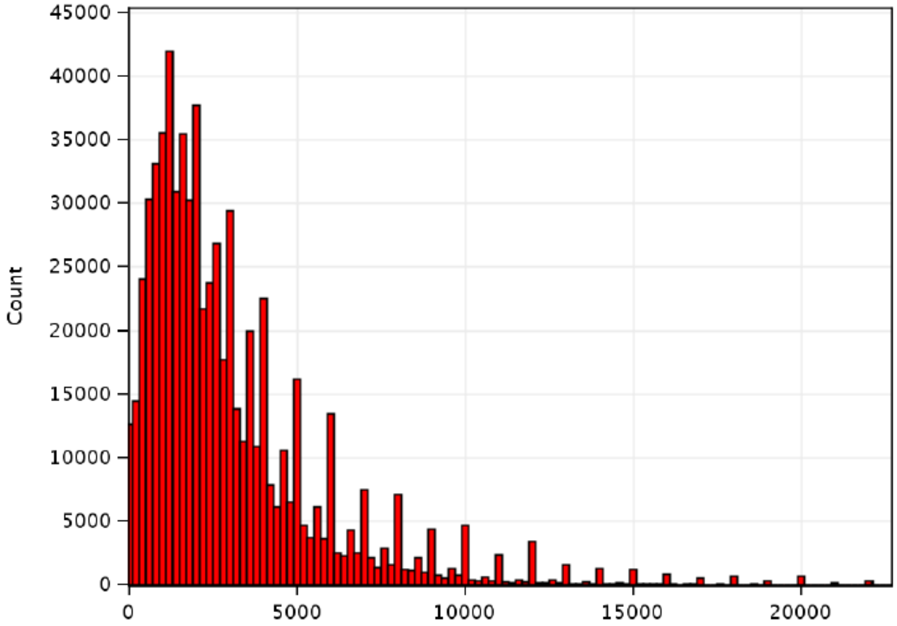
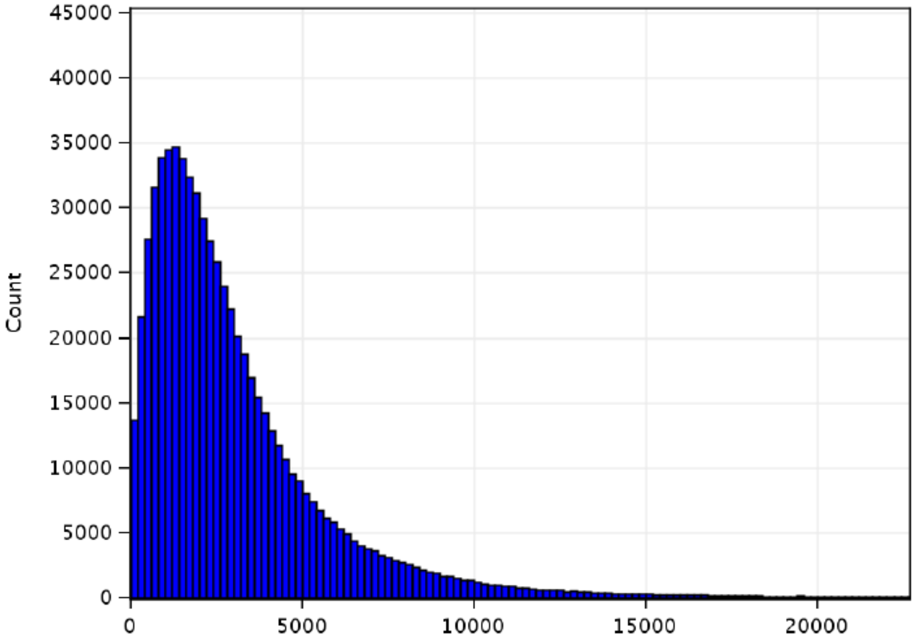


Figure 2. Histogram of CoreLogic Property Taxes ($) from Records Linked to 2010 ACS



Source: 2010 ACS single-family, owner-occupied households linked to 2008-2010 CoreLogic records. 676,842 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, Benitez-Silva et al. 2008, Kingkade 2013). 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, 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) percentile of the percentage difference for households overall.

Table 8. Distribution of Percentage Difference of ACS Property Taxes from CoreLogic Property Taxes by Household Characteristics

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

While comparisons by household characteristics may reflect patterns in ACS response error, comparing ACS and CoreLogic property taxes 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 the distribution of the percentage difference by state in Table 9 and by large county in Table 10. These tables also provide the correlation between ACS and CoreLogic records by geographic area. Boxplots of the percentage difference are presented by state in Figure 3 and by county in Figure 4 for large counties, with the whiskers representing the 5th and 95th percentiles.

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.

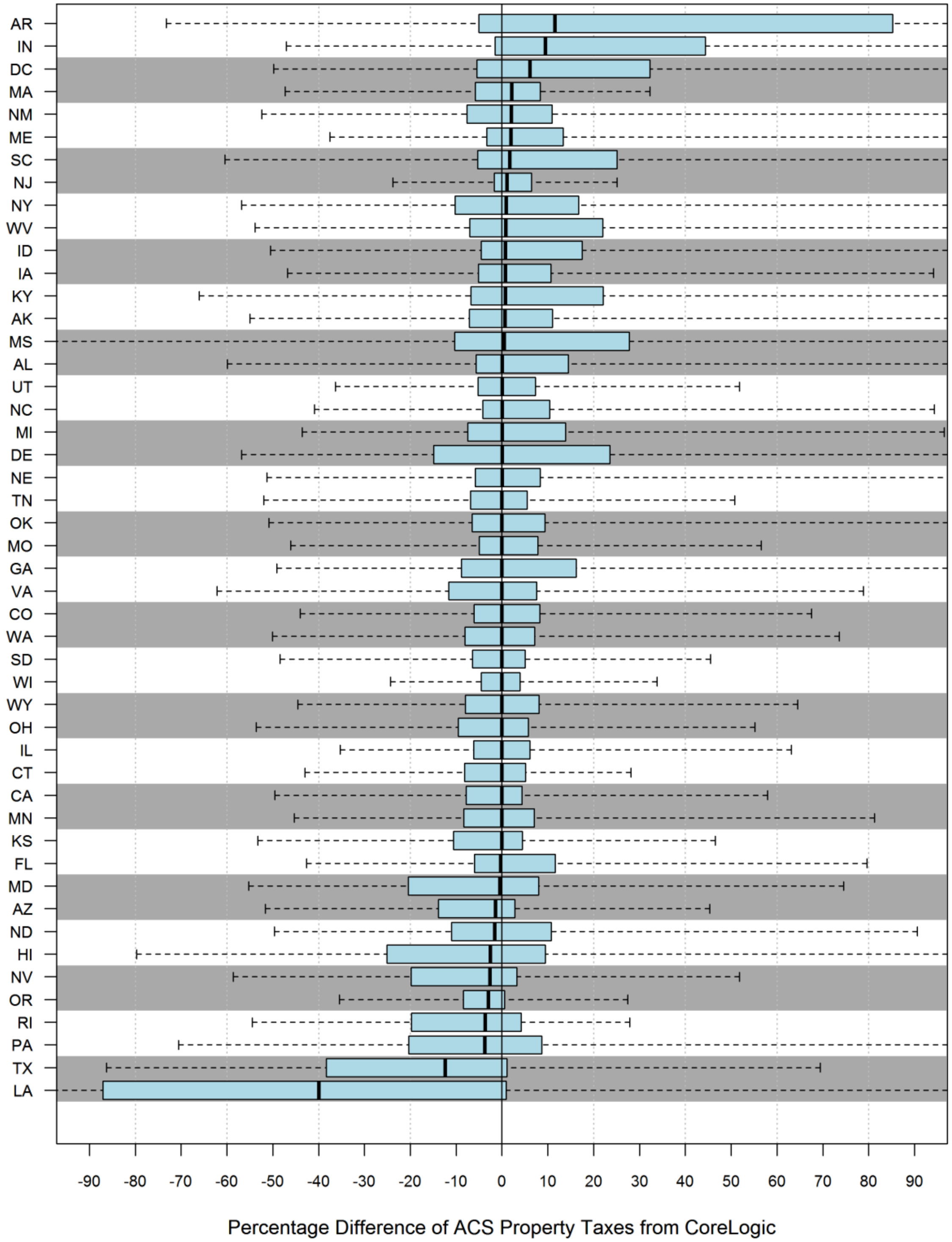
Examining the interquartile range as a measure of spread of the percentage difference can help with assessing the accuracy of CoreLogic property taxes in different states. Among the smallest interquartile ranges are those of Milwaukee County, WI (5.9 percent) and Wake County, NC (6.8 percent). On the 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.

