DEMOGRAPHIC RESEARCH
A peer-reviewed, open-access journal of population sciences

DEMOGRAPHIC RESEARCH

VOLUME 47, ARTICLE 27, PAGES 843–882
PUBLISHED 6 DECEMBER 2022
https://www.demographic-research.org/Volumes/Vol47/27/
DOI: 10.4054/DemRes.2022.47.27

Research Article

Small-area estimates from consumer trace data

Arthur Acolin

Ari Decter-Frain

Matt Hall

Contents

Demographic Research: Volume 47, Article 27
Research Article

Small-area estimates from consumer trace data

Arthur Acolin[1]
Ari Decter-Frain[2]
Matt Hall[2]

Abstract

BACKGROUND
Timely, accurate, and precise demographic estimates at various levels of geography are crucial for planning, policymaking, and analysis. In the United States, data from the decennial census and annual American Community Survey (ACS) serve as the main sources for subnational demographic estimates. While estimates derived from these sources are widely regarded as accurate, their timeliness is limited and variability sizable for small geographic units like towns and neighborhoods.

OBJECTIVE
This paper investigates the potential for using nonrepresentative consumer trace data assembled by commercial vendors to produce valid and timely estimates. We focus on data purchased from Data Axle, which contains the names and addresses of over 150 million Americans annually.

METHODS
We identify the predictors of over- and undercounts of households as measured with consumer trace data and compare a range of calibration approaches to assess the extent to which systematic errors in the data can be adjusted for over time. We also demonstrate the utility of the data for predicting contemporaneous (nowcasting) tract-level household counts in the 2020 Decennial Census.

RESULTS
We find that adjusted counts at the county, ZIP Code Tabulation Areas (ZCTA), and tract levels deviate from ACS survey-based estimates by an amount roughly equivalent to the ACS margins of error. Machine-learning methods perform best for calibration of county- and tract-level data. The estimates are stable over time and across regions of the country. We also find that when doing nowcasts, incorporating Data Axle estimates improved prediction bias relative to using the most recent ACS five-year estimates alone.

[1] University of Washington, USA. Email: acolin@uw.edu.
[2] Cornell University, USA.

CONTRIBUTION

Despite its affordability and timeliness compared to survey-based measures, consumer trace data remains underexplored by demographers. This paper examines one consumer trace data source and demonstrates that challenges with representativeness can be overcome to produce household estimates that align with survey-based estimates and improve demographic forecasts. At the same time, the analysis also underscores the need for researchers to examine the limits of the data carefully before using them for specific applications.

1. Introduction

Demographic research, planning, and policymaking rely on having demographic estimates that are not only accurate, but that are timely, granular in scale, and statistically precise. Despite large-scale governmental efforts to produce such data, the availability of small-area estimates is limited by costly survey operations and data processing time (Bryan 2004). In the United States, the Census Bureau serves as the main source for demographic data, with the decennial census enumerating the full population every ten years, and the American Community Survey (ACS) providing sample-based estimates annually. While decennial data is as close to the ground-truth population as we can get, it can quickly become dated and not reflective of the current situation, particularly in areas undergoing rapid change. The ACS, on the other hand, provides estimates annually, but because it relies on sampling, it is able only to provide annual estimates for the largest areas with small-area estimates produced over five-year rolling periods and containing substantial uncertainty.

The need for timely, accurate, and precise small-area demographic estimates has long been of crucial importance. These data are needed to make effective planning and decision-making for school and transportation systems, caseload management, housing and commercial development, and other applications. They are also needed for a wide array of research applications, either as a source for studying population change or in establishing population baselines. These needs are particularly important in the context of rapidly changing populations, such as those where gentrification or displacement from natural disaster occur. They are also important in the context of short-term shifts in populations that may alter the incidence of particular events, including temporary migration during the COVID-19 pandemic that impacted rate denominators.

The ability to produce small-area estimates based on sources other than publicly available decennial census and ACS data is an area that deserves particular attention given new disclosure avoidance systems implemented in the 2020 Census that impact the accuracy of small-area estimates by injecting noise in the estimates. The Census Bureau's concern for the privacy of citizens has led to differential privacy standards that may limit the usability of small-area estimates for many research and policy applications (e.g., Hauer and Santos-Lozada 2021; Winkler et al. 2021; Ruggles et al. 2019).