Table 9. Distribution of Percentage Difference of ACS Property Taxes from CoreLogic Property Taxes by State

	Percentiles for % Difference of ACS from CoreLogic					Interquartile	ACS-CoreLogic	Number of
State	**5th**	**25th**	**50th**	**75th**	**95th**	**Range**	**Correlation**	**Records**
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 3. Boxplots of Percentage Difference of ACS Property Taxes from CoreLogic Property Taxes by State



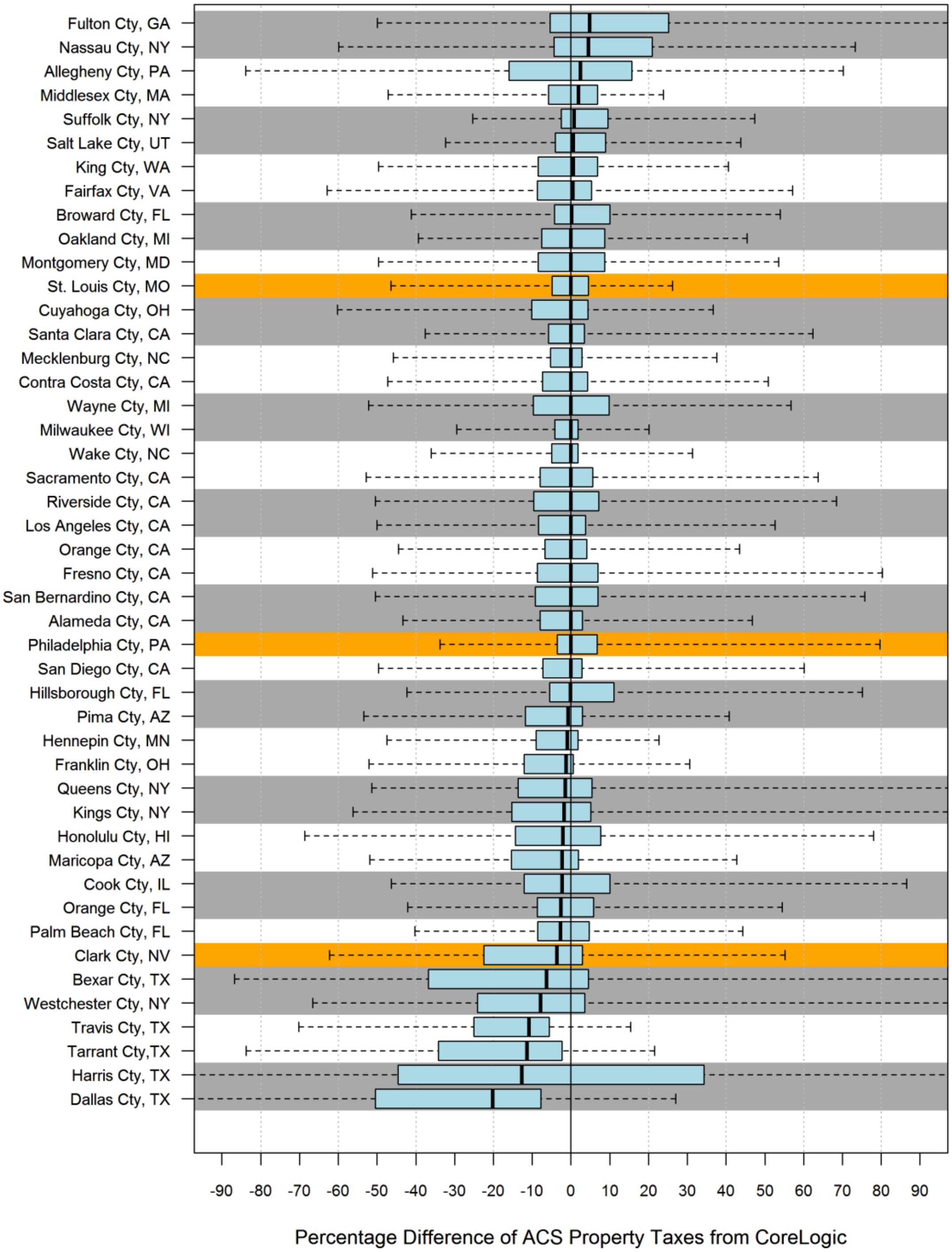
Source: 2010 ACS single-family, owner-occupied households linked to 2008-2010 CoreLogic records. 676,842 Records. Whiskers indicate 5th and 95th Percentiles.

Table 10. 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	**5th**	**25th**	**50th**	**75th**	**95th**	**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, 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,	-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,	-45.8	-5.3	0.0	2.8	37.6	8.0	0.84	2,055
	Contra Costa Cty,	-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,	-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.

Figure 4. Boxplot of Percentage Difference of ACS Property Taxes from CoreLogic Property Taxes by County



Source: 2010 ACS single-family, owner-occupied households linked to 2008-2010 CoreLogic records in select large counties. Whiskers indicate 5th and 95th Percentiles.

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 10 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).

In addition, the research found that CoreLogic may be useful in providing property tax information to substitute for ACS responses when respondents report their property taxes as zero. In 2010, there were 21,358 households that reported their property taxes as zero in the ACS. Of those, 39.4 percent linked to nonzero CoreLogic property tax information, while 47.8 percent were unable to be linked to a CoreLogic record with property tax information. Thus, there is evidence that many of the ACS respondents who report no property taxes are actually paying taxes, and CoreLogic may help provide a value to substitute for the response.

3.4.Comparisons in Clark County, Philadelphia County and St. Louis County

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. In general, I do not take the view that the CoreLogic data are a "gold standard" for property tax amounts. However, 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.[6] Further, the research found that the tax code area determined 98.9 percent of the

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

variation in tax rates for St. Louis County CoreLogic records linked to the 2010 ACS.[7] This information combined with the low error 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, the use of CoreLogic in counties with high quality CoreLogic data to counties with possibly larger errors in CoreLogic can be compared.

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. Table 11 presents the distribution of ACS-CoreLogic percentage differences for ACS responses and for allocations separately in these three counties. The table does not adjust for differences in characteristics between respondents and nonrespondents. In all three cases, the differences are larger for the allocations than for the responses. Among ACS respondents, the percentage differences from CoreLogic are greatest in Clark County, and the median percentage difference in Clark County is -3.7 percent.

Table 11. Distribution of Percentage Difference of ACS Property Taxes from CoreLogic Property Taxes by County for Responses and Allocations

	Percentiles for % Difference of ACS from CoreLogic					Interquartile	Number of
	5th	25th	50th	75th	95th	Range	Records
			Clark County, NV				
Responses	-62.3	-22.5	-3.7	2.9	55.2	25.4	3,514
Allocations	-77.1	-37.3	-12.0	20.5	101.3	57.8	867
			Philadelphia County, PA				
Responses	-33.8	-3.5	-0.1	6.7	79.7	10.2	2,925
Allocations	-86.8	-35.0	-3.8	24.7	204.4	59.6	516
			St. Louis County, MO				
Responses	-46.4	-4.9	0.0	4.5	26.2	9.4	3,592
Allocations	-83.4	-28.5	-5.1	19.4	80.9	47.9	492

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

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

The mean absolute percentage differences are presented for the three counties separately for responses and allocations in

Table 12. Again, in all three cases, the allocations differ more from the CoreLogic information than do the ACS responses. Clark County (27.0 percent) and Philadelphia County (26.1 percent) have much greater mean absolute percentage differences for respondents than does St. Louis County (13.6 percent).

Table 12. Mean Absolute Percentage Difference of ACS Property Taxes from CoreLogic Property Taxes by County for Responses and Allocations

	Clark Cty, NV	Philadelphia Cty, PA	St. Louis Cty, MO
Responses	27.0	26.1	13.6
Allocations	45.1	62.8	39.4

Source: 2010 ACS single-family, owner-occupied households linked to 2008-2010 CoreLogic records in three counties. Number of records available in Table 11.

4. Methods

4.1.Overview

This research compares different methods of using ACS and CoreLogic information to produce county estimates of mean property taxes. I evaluate the use of CoreLogic to address two concerns with the ACS estimates, response error and nonresponse. Estimates are compared for Clark, Philadelphia and St. Louis counties. Recognizing that the CoreLogic data come from different local tax authorities, all modeling is conducted within county.

To study response error, estimates are compared that either use the ACS responses whenever present or use the CoreLogic information whenever present. As such, I can determine how much the estimates would be affected viewing one data source or the other as more trustworthy. To study the use of CoreLogic data in nonresponse modeling, I compare different methods of imputation for item missingness in either dataset.[8] These methods include:

A) for ACS-based estimates, the hot deck allocations currently used by the ACS,

B) for both ACS- and CoreLogic-based estimates, imputing by direct substitution with data from the alternative data source,

C) for ACS-based estimates, multiple imputation by either a parametric or semiparametric modeling approach **using ACS data alone**,

D) for both ACS- and CoreLogic-based estimates, multiple imputation by either a parametric or semiparametric modeling approach **using both ACS and CoreLogic data** and

E) for both ACS- and CoreLogic-based estimates, partially identified interval estimates that relax assumptions about the missingness mechanism and provide bounds for estimates.

For Method A, the ACS currently uses a hot deck single imputation method for item nonresponse, a nonparametric method. The methodology is described in Stiller and Dalzell (1998). The data are sorted based on geographic identifiers. Then, using designated categorical variables, imputation cells are defined. For each nonresponse, a donor is chosen from within the imputation cell of the missing response. For property taxes, lot size and building type are the variables that define the imputation cells. Property taxes are imputed jointly with home value and insurance amount, meaning that if two or more of those variables are missing for a respondent, then the same donor is chosen to impute for all missing values for the nonrespondent.

Method B, for the ACS-based estimates, directly substitutes for missing ACS responses with CoreLogic information. While straightforward to implement, the method treats the CoreLogic

[8] While I only study item nonresponse, many of the methods presented could be used to address unit nonresponse. Different variables would need to be used in modeling imputations, as some ACS data that are available for item nonrespondents are not available for unit nonrespondents.

values as a “gold standard” when CoreLogic property taxes are available and ACS property taxes are not. For either ACS- or CoreLogic-based estimates, when neither ACS responses nor CoreLogic information are available, the hot deck allocations are used. Other approaches use CoreLogic data for estimates but do not require the CoreLogic data to be a “gold standard.” Method D uses CoreLogic for modeling imputations, and Method E uses CoreLogic information by census block group to provide bounds for property tax estimates.

Comparing Methods C and D for the ACS-based estimates allows for studying how nonresponse adjustment changes when CoreLogic data is or is not used in imputation models. By using multiple imputation, I can investigate both whether using CoreLogic changes estimates and whether it improves the precision of the estimates. Differences in the fit of imputation models indicate differences in predictive power. Slud (2015) studied using model-based imputation approaches for categorical ACS variables, including with logistic regression and recursive partitioning methods. He found that the model-based imputations were different from the ACS’s hot deck allocations (Method A). Still, he advocated for studying model-based imputation to evaluate potential changes to imputation modeling as the fit and performance of model-based methods could be more readily studied than for hot deck approaches.

Estimation was conducted in SAS® 9.2 for all estimates that did not use multiple imputation and in *R* with package *mice* for the estimates that used multiple imputation. All estimates of county property taxes presented account for the complex sample design of the ACS and the use of successive difference replication (Fay and Train 1995) to estimate standard errors and confidence intervals (Census Bureau 2014b), with either PROC SURVEYMEANS in SAS® or the *survey* package in *R* (Lumley 2010).

4.2. Modeling Nonresponse with Multiple Imputation

Methods C and D use multiple imputation to impute for missing ACS responses with modeling. Multiple imputation generates m complete datasets with each dataset having imputations for all missing values. Estimates are then constructed on the complete data for each dataset, and the estimates are combined to create one overall estimate and a standard error. Thus, multiple imputation accounts for uncertainty due to nonresponse by accounting for the variation between estimates across the m imputed datasets.

Method C uses ACS information alone, while method D also uses CoreLogic data for the ACS-based estimates. Method C assumes that the data are *missing at random* conditional on ACS information alone. Specifically, let $X_{tax,ACS}$ represent ACS reported property taxes and R be the vector of missingness indicators taking the value 1 when $X_{tax,ACS}$ is observed and 0 when not. Let $X_{other,ACS}$ be the matrix of other variables from the ACS and X_{CL} be the matrix of auxiliary data available from CoreLogic. Methods A and C assume that the data are *missing at random* conditional on other ACS characteristics. This implies that the conditional distribution of the ACS property tax measure is independent of whether the data are observed, $P\left(X_{tax,ACS} \mid R=0, X_{other,ACS}\right) = P\left(X_{tax,ACS} \mid R=1, X_{other,ACS}\right)$.

Method D also assumes that the data are missing at random, but the assumption is less strict than for Method C as CoreLogic variables are used to model the missing data. Specifically, Method D assumes $P\left(X_{tax,ACS} \mid R=0, X_{other,ACS}, X_{CL}\right) = P\left(X_{tax,ACS} \mid R=1, X_{other,ACS}, X_{CL}\right)$. As X_{CL} includes another measure of property taxes and other variables that are correlated with property taxes, using the auxiliary information from CoreLogic may be valuable for nonresponse modeling (Zanutto and Zaslavsky 2002).

An approach that accounts for the uncertainty due to missing data in $X_{tax,ACS}$, $X_{other,ACS}$ and X_{CL} can be advantageous, as several variables in both datasets are missing data. Both Methods C and D use multiple imputation by chained equations (MICE) (van Buuren 2012, van Buuren and Groothuis-Oudshoorn 2011) to compute estimates of property taxes in the three counties and their standard errors. MICE uses Gibbs sampling, a Bayesian Markov Chain Monte Carlo simulation technique, to iteratively sample over the conditional distribution of each variable with missing data, conditioning on other variables used in modeling. While MICE is often advocated for conducting multivariate analyses where more than one variable of interest has missing data, MICE is also useful for studying estimates using a single variable, as this research does for property taxes. MICE algorithms also can account for uncertainty in the parameters of the imputation models.

Now, I describe the MICE algorithms used for this research. Let $X_1,...,X_p$ refer to the vectors of hypothetical responses for all variables in $X_{tax,ACS}$, $X_{other,ACS}$ and X_{CL}. That is, for each $j=1,...,p,$ let X_j includes both the observed responses and the unobserved nonresponses. Write $X_j=\left(X_j^{obs},X_j^{mis}\right),$ where X_j^{obs} are the observed responses and X_j^{mis} are the unobserved nonresponses. Let there be T iterations for each multiply imputed dataset. Let $\tilde{X}_j^t$ refer to completed values for variable X_j on the tth iteration. MICE algorithms typically assume a functional form for the imputation model, but there is uncertainty in the parameters of the model due to missing data. Let ϕ_j be the parameters of the imputation model for variable X_j. Let $\tilde{\phi}_j^t$ refer to the estimate of ϕ_j on the tth iteration. For each dataset, MICE begins by randomly drawing values from the observed data to fill in missing values of the missing data. For each of the imputations, MICE algorithms iteratively

- Draw $\tilde{\phi}_j^t$ from $P\left(\phi_j^t \mid X_j^{obs}, \tilde{X}_{-j}^t, R\right)$, where $\tilde{X}_{-j}^t$ are the completed data for all other variables at that iteration, and then
- Calculate $\tilde{X}_j^t$ by drawing imputations from $P\left(X_j^{mis,t} \mid X_j^{obs}, \tilde{X}_{-j}^t, R, \tilde{\phi}_j^t\right)$.

The algorithm repeats these steps iterating over all variables $X_1, ..., X_p$ and then repeating T times. After completing all iterations, analysis is conducted on the final complete dataset before combining the estimates from each completed dataset.

To implement MICE, choices must be made for the conditional distributions of all ϕ_j and X_j. I conduct estimates using three difference approaches:

- Bayesian imputation using the normal linear model (NORM) (Rubin 2004, van Buuren 2012).
- Predictive mean matching with Bayesian estimation of regression coefficients and a stochastic matching distance (PMM) (Little 1988, van Buuren 2012).
- Recursive partitioning (CART) (Burgette and Reiter 2010; van Buuren 2012; Doove, van Buuren and Dusseldorp 2014).

The NORM procedure is a parametric approach with similarities to that used by Benedetto, Motro and Stinson (2016) for SIPP imputations of program receipt. NORM estimates a linear model at each iteration, estimating $P\left(X_j^{mis,t} \mid X_j^{obs}, \tilde{X}_{-j}^t, R, \tilde{\phi}_j^t\right)$ as a linear function of covariates with a normally distributed error term. Imputations are drawn from this distribution. The parameter vector for the imputation model ϕ_j consists of the regression coefficients for the linear model and the variance of the error term. $P\left(\phi_j^t \mid X_j^{obs}, \tilde{X}_{-j}^t, R\right)$reflects uncertainty in the values of the regression coefficients and the error variance due to missing data.

PMM is a semiparametric approach and can be considered to be a hot deck procedure. The steps of the NORM procedure are used to determine predicted values for each observation for the variable being modeled, for both the observed and missing data. Then, for the predicted value of each nonresponse, the five nearest donors are determined based on the observed values of the five closest predictions among the observed data. A donor is chosen randomly with an equal probability of each donor being chosen.

CART allows for modeling more complex relationships among the variables with and without missing data. For imputation of each $X_j^{mis,t}$, a regression tree model (Breiman et al. 1984) is fit to the data where X_j is observed.[9] Using regression trees allows for important interactions among the variables in $\tilde{X}_{-j}^t$ to influence the imputations. After all observed and missing observations of X_j are assigned to the appropriate terminal nodes, imputations are taken from donors from the observed data in the same terminal node as a missing observation. The donor is selected randomly, with equal probabilities for all potential donors in the terminal node.

For all methods, the ACS variables for property taxes, home value, home insurance amount, mortgage status, age, race, education, household income, year structure built, lot size, number of bedrooms, number of rooms, family size, block group and the survey weight are used in modeling. As Method A, the method used by the Census Bureau, imputes property taxes, home value and home insurance amount jointly, the methods impute the missing data for each of the variables. For the other variables, the ACS's allocations are used for the missing data, and nonresponse is not modeled. The NORM and PMM procedures assume that the error terms of the linear regression

[9] The algorithm attempts to split any node of a regression tree with 15 or more observations, but requires at least 5 observations in each terminal node. In order for a split to be completed, R^2 must increase by at least 0.01%. Each tree has a maximum depth of 30.

models are normally distributed, although PMM is somewhat robust to departures from this assumption (van Buuren 2012). I investigated linear regression models using different transformations for continuous variables and found that residuals were approximately normally distributed when I used square root transformations. Thus, I applied the square root transformation to the property tax, home value, home insurance amount and household income variables for all three methods.

In the three counties studied, different CoreLogic variables are used in modeling due to differences in the CoreLogic data among the counties. CoreLogic property tax amount is always used in modeling. In St. Louis County and Clark County CoreLogic home value is also used. Additionally, Clark County had the square feet of the unit on the file. MICE algorithms account for the uncertainty due to missing data in all three of these variables, and square root transformation are used.