In light of these issues, alternative data sources have growing appeal for overcoming the challenges associated with traditional sources. Data from credit records, social media, internet activity, and consumer traces have emerged as alternatives for constructing demographic estimates, often with vast improvements in temporal and spatial frequency (Billari and Zagheni 2017). Yet, the potential advantages of these data are potentially offset by sizable risks associated with the nonrepresentative aspects arising from the non-research-related purpose of the underlying data (Salganik 2019). Whereas the Census Bureau's annual measures of annual households are imprecise but mostly unbiased (being available for only large counties), consumer data are highly precise, but with some bias not known a priori highlighting the trade-off between variance and bias in producing small-area estimates.

This paper considers the potential of consumer trace data purchased from Data Axle (previously known as Infogroup) for producing small-area population estimates with high precision, high accuracy, annual frequency, and at a cost ranging in the low five figures that makes it accessible to purchase for a range of research teams. Consumer trace data are constructed by commercial vendors primarily for marketing purposes. The data contain near-exhaustive lists of US households, including current addresses and counts of household members. Companies assemble and maintain the data through proprietary processes, sometimes involving hundreds of sources, which are effectively "black-boxes" to researchers (Pasek et al. 2014). Nonetheless, the data hold tremendous promise due to their massive scale.

We seek to address the aforementioned risks associated with data of this kind. We focus on three questions: (1) How biased are estimates of households from Data Axle compared to survey-based estimates?; (2) Can biases in the coverage of consumer data be mitigated by calibrating against the decennial census?; and (3) Can more timely consumer trace data be used to produce nowcasts (prediction of current estimates) of population levels that improve on traditional forecasts?

To answer these questions, we apply a series of statistical tools that leverage state-of-the-art techniques in demographic analysis and computational social science. Moving beyond related work, we test a variety of different modeling approaches for calibration. We find that Lasso regression, a machine-learning approach that finds the best parameters to include in a linear specification, substantially outperforms various fixed specifications at adjusting tract-level household counts. This highlights the value of machine learning for calibrating nontraditional data in demographic analysis since the idiosyncrasies of these data are often unknown to researchers.

Our findings indicate that raw household estimates based on consumer data compared to survey-based estimates have mean absolute proportional error (MAPE) of 8% at the county level and 18% and 17% at the zip code and census tract levels, respectively. We compare a range of linear calibration models to the data (Billari and Zagheni 2017; Zagheni, Weber, and Gummadi 2017; Alexander, Polimis, and Zagheni 2020), the best performing of which reduce MAPE to 15% for zip codes and 11% for tracts. These deviations appear reasonable in comparison to the margins of error for the ACS five-year estimates (on average, 13% for ZIP Code Tabulation Areas (ZCTAs) and 7% for tracts). We therefore argue that the data are, on the whole, well aligned with existing survey-based estimates. We recommend using the data with caution, however, given that (a) over- and undercounts are associated with neighborhood characteristics, and (b) we found a substantial, undocumented, year-over-year shift in the data that appears to be an artifact of the data construction process.

The next section reviews the literature on small-area demographic estimates and the emerging use of non-survey data to derive estimates. Section 3 presents the consumer data used in the analysis. Section 4 reports demographic estimates at various levels of geography and examines neighborhood characteristics associated with under- or overestimations. Section 5 presents the adjustments based on linear calibration models, showing substantial improvements in the estimates fit to the ACS. Finally, section 6 concludes on the potential and limits of consumer data to producing small-area demographic estimates and the value of using adjustment technique to make them more robust for using the estimates for research purposes.

2. Background

Existing small-area demographic methods have traditionally relied on the production of reliable estimates based on surveys with limited sample sizes or available only on a decennial basis (decennial census) (Pfefferman 2013; Rao 2003). In the United States, demographic data are sourced almost exclusively from the Census Bureau’s large-scale data efforts. These include the decennial full population enumerations and large national surveys (e.g., the ACS and Current Population Survey). These data are widely regarded and evaluated as the most accurate representation of the US population (Alexander 2002). However, population-level data are produced only every ten years. The development of the ACS in the first decade of the 21st century substantially expanded the availability of small-area estimates between decennial censuses, but despite the large scale of the ACS, the precision of sample-based annual estimates is limited by their limited size for small geographic areas. The ACS collects interviews for about 1.6% of US households annually. While the final sample is sufficiently large for large areas (states and metropolitan regions), samples are comparatively small for small areas. Accordingly, annual estimates are available only for areas with populations above 65,000 residents, while estimates for smaller units – like counties and census tracts – are smoothed over a five-year period and have a substantial level of uncertainty (Alexander 2002; Folch et al. 2016).