For each model for Methods C and D, $m = 20$ complete datasets are imputed using $T = 30$ iterations over the variables with missing data for each multiple imputation. Analysis is conducted on each complete dataset. The estimates from analysis on each complete imputed dataset are combined to form an overall estimate of a county's property taxes with a standard error (Rubin 2004). Let Q_i be the estimate of a county's property taxes and U_i the estimated variance of that estimate for complete dataset k, $k = 1, ..., m.$ The overall estimate of a county's property taxes is

$$\bar{Q} = \frac{1}{m}\sum_{k=1}^{m} Q_k . \qquad (2)$$

The average within imputation variance is

$$\bar{U} = \frac{1}{m}\sum_{k=1}^{m} U_k . \qquad (3)$$

The between imputation variance is

$$B = \frac{1}{m-1}\sum_{k=1}^{m}\left(Q_k - \overline{Q}\right)^2 . \tag{4}$$

The total variance of the estimate of county property taxes is

$$T = \bar{U} + \left(1 + m^{-1}\right)B, \tag{5}$$

of which one can take the square root to estimate the standard error. The relative increase in variance due to the missing data is then

$$r = \frac{\left(1 + m^{-1}\right)B}{\bar{U}} . \tag{6}$$

4.3.Partially Identified Interval Estimates Using Neighborhood Information

Methods C and D, as well as Method A, assume that the data are missing at random. Specifically, the methods assume that the distribution of the missing responses is the same as the distribution of the observed responses conditional on other ACS variables (Methods A and C) or on both ACS and CoreLogic information together (Method D). All of these methods produce *point estimates* of county property taxes that are possible to obtain by these assumptions about the missingness mechanism. While Method D may be an improvement upon Methods A and C, the assumption that the data are missing at random is not testable without observing the missing data.

Method E uses CoreLogic to provide *partially identified interval estimates* (Manski 2007, Manski 2015, Manski 2016) of county property taxes to account for the possibility that the data are not missing at random, or that the missingness mechanism may depend upon unobservable variables. Specifically, instead of determining a point estimate, an interval is provided for the estimate by making weaker assumptions about the missing data that may be more reasonable than that the data are missing at random.

To avoid having bounds that are so wide as to be uninformative, some assumptions are needed when constructing an interval estimate for a mean. Let $X_{tax,ACS}$ be a measure of property taxes with missing data and R an indicator that $X_{tax,ACS}$ is not missing. By the Law of Iterated Expectations, the mean of $X_{tax,ACS}$ is

$$E\left[X_{tax,ACS}\right]=E\left[X_{tax,ACS}\mid R=1\right]\Pr\left(R=1\right)+E\left[X_{tax,ACS}\mid R=0\right]\left[1-\Pr\left(R=1\right)\right]. \tag{7}$$

The data reveal $E\left[X_{tax,ACS}\mid R=1\right]$, the mean taxes among respondents, and $\Pr\left(R=1\right)$, the fraction responding. However, $E\left[X_{tax,ACS}\mid R=0\right]$, the mean taxes of nonrespondents, is unknown and, without any assumptions or additional information, could theoretically be any value greater than or equal to 0. Therefore, without any assumptions, whenever there are any missing data, or when $\Pr\left(R=1\right)<1$, all that is known is that $E\left[X_{tax,ACS}\right]\geq E\left[X_{tax,ACS}\mid R=1\right]\Pr\left(R=1\right)$. If it is assumed that the data are missing at random conditional on a set of covariates X, then $E\left[X_{tax,ACS}\mid R=0,X\right]=E\left[X_{tax,ACS}\mid R=1,X\right]$, which is observed. Thus, the strong assumption means that the data reveal $E\left[X_{tax,ACS}\mid R=0\right]$, and $E\left[X_{tax,ACS}\right]$ is point identified.

Manski (2016) discusses reasons why one may find the assumption of missingness at random to be too strong but be willing to impose some weaker assumptions to learn about values such as $E\left[X_{tax,ACS}\right]$.One option is to place restrictions on the distribution of the missing observations, or on just the expected value $E\left[X_{tax,ACS}\mid R=0\right]$. This research takes advantage of the similarity of housing characteristics within a neighborhood (Basu and Thibodeau 1998) to place restrictions on $E\left[X_{tax,ACS}\mid R=0\right]$. Specifically, I consider census block groups, which are available in the CoreLogic data. In each of the three counties studied, CoreLogic data are available for more than

90 percent of the single-family, owner-occupied ACS records. As CoreLogic data are not a sample but theoretically a census of households in the county, I posit that in these counties $E\left[X_{tax,ACS}^{BG} \mid R=0\right]$, the expected value of the missing data in the block group, is greater than or equal to the 5th percentile but less than or equal to the 95th percentile of all reported CoreLogic taxes reported in the census block group.[10] As such, bounds are provided for $E\left[X_{tax,ACS} \mid R=0\right]$, and a partially identified interval estimate is determined for $E\left[X_{tax,ACS}\right]$.[11]

Note that this method uses a variable of X, the census block group, to inform the bounds for the interval estimate. Specifically, the approach uses quantiles from the distribution of the CoreLogic data conditional on the block group for estimation. This could be extended by studying quantiles of the distribution of CoreLogic taxes conditional on combinations of covariates from $X_{other,ACS}$ and/or X_{CL}. Then, the effect of the resulting changes in bounds for $E\left[X_{tax,ACS} \mid R=0\right]$ on the partially identified interval estimates of $E\left[X_{tax,ACS}\right]$ can be examined.

5. Results

5.1.County Property Tax Estimates

This section presents estimates of mean property tax estimates in Clark County, NV, Philadelphia County, PA and St. Louis County, MO under different methods either using or not

[10] For some ACS records, there are too few CoreLogic records available in the block group. I assume that the expected property taxes for these records is between the 5th and 95th percentile of CoreLogic property tax records for the entire county.

[11] The estimates of the upper and lower bounds of the interval estimate can be obtained by imputing each missing observation with either the 5th or 95th percentile of the block group's CoreLogic property taxes and using the survey weights to take the weighted average of all records. To obtain standard errors, Horowitz and Manski (2000) advocate using the bootstrap as a resampling technique. As successive difference replication is a resampling technique that accounts for the survey design, the ACS replicate weights are used to determine standard errors for the bounds of the interval estimates. Bonferroni joint confidence intervals (Dunn 1961) are calculated to account for that estimates of the upper and lower bounds are jointly estimated.

using the CoreLogic records. Results can be found in Table 13 and in Figure 5.[12] For all three counties studied, the largest differences in estimates are due to whether either the ACS responses or the CoreLogic records are primarily used to construct estimates. In St. Louis County, the estimates primarily using the CoreLogic records are 2.4 to 2.6 percent larger than the ACS estimate of Method A. 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 2.4 to 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 6.8 to 6.9 percent higher than the ACS estimate in Clark County and 8.3 to 11.1 percent lower in Philadelphia County.

The county property tax estimates are also sensitive as to whether CoreLogic information is directly substituted for imputations (Method B) or whether a model-based imputation approach is used (Methods C and D). In general, using direct substitution of CoreLogic values changes the estimate from the Method A estimate in the same direction as constructing the estimates primarily using CoreLogic data. In Clark County, using direct substitution increased the mean property tax estimate by 2.2 percent, compared to a 0.2 to 0.9 percent change for the model-based imputation methods. In St. Louis County, using direct substitution increased the mean property tax estimate by 0.7 percent, compared to a -0.8 to +0.1 percent change for the model-based imputations. In Philadelphia County, direct substitution decreased the mean property tax estimate by 4.2 percent. In this case, the estimates using model-based imputation were also much lower than the ACS

[12] All estimates presented in Section 5 use the survey weights as well as the ACS replicate weights with jackknife replication for confidence intervals.

estimate by 3.1 to 4.5 percent. These findings provide reason to be wary of imputing for ACS nonresponse directly with the CoreLogic data. There may be errors in the CoreLogic data in Clark and Philadelphia counties, and direct substitution biases estimates in the direction of those errors.

Table 13. 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 – Hot Deck Allocations	**2160 (28)**	**1526 (28)**	**2788 (39)**
[95% Confidence Interval]	*[2105, 2216]*	*[1471, 1581]*	*[2711, 2864]*
ACS – CoreLogic Substitutions	**2207 (25)**	**1462 (26)**	**2806 (34)**
[95% Confidence Interval]	*[2157, 2257]*	*[1411, 1513]*	*[2737, 2875]*
ACS – NORM MI without CoreLogic	**2170 (27)**	**1464 (24)**	**2765 (37)**
[95% Confidence Interval]	*[2115, 2224]*	*[1416, 1512]*	*[2692, 2860]*
ACS – PMM MI without CoreLogic	**2171 (29)**	**1460 (24)**	**2762 (37)**
[95% Confidence Interval]	*[2113, 2228]*	*[1412, 1509]*	*[2689, 2836]*
ACS – CART MI without CoreLogic	**2181 (28)**	**1457 (24)**	**2773 (37)**
[95% Confidence Interval]	*[2124, 2237]*	*[1408, 1505]*	*[2700, 2847]*
ACS – NORM MI with CoreLogic	**2166 (27)**	**1472 (25)**	**2781 (36)**
[95% Confidence Interval]	*[2113, 2220]*	*[1421, 1522]*	*[2709, 2852]*
ACS – PMM MI with CoreLogic	**2165 (27)**	**1469 (26)**	**2780 (36)**
[95% Confidence Interval]	*[2111, 2220]*	*[1417, 1520]*	*[2709, 2851]*
ACS – CART MI with CoreLogic	**2166 (27)**	**1478 (26)**	**2790 (36)**
[95% Confidence Interval]	*[2112, 2220]*	*[1426, 1531]*	*[2719, 2860]*
ACS – Neighborhood-Based Bounds	**[1816 (31), 2749 (29)]**	**[1382 (23), 1541 (26)]**	**[2649 (38), 3006 (36)]**
[95% Confidence Interval]	*[1746, 2815]*	*[1329, 1600]*	*[2561, 3089]*
CoreLogic – ACS Substitutions	**2309 (24)**	**1399 (21)**	**2860 (33)**
[95% Confidence Interval]	*[2262, 2356]*	*[1357, 1441]*	*[2794, 2925]*
CoreLogic – NORM MI	**2309 (23)**	**1357 (18)**	**2855 (33)**
[95% Confidence Interval]	*[2263, 2355]*	*[1320, 1393]*	*[2790, 2920]*
CoreLogic – PMM MI	**2307 (23)**	**1357 (18)**	**2854 (33)**
[95% Confidence Interval]	*[2260, 2353]*	*[1320, 1393]*	*[2789, 2919]*
CoreLogic – CART MI	**2309 (23)**	**1372 (20)**	**2857 (33)**
[95% Confidence Interval]	*[2263, 2355]*	*[1333, 1411]*	*[2791, 2923]*
CoreLogic – Neighborhood-Based Bounds	**[2213 (23), 2432 (27)]**	**[1226 (17), 1509 (19)]**	**[2743 (42), 3015 (37)]**
[95% Confidence Interval]	*[2160, 2492]*	*[1186, 1553]*	*[2671, 3099]*
Number of Records	4,650	3,815	4,274

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

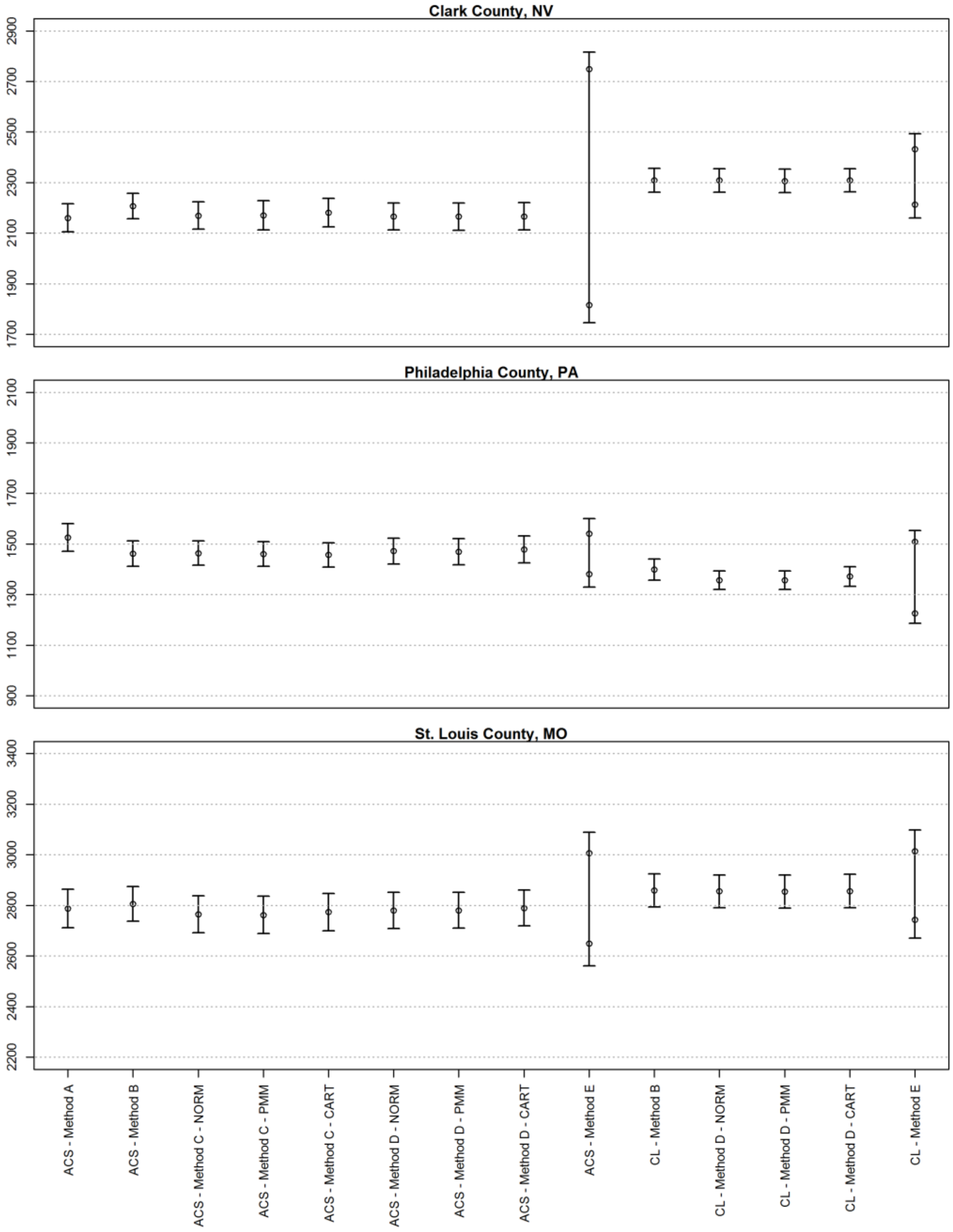
Using a model-based imputation approach, the county estimates are neither sensitive to the choice of imputation method nor whether CoreLogic data is used in imputation modeling. The model-based imputation methods range between $2,165 and $2,181 in Clark County, $1,460 and $1,478 in Philadelphia County and $2,762 and $2,790 in St. Louis County. In Clark County and St. Louis County, the estimates are not very different from the ACS estimates using hot deck allocations. The estimates are even more similar among the three multiple imputation approaches either using CoreLogic data (Method D) or not using CoreLogic data (Method C). In all three counties, these estimates never differ by more than $11 whether NORM, PMM or CART is used. Estimates across the three multiple imputation methods for the CoreLogic-based estimates are also very similar to each other. The estimated standard errors, which for the model-based estimates account for the uncertainty due to imputation, are similar whether or not CoreLogic data are used in modeling, indicating that the imputation approach has little impact on the uncertainty of estimates. The relative increase in variance in all multiple imputation estimates is less than 1 percent.

Partially identified interval estimates using the neighborhood-based bounds (Method E) are presented for all three counties for both ACS- and CoreLogic-based estimates. These intervals relax the assumption that the survey responses are missing at random. In all cases, the intervals cover the range of the other point-identified estimates. The intervals are never wider than $400 except in Clark County, which has the largest rate of survey nonresponse.

In addition, estimates are presented for mean property taxes by mortgage status in each of the three counties in Tables 14 and 15 as well as in Figure 6. Mostly, the same patterns emerge as for the overall mean property tax estimates for the county. There are large differences between the ACS- and CoreLogic-based estimates. In fact, for St. Louis 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. Estimates are mostly insensitive to the choice of imputation method or whether CoreLogic data are used, although for St. Louis County the multiple imputation estimates using

CoreLogic data are closer to the hot deck allocation estimates than the multiple imputation estimates not using CoreLogic data.

Figure 5. Comparison of Different Estimates of Mean Property Tax Amounts ($) for Single-Family, Owner-Occupied Homes in Three Counties



Source: 2010 ACS single-family, owner-occupied households and linked CoreLogic records in three counties. Error bars represent 95 percent confidence intervals. Partially identified interval estimates for neighborhood-based bounds presented for Method E with two points for lower and upper bounds of interval estimates.

Table 14. 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 – Hot Deck Allocations	**2156 (31)**	**1529 (26)**	**2774 (49)**
[95% Confidence Interval]	*[2094, 2217]*	*[1477, 1582]*	*[2676, 2872]*
ACS – CoreLogic Substitutions	**2206 (29)**	**1524 (27)**	**2793 (46)**
[95% Confidence Interval]	*[2147, 2264]*	*[1470, 1578]*	*[2701, 2885]*
ACS – NORM MI without CoreLogic	**2164 (30)**	**1535 (27)**	**2756 (50)**
[95% Confidence Interval]	*[2104, 2224]*	*[1481, 1589]*	*[2657, 2856]*
ACS – PMM MI without CoreLogic	**2164 (33)**	**1532 (27)**	**2754 (50)**
[95% Confidence Interval]	*[2099, 2230]*	*[1478, 1586]*	*[2654, 2853]*
ACS – CART MI without CoreLogic	**2176 (32)**	**1531 (27)**	**2768 (50)**
[95% Confidence Interval]	*[2112, 2241]*	*[1477, 1585]*	*[2669, 2867]*
ACS – NORM MI with CoreLogic	**2162 (30)**	**1537 (27)**	**2767 (48)**
[95% Confidence Interval]	*[2102, 2223]*	*[1483, 1591]*	*[2671, 2862]*
ACS – PMM MI with CoreLogic	**2163 (31)**	**1534 (27)**	**2765 (48)**
[95% Confidence Interval]	*[2101, 2225]*	*[1480, 1588]*	*[2671, 2860]*
ACS – CART MI with CoreLogic	**2161 (31)**	**1551 (31)**	**2777 (47)**
[95% Confidence Interval]	*[2099, 2223]*	*[1490, 1613]*	*[2684, 2870]*
ACS – Neighborhood-Based Bounds	**[1774 (31), 2795 (35)]**	**[1445 (26), 1607 (26)]**	**[2643 (50), 2983 (44)]**
[95% Confidence Interval]	*[1703, 2875]*	*[1386, 1668]*	*[2529, 3083]*
CoreLogic – ACS Substitutions	**2318 (29)**	**1456 (22)**	**2859 (43)**
[95% Confidence Interval]	*[2261, 2375]*	*[1412, 1500]*	*[2774, 2943]*
CoreLogic – NORM MI	**2316 (28)**	**1411 (22)**	**2862 (42)**
[95% Confidence Interval]	*[2260, 2373]*	*[1368, 1454]*	*[2778, 2946]*
CoreLogic – PMM MI	**2314 (28)**	**1410 (21)**	**2859 (42)**
[95% Confidence Interval]	*[2258, 2371]*	*[1367, 1452]*	*[2775, 2943]*
CoreLogic – CART MI	**2315 (28)**	**1433 (22)**	**2857 (42)**
[95% Confidence Interval]	*[2259, 2371]*	*[1390, 1477]*	*[2773, 2941]*
CoreLogic – Neighborhood-Based Bounds	**[2222 (28), 2427 (32)]**	**[1284 (22), 1566 (21)]**	**[2754 (42), 3025 (46)]**
[95% Confidence Interval]	*[2159, 2500]*	*[1235, 1614]*	*[2659, 3131]*
Number of Records	3,709	2,201	2,945