The relatively low frequency of data collection and limited availability for small areas of existing survey-based demographic data pose challenges for applications to the analysis of situations where underlying populations are rapidly changing, such as neighborhoods undergoing rapid change or turnover (Schnake-Mahl et al. 2020). Increasingly social science research is relying on highly temporal data that allow data to be represented in near real time (Li 2021; Chetty et al. 2020; Pennington 2020). In addition, alternative data sources are not limited by the constraints of administrative boundaries – such as counties or census tracts – that official statistics are typically expressed in. Having the flexibility to deviate from these boundaries would allow researchers and planners to understand population dynamics in continuous space (e.g., to align with flood zones, forest fire spread) or in isolated areas (e.g., in response to building demolition). Temporally and spatially precise demographic data have extensive application potential: examining the link between new construction and a range of outcomes from health to economic mobility (Li 2021; Asquith et al. 2021; Damiano and Frenier 2020; Pennington 2020), the relationship between transit investments and surrounding population change (Boarnet et al. 2018), or the impact of zoning changes on population composition (Freemark 2020). In addition, high frequency, spatially precise population data can be used as denominators for temporally variant event rates, such as with criminal activity or disease contagion (though the use of mixed numerators and denominators can create major complications; see, e.g., Lopez-Vizcaıno, Lombardıa, and Morales 2013).

Demographic estimates based on nontraditional data have been the object of recent but sustained interest to address the shortcomings of traditional data sources (Alburez-Gutierrez et al. 2019; Alexander, Polimis, and Zagheni 2020; Billari and Zagheni 2017; Cesare et al. 2018; Lee and Van der Klauw 2010; Penner and Dodge 2019; Williams et al. 2015; Zagheni et al. 2018; Zagheni et al. 2014; Zagheni and Weber 2012). These nontraditional data sources are generally not collected through surveys for the purpose of creating population estimates. Rather, they result from activities in administrative databases (Penner and Dodge 2019), geo-located phone records (Williams et al. 2015), or digital traces of online activity (Alexander, Polimis, and Zagheni 2020; Zagheni et al. 2018; Zagheni et al. 2014; Zagheni and Weber 2012).

Consumer data are one such type of nontraditional trace data that provide a new and potentially valuable source of timely and precise demographic estimates. The data are constructed via proprietary processes for aggregating traces left by consumers when using credit cards and loyalty programs, enrolling in public or private services, renewing subscriptions, identifying billing addresses on purchases, registering vehicles, or filing change-of-address information (Phillips 2020). These traces accumulate on a recurrent basis, allowing for continual updates of the resulting data and thus the potential for timely population estimation. Recent work has used consumer data to measure housing stability (Phillips 2020). In theory, the providers of this type of data have strong incentives for correctly identifying consumers' residential locations. However, consumer data – like other nontraditional sources – are not designed to be representative of populations and have known and presumed coverage biases.[3]

The development of demographic measures based on nontraditional data has been limited by the biases inherent to these data sources and concerns about the validity of the estimates produced (Alburez-Gutierrez et al. 2019; Cesare et al. 2018; Phillips 2020; Zagheni et al. 2018). As a result, the potential of these data to generate demographic estimates to complement existing survey-based measures has not been fully realized. Nontraditional data sources, including the data proposed for use in this project, all have biases in terms of coverage and representativeness (e.g., underrepresentation of those with limited credit or online activity) requiring statistical adjustments for valid population estimates (DeWaard, Johnson, and Whitaker 2019; Dragan, Ellen, and Glied 2020; Lee and Van der Klaauw 2010; Phillips 2020). While nontraditional data sources are prone to those biases, consumer data benefit from being aggregated from different data sources and have plausibly wider reach than sources such as digital trace data that are generally drawn from smaller nonrandom subsets of populations.

A commonality in the barrier to using nontraditional data for demographic estimation is that the data-generation process was not designed to produce data representative of the entire population. Accordingly, a necessary task when assessing the potential of any nontraditional source is to document and minimize the deviations between it and well-established survey-based estimates (Alexander, Polimis, and Zagheni 2020; Dever, Rafferty, and Valliant 2008; DeWaard, Johnson, and Whitaker 2019; Penner and Dodge 2019; Zagheni et al. 2018; Zagheni and Weber 2012). In this paper we conduct such an exercise. We first document the coverage rates between small-area population estimates derived from consumer data and census data. Next, we show that over- and undercounts are systematically associated with a wide range of neighborhood characteristics. Finally, we compare a series of models for producing adjusted estimates that account for the association between errors and neighborhood characteristics. Our approach to calibration follows previous research in relying on linear models (Zagheni et al. 2018; Alexander, Polimis, and Zagheni 2020) and pushes further by considering alternative specifications, including introducing penalty terms for parameter selection.