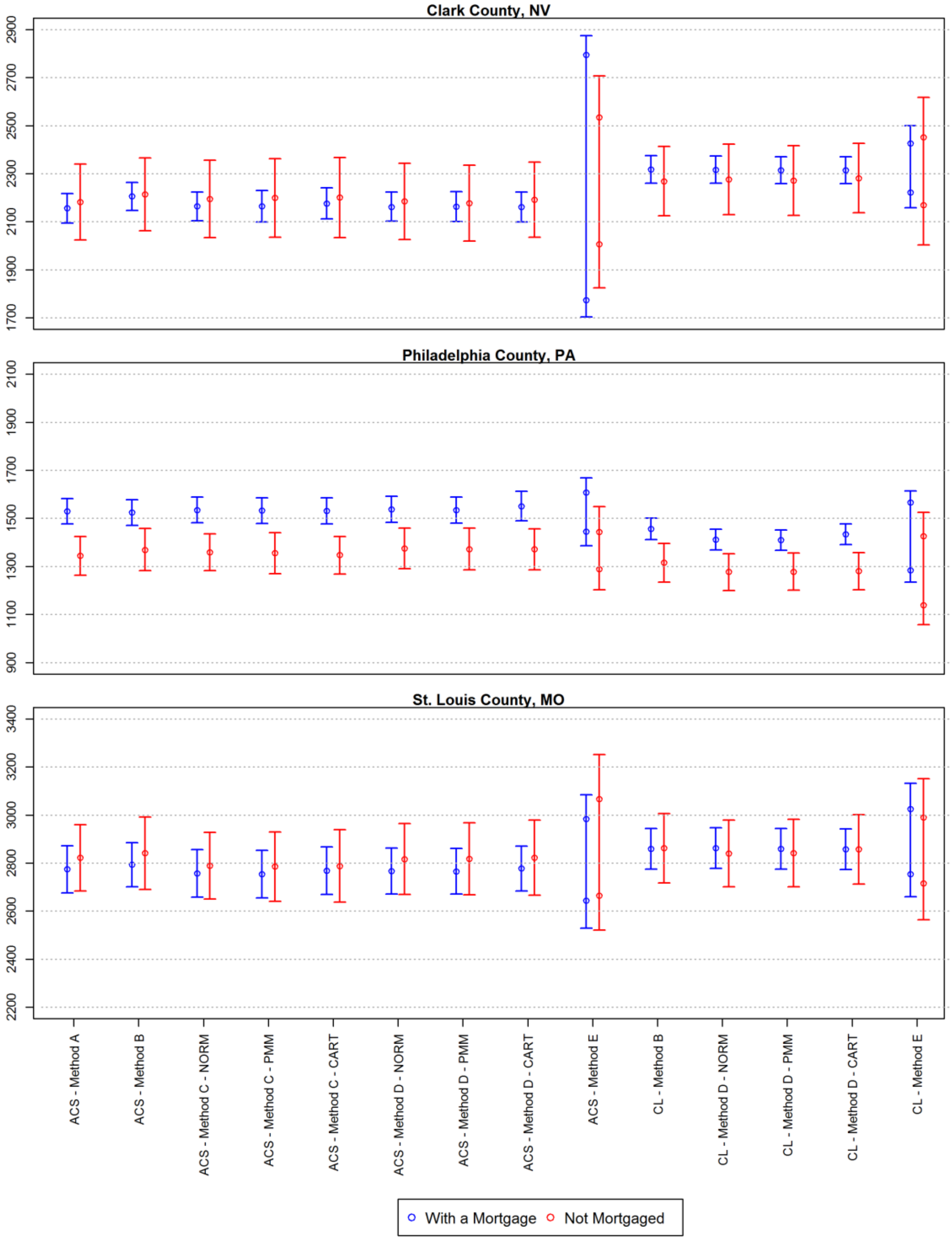
Source: 2010 ACS single-family, owner-occupied households with a mortgage and linked CoreLogic records in three counties.

Table 15. 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 – Hot Deck Allocations	**2182 (79)**	**1344 (41)**	**2822 (69)**
[95% Confidence Interval]	*[2024, 2340]*	*[1263, 1424]*	*[2684, 2960]*
ACS – CoreLogic Substitutions	**2214 (76)**	**1369 (44)**	**2841 (76)**
[95% Confidence Interval]	*[2062, 2366]*	*[1282, 1457]*	*[2690, 2991]*
ACS – NORM MI without CoreLogic	**2195 (81)**	**1358 (39)**	**2788 (70)**
[95% Confidence Interval]	*[2033, 2356]*	*[1282, 1435]*	*[2650, 2927]*
ACS – PMM MI without CoreLogic	**2200 (82)**	**1355 (43)**	**2785 (72)**
[95% Confidence Interval]	*[2036, 2363]*	*[1270, 1440]*	*[2641, 2929]*
ACS – CART MI without CoreLogic	**2201 (84)**	**1347 (39)**	**2787 (76)**
[95% Confidence Interval]	*[2034, 2368]*	*[1268, 1425]*	*[2637, 2938]*
ACS – NORM MI with CoreLogic	**2185 (80)**	**1375 (43)**	**2816 (74)**
[95% Confidence Interval]	*[2025, 2344]*	*[1290, 1460]*	*[2669, 2964]*
ACS – PMM MI with CoreLogic	**2178 (79)**	**1372 (44)**	**2818 (75)**
[95% Confidence Interval]	*[2019, 2336]*	*[1285, 1460]*	*[2668, 2968]*
ACS – CART MI with CoreLogic	**2191 (79)**	**1371 (43)**	**2822 (75)**
[95% Confidence Interval]	*[2035, 2348]*	*[1286, 1456]*	*[2666, 2978]*
ACS – Neighborhood-Based Bounds	**[2006 (80), 2535 (75)]**	**[1288 (37), 1444 (46)]**	**[2664 (62), 3066 (81)]**
[95% Confidence Interval]	*[1824, 2707]*	*[1203, 1548]*	*[2521, 3251]*
CoreLogic – ACS Substitutions	**2269 (72)**	**1316 (41)**	**2862 (72)**
[95% Confidence Interval]	*[2125, 2414]*	*[1235, 1396]*	*[2717, 3006]*
CoreLogic – NORM MI	**2276 (74)**	**1277 (38)**	**2839 (70)**
[95% Confidence Interval]	*[2129, 2423]*	*[1200, 1353]*	*[2701, 2978]*
CoreLogic – PMM MI	**2271 (73)**	**1278 (38)**	**2841 (70)**
[95% Confidence Interval]	*[2126, 2416]*	*[1201, 1356]*	*[2701, 2981]*
CoreLogic – CART MI	**2281 (73)**	**1280 (39)**	**2857 (73)**
[95% Confidence Interval]	*[2137, 2426]*	*[1203, 1357]*	*[2712, 3001]*
CoreLogic – Neighborhood-Based Bounds	**[2169 (73), 2452 (72)]**	**[1139 (36), 1426 (43)]**	**[2716 (67), 2989 (71)]**
[95% Confidence Interval]	*[2004, 2618]*	*[1057, 1524]*	*[2563, 3151]*
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.

Figure 6. Comparison of Different Estimates of 2010 Mean Property Tax Amounts ($) by Mortgage Status for Single-Family, Owner-Occupied Homes in Three Counties



Source: 2010 ACS single-family, owner-occupied households and linked CoreLogic records in three counties. Error bars represent 95 percent confidence intervals. Partially identified interval estimates for neighborhood-based bounds presented for Method E with two points for lower and upper bounds of interval estimates.

5.2.Imputation Model Diagnostics

Using CoreLogic data in imputation modeling modestly improves the fit of imputation models. Table 16 presents the R^2 and Adjusted R^2 for the imputation models used in linear regression with the NORM procedure. These estimates were determined by combining results from regressions across the multiply imputed data. In all three counties, using CoreLogic covariates for imputation improves the fit of the models. In Clark County, R^2 increases by 0.044 and Adjusted R^2 by 0.048. In Philadelphia, R^2 and Adjusted R^2 increase by 0.080 and 0.089 respectively. In St. Louis County, R^2 increases by 0.037 and Adjusted R^2 by 0.039. These increases are modest. Still, they indicate the value of using the CoreLogic data for imputation modeling. Of note are the high values of R^2 in St. Louis County. Of the three counties, St. Louis County had the highest correlation between ACS and CoreLogic property taxes of 0.92. However, the model with only ACS covariates also has a strong goodness of fit, with R^2 of almost 0.8.

Table 16. R^2 (Adjusted) for Regression Models of ACS Property Taxes With and Without CoreLogic Covariates Using Multiply Imputed Data

	Clark Cty, NV	**Philadelphia Cty, PA**	**St. Louis Cty, MO**
With ACS and CoreLogic Covariates	0.667 (0.628)	0.758 (0.731)	0.836 (0.827)
With ACS Covariates Only	0.623 (0.580)	0.678 (0.642)	0.799 (0.788)
Number of Records	4,650	3,815	4,274

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

Beyond goodness of fit, determining diagnostics for imputation models can be a challenge. I present the distributions of the percentage difference of the imputations from the CoreLogic data. In addition, I conduct a simulation randomly setting some ACS responses to missing and study how the imputations using the modified dataset compare to the actual ACS responses. Nonetheless, I recognize the limitations of this approach for evaluating imputations. Imputation is not prediction. All methods using multiple imputation presented in Section 4.2 incorporate a stochastic component related to the uncertainty of the imputation models. This allows standard errors for estimates to account for the uncertainty due to missing data. Therefore, the ideal imputation method is not the one that perfectly predicts the missing values. Nonetheless, studying the percentage difference of imputations from either the CoreLogic or ACS value can be informative for exploratory purposes. First, the various methods can be compared to the hot deck allocations. Second, goodness of fit is desirable for imputation models. To the extent that reduced variation of imputations about the CoreLogic or ACS numbers indicates goodness of fit of the imputation model, the results provide evidence to explore which models perform better.

Table 17 presents the mean absolute percentage differences of the imputation methods from CoreLogic for all three counties, comparing the hot deck allocations to the three imputation methods with and without CoreLogic data used in modeling. Table 18 presents quantiles of the percentage difference of imputations from the CoreLogic values. For both tables, when multiple imputation is conducted, the average is taken across the multiple imputations. In addition, both tables compare the results to the distribution of the percentage differences for ACS responses from CoreLogic. In all cases, the mean absolute percentage differences tend to be higher for the multiple imputation procedures not using CoreLogic than for the ACS's hot deck allocation. Yet, the mean absolute percentage differences are less for the multiple imputation procedures using CoreLogic than for hot deck allocation, particularly so for the PMM and CART imputations. For example, in St. Louis County, the mean absolute percentage difference is 39.4% for the ACS hot deck allocations, 33.1 percent for NORM using CoreLogic, 25.9 percent for PMM using CoreLogic and 19.2 percent for CART using CoreLogic. The patterns are similar in the other two counties. Using CoreLogic data in imputation modeling substantially brings imputations closer to the CoreLogic values. The strong results for PMM and CART indicate that these methods should be considered seriously in comparison to a linear regression or hot deck approach. The semiparametric approach of PMM as well as the model selection aspect of CART may be advantageous. Note that in none of the cases is the mean absolute percentage difference smaller than that for the ACS responses.

Table 17. Mean Absolute Percentage Difference of Imputations from CoreLogic Property Taxes for ACS-Based Estimates

	Clark Cty, NV	Philadelphia Cty, PA	St. Louis Cty, MO
Hot Deck Allocations	45.1	62.8	39.4
NORM MI without CoreLogic	54.7	76.6	47.8
PMM MI without CoreLogic	47.8	62.1	40.8
CART MI without CoreLogic	47.8	95.2	46.3
NORM MI with CoreLogic	45.1	58.5	33.1
PMM MI with CoreLogic	38.1	44.8	25.9
CART MI with CoreLogic	32.3	37.9	19.2
Responses	27.0	26.1	13.6

Source: 2010 ACS single-family, owner-occupied households and linked CoreLogic records in three counties. See Table 18 for number of records.

One concern with evaluating the imputation models by comparing imputations to CoreLogic data is that the evaluation could unfairly advantage the methods using the CoreLogic data, while the ACS measures the concept truly of interest. Thus, I also conducted some simple simulations to compare imputations to the ACS responses. For these simulations, I removed the records with allocations for ACS property taxes from the datasets. Then, each ACS property tax response has 15 percent probability of being simulated as missing. I then ran the six imputation methods of Methods C and D on these datasets within each county using $m = 2$ multiple imputations.

Table 19 presents the mean absolute percentage different of the imputations in the simulation from the actual ACS responses, and Table 20 presents quantiles of the percentage difference. In both cases, the mean of statistics across the multiple imputations was used. For all three counties, the CART imputation using CoreLogic data performs the best of any of the six imputation models. Predictive mean matching using CoreLogic data performs very well in Clark County. It also performs well in St. Louis County, particularly compared to predictive mean matching without using CoreLogic data, but performs similarly to CART imputation not using CoreLogic data.

Predictive mean matching imputation with or without CoreLogic data perform similarly in Philadelphia County. Interestingly, linear regression (NORM) imputation using CoreLogic data does not perform materially better than linear regression imputation without CoreLogic data. While this initial evidence suggests that using CoreLogic data with either a CART imputation approach or sometimes a PMM approach improves imputation models, this hypothesis can be investigated further with additional simulations using more complex patterns of missing data.

Some caution is needed for interpreting these results. As imputation is not prediction, these tables should not provide the deciding criteria for choosing among imputation methods. Ultimately what matters is the impact of the imputation method on the performance of estimates of interest, specifically regarding bias and coverage. Estimates of mean county property taxes and mean county taxes by mortgage status are mostly insensitive to the imputation model. However, there may be multivariate estimates or estimates for smaller geographic areas that are sensitive to the imputation method.

Table 18. Distribution of Percentage Difference of Imputations from CoreLogic Property Taxes for ACS-Based Estimates

	Percentiles for % Difference of Imputation from CoreLogic					Interquartile	Number of
	5th	25th	50th	75th	95th	Range	Records
			Clark County, NV				
Hot Deck Allocations	-77.1	-37.3	-12.0	20.5	101.3	57.8	867
NORM MI without CoreLogic	-87.5	-50.7	-14.0	32.8	135.7	83.5	
PMM MI without CoreLogic	-81.4	-40.9	-10.9	26.7	118.6	67.6	
CART MI without CoreLogic	-79.4	-36.7	-7.6	27.7	126.4	64.5	
NORM MI with CoreLogic	-82.6	-45.3	-13.6	24.2	102.6	69.5	
PMM MI with CoreLogic	-73.4	-34.8	-10.6	18.0	85.7	52.8	
CART MI with CoreLogic	-74.5	-26.2	-6.3	7.9	68.1	34.0	
Responses	-62.3	-22.5	-3.7	2.9	55.2	25.4	3,514
			Philadelphia County, PA				
Hot Deck Allocations	-86.8	-35.0	-3.8	24.7	204.4	59.7	516
NORM MI without CoreLogic	-85.7	-47.2	-4.6	54.7	226.1	101.9	
PMM MI without CoreLogic	-74.9	-36.7	-3.6	41.9	196.6	78.5	
CART MI without CoreLogic	-83.9	-48.4	-7.1	55.7	332.4	104.1	
NORM MI with CoreLogic	-78.4	-36.3	-0.8	45.3	165.0	81.6	
PMM MI with CoreLogic	-64.9	-26.9	-1.2	31.0	124.4	58.0	
CART MI with CoreLogic	-57.7	-9.0	0.4	13.1	108.1	22.1	
Responses	-33.8	-3.5	-0.1	6.7	79.7	10.2	2,925
			St. Louis County, MO				
Hot Deck Allocations	-83.4	-28.5	-5.1	19.4	80.9	47.9	492
NORM MI without CoreLogic	-82.7	-46.5	-14.3	24.2	105.2	70.8	
PMM MI without CoreLogic	-78.5	-37.7	-9.6	19.3	83.8	56.9	
CART MI without CoreLogic	-76.4	-36.4	-6.5	25.7	107.2	62.1	
NORM MI with CoreLogic	-64.4	-33.9	-10.0	16.7	68.3	50.6	
PMM MI with CoreLogic	-55.8	-24.7	-6.4	11.1	48.2	35.8	
CART MI with CoreLogic	-60.9	-10.4	-0.8	6.5	32.8	16.9	
Responses	-46.4	-4.9	0.0	4.5	26.2	9.4	3,592

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

Table 19. Mean Absolute Percentage Difference of Imputations from ACS Response in Simulation Study

	Clark Cty, NV	Philadelphia Cty, PA	St. Louis Cty, MO
NORM MI without CoreLogic	58.1	79.8	62.7
PMM MI without CoreLogic	56.5	72.7	60.0
CART MI without CoreLogic	54.8	85.1	48.4
NORM MI with CoreLogic	59.0	75.5	60.3
PMM MI with CoreLogic	48.1	70.9	48.3
CART MI with CoreLogic	47.3	58.3	40.5
Number of Records	557	478	583

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

6. 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.

Table 20. Distribution of Percentage Difference of Imputations from ACS Responses in Simulation Study

	Percentiles for % Difference of Imputation from ACS Response					Interquartile Range	Number of Records
	5th	25th	50th	75th	95th		
			Clark County, NV				
NORM MI without CoreLogic	-70.4	-37.7	-6.1	45.4	170.6	83.1	557
PMM MI without CoreLogic	-67.6	-33.9	-4.7	41.4	149.7	75.3	
CART MI without CoreLogic	-64.3	-31.7	-3.1	31.4	157.6	63.1	
NORM MI with CoreLogic	-71.6	-36.2	-3.0	42.5	177.0	78.7	
PMM MI with CoreLogic	-65.7	-30.5	-4.0	31.6	144.7	62.1	
CART MI with CoreLogic	-60.0	-24.5	0.0	26.7	155.7	51.2	
			Philadelphia County, PA				
NORM MI without CoreLogic	-73.2	-35.1	0.1	49.9	215.9	85.0	478
PMM MI without CoreLogic	-63.3	-27.4	3.7	46.2	211.0	73.6	
CART MI without CoreLogic	-67.8	-31.0	0.1	49.5	283.7	80.5	
NORM MI with CoreLogic	-67.2	-27.0	7.1	51.0	154.4	78.0	
PMM MI with CoreLogic	-59.3	-23.2	0.0	33.9	233.3	57.1	
CART MI with CoreLogic	-48.7	-13.3	0.0	15.4	165.0	28.7	
			St. Louis County, MO				
NORM MI without CoreLogic	-58.1	-27.2	-0.7	34.3	121.6	61.5	583
PMM MI without CoreLogic	-48.6	-18.9	-0.1	26.3	114.6	45.2	
CART MI without CoreLogic	-55.8	-22.1	-0.3	27.8	95.9	49.9	
NORM MI with CoreLogic	-52.0	-21.8	-1.0	22.9	87.2	44.7	
PMM MI with CoreLogic	-48.7	-15.0	1.6	18.6	77.1	33.6	
CART MI with CoreLogic	-49.2	-11.1	-0.1	12.1	64.0	23.2	

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

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 2.5 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.

In many instances where ACS respondents reported their property taxes as zero, the CoreLogic file recorded that the respondents paid property taxes. The Census Bureau should consider using CoreLogic for imputations for these specific respondents in any counties where the CoreLogic data are believed to be of sufficient quality.

One promising method of using the CoreLogic data to improve survey estimates is to use CoreLogic as auxiliary data to improve imputations and nonresponse adjustments. This approach does not require that the commercial data are a "gold standard," but it does require that the CoreLogic data offer good coverage of the population and that CoreLogic variables improve the goodness of fit of imputation models for ACS taxes, which is not true of all counties and townships. Other research on uses of administrative records and commercial data indicate that this approach may be promising (Zanutto and Zaslavsky 2002, Peytchev and Raghunathan 2013, West and Little 2013, Benedetto et al. 2016). The research identified Clark County, Philadelphia County and St. Louis County as examples of counties fulfilling these criteria and found that using CoreLogic in imputation modeling modestly improves the goodness of fit of the imputation models. The county and county by mortgage status estimates were not sensitive to the imputation method nor the use of CoreLogic data for nonresponse adjustment. Yet, some initial evidence finds using CoreLogic data with either a predictive mean matching or recursive partitioning approach for imputation modeling decreases the percentage difference between imputations and CoreLogic values relative to the current hot deck approach. Future evaluations can examine how using CoreLogic in modeling imputations affects other estimates of interest.

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. Third, all nonresponse and imputation modeling is conducted with models run within each of the three counties, as the research recognizes the differences in the CoreLogic data between counties. This approach would not be feasible for the entire ACS file. Future work could study how to adapt imputation models for the ACS file covering the entire U.S. Finally, the imputation models used all rely on the assumption that the survey responses are missing at random, conditional on either ACS variables alone or on ACS and CoreLogic variables together. This assumption is not testable. Partially identified interval estimates using the CoreLogic data provide one way of relaxing this assumption to produce estimates.

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. Still, there are principled ways that the commercial data can be used to benefit statistical products even when the data are not a “gold standard.” As new approaches toward federal statistical products are considered in the future, careful evaluations of “found” data will continue to be needed.

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.

Basu, S. & Thibodeau, T. G. (1998). Analysis of spatial autocorrelation in house prices. *The Journal of Real Estate Finance and Economics*, *17*(1), 61–85.

Bee, C. A., Gathright, G. M. R. & Meyer, B. D. (2015). Bias from unit non-response in the measurement of income in household surveys. University of Chicago Working Paper. <http://harris.uchicago.edu/sites/default/files/JSM2015%20BGM%20Unit%20Non-Response%20in%20CPS.pdf>. Accessed May 3, 2016.

Benedetto, G., Motro, J. & Stinson, M. (2016). Introducing parametric models and administrative records into 2014 SIPP imputations. Proceedings of the 2015 Federal Committee on Statistical Methodology Research Conference.

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.

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984). *Classification and regression trees*. Boca Raton, FL: Chapman and Hall/CRC.

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

Burgette, L. F. & Reiter, J. P. (2010). Multiple imputation for missing data via sequential regression trees. *American Journal of Epidemiology*, 172, 1070–1076.

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 (2014). Design and methodology: American Community Survey. <http://www2.census.gov/programs-surveys/acs/methodology/design_and_methodology/acs_design_methodology_report_2014.pdf>. 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.

Census Bureau (2016). American Community Survey response rates. <http://www.census.gov/acs/www/methodology/sample-size-and-data-quality/response-rates/>.

Accessed May 3, 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.

Doove, L. L., van Buuren, S. & Dusseldorp, E. (2014). Recursive partitioning for missing data imputation in the presence of interaction effects. *Computational Statistics & Data Analysis*, *72*, 92–104.

Dunn, O. J. (1961). Multiple comparisons among means. *Journal of the American Statistical Association*, *56*(293), 52–64.

Fay, R. E. & Train, G. (1995). Aspects of survey and model-based postcensal estimation of income and poverty characteristics for states and counties. *Proceedings of the Section on Government Statistics, American Statistical Association,* 154-159. Alexandria, VA.

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.

Hokayem, C., Bollinger, C. & Ziliak, J. P. (2014). The role of CPS nonresponse on the level and trend in poverty. SEHSD Working Paper #2014-14. Washington, D.C.: U.S. Census Bureau.

Horowitz, J. L. & Manski, C. F. (2000). Nonparametric analysis of randomized experiments with missing covariate and outcome data. *Journal of the American Statistical Association*, *95*(449), 77–84.

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.

Keller, A. (2016). Imputation reesearch for the 2020 Census. *Statistical Journal of the IAOS*, *32*, 189-198.

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.

Little, R. J. A. (1988). Missing-data adjustments in large surveys. *Journal of Business & Economic Statistics, 6*(3), 287–296.

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

Manski, C. F. (2007). *Identification for prediction and decision*. Harvard University Press.

Manski, C. F. (2015). Communicating uncertainty in official economic statistics: An appraisal fifty years after Morgenstern. *Journal of Economic Literature, 53*(3), 631–653.

Manski, C. F. (2016). Credible interval estimates for official statistics with survey nonresponse. *Journal of Econometrics*, *191*(2), 293–301.

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.

Morris, D. S., Keller, A. & Clark, B. (2016). An approach for using administrative records to reduce contacts in the 2020 Decennial Census. *Statistical Journal of the IAOS*, *32*, 177-188.

Mule, V. T. & Keller, A. (2014). Using administrative records to reduce census nonresponse followup operations. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 3601-3608. Alexandria, VA: American Statistical Association

Mulry, M. H., Nichols, E. M. & Childs, J. H. (2014). Study of error in survey reports of move month using the U.S. Postal Service change of address records. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 3109-3123. Alexandria, VA: American Statistical Association.

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.

National Research Council (2009). *Reengineering the Survey of Income and Program Participation*. Panel on the Census Bureau's Reengineered Survey of Income and Program Participation, J. K. Scholz & C. F. Citro, Eds. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

National Research Council (2013). *Nonresponse in social science surveys: A research agenda*. Panel on a Research Agenda for the Future of Social Science Data Collection, T. J. Plewes & R. Tourangeau, Eds. Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, D.C.: National Academy Press.

Peytchev, A. & Raghunathan, T. (2013). Evaluation and use of commercial data for nonresponse bias adjustment. Paper presented at the Annual Conference of the American Association for Public Opinion Research, Boston, MA.

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.

Rubin, D. (2004). *Multiple imputation for nonresponse in surveys*. New York: Wiley.

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.

Saint Louis County Department of Revenue. Tax rates. <https://revenue.stlouisco.com/Collection/YourTaxRates.aspx>. Accessed April 17, 2016.

Seaman, S. R. & White, I. R. (2013). Review of inverse probability weighting for dealing with missing data. *Statistical Methods in Medical Research*, *22*(3), 278–95.

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.”

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: Chapman & Hall/CRC.

van Buuren, S. & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3).

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

West, B. T. & Little, R. J. A. (2013). Non-response adjustment of survey estimates based on auxiliary variables subject to error. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 62*(2), 213–231.

West, B. T., Wagner, J., Hubbard, F. & Gu, H. (2015). The utility of alternative dommercial data sources for survey operations and estimation: Evidence from the National Survey of Family Growth. *Journal of Survey Statistics and Methodology*, *3*(2), 240–264.

Zanutto, E. & Zaslavsky, A. (2002). Using administrative records to impute for nonresponse. *Survey Nonresponse.* New York: John Wiley and Sons, 403–415.

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