[3] Previous research also finds that the demographic characteristics reported in consumer data differ substantially from self-reports (Pasek et al. 2014).

3. Consumer data

This analysis uses historical consumer records aggregated by Data Axle (previously Infogroup), a private-sector marketing data provider. The Data Axle data – like other sources of consumer data – are derived from a variety of sources, including loyalty programs, credit records, utility records, voter registrations, real estate tax assessments and deed transfers, public records (bankruptcies, licenses, and registrations), and mailing address changes. Table 1 lists the key sources as disclosed by Data Axle.

Table 1: Data sources for Data Axle database

Sources	Nature
Utility connections and changes	Public
Change-of-address notifications	Private
Real estate tax assessments and deed transfers	Public
Voter registrations (where available for marketing applications)	Public
Credit card billing statements	Private
Public records, such as bankruptcies, pilot licenses, hunting licenses, and boat registrations	Public
Telephone white page directories	Private
Newspaper and magazine subscription lists	Private

Note: Data Axle indicates that it has over 100 contributing sources but does not list them all.

Data Axle data are widely used by businesses, with 20% of Fortune 500 as customers, according to its website. The data serve the main purpose of identifying consumer locations for targeted marketing, administrative relations (e.g., follow-up contact or bill collection), and other commercial communication. To meet this need, Data Axle has a strong incentive in ensuring that residential information is current and accurate. In other words, a crucial feature of the data is the ability to find households' residential addresses in real time, providing a critical asset for demographic data, which is often hobbled by the lack of accurate, timely, and wide-ranging small-area data. It is similar to other consumer datasets, such as Infutor and LexisNexis, that have also been used in academic research (Diamond, McQuade, and Qian 2019; Ling et al. 2019; Greenlee 2019; Phillips 2020).

The full Data Axle dataset is provided as a series of annual snapshots between 2006 and 2019 at the 'consumer-unit' (i.e., family) level. The number of records increases from 130 million in 2006 to 175 million in 2019. In addition to a time-constant consumer-unit ID, the dataset includes geo-coded addresses, as well as directly sourced, inferred, or imputed demographic characteristics, such as race, family size, marital status, presence and number of children, and householder age.

We aggregate the consumer data to produce household-like measures based on individual dwelling unit information. This allows for the Data Axle data to be as similar as possible to the US Census Bureau's definition of households as all the people who occupy a housing unit (US Census 2015). This count of households can be produced for any level of spatial aggregation on an annual basis, and in this paper, we report results for counties, ZCTAs, and census tract (see Appendix 1).

A limit of these data is the limited transparency on how they are collected and on their level of accuracy, which requires substantial validation before using them for research projects (Kennel and Li 2009; Phillips 2020). For instance, the documentation does not list individual sources for each year or how signals from different sources are prioritized. In addition, changes to the data-generation process can affect estimates based on this data. For example, in 2013–2014, Data Axle incorporated data from new sources, including through a merger, but didn't fully deduplicate individual observations. As a result, observation counts for these two years are substantially higher than for either 2012 or 2015. Data Axle provided the explanation for this data bump after being contacted, but it was not part of the metadata. This highlights a limit of these data sources that are subject to variability in quality and coverage that are not always immediately visible to data users or disclosed in the documentation.[4]

4. Coverage

The first step in establishing the usability of the consumer data to produce demographic estimates is to compare estimates from Data Axle to estimates from the Census Bureau for different levels of geography and time periods. In the second step, we explore sociodemographic and geographical factors that could be associated with divergent estimates.

[4] There are also important ethical considerations given the sensitivity of the data points made available. Phillips (2020: 1340) argues that "given their availability to customers outside research, these data may be viewed in discussion of ethics as public information." However, risks of disclosure of personally identifiable information should still be considered, and all representation of the data should be aggregated at a sufficient level to preserve anonymity.

4.1 Description of over- and undercounts: Spatially and over time

Figure 1 displays the coverage ratio calculation by dividing county-level Data Axle estimate averages over a five-year period with equivalent estimates from five-year ACS estimates over the same period for both 2006–2010 (2010 estimate on top panel) and 2014–2018 (2018 estimate on bottom panel).[5] The maps show that in most cases the Data Axle estimates are between 90% and 110% of the ACS estimates. The share of estimates within that 90% to 110% range increases from 61% to 71% between 2010 and 2018, and the share within an 80% to 120% range reaches 90% in 2010 and 94% in 2018. This indicates that at the county level, the Data Axle household estimates generally are closely related to the ACS counts and that overtime the Data Axle estimates have generally become closer.

The maps also show some level of spatial correlation in areas in which the Data Axle estimates are substantially above or under those of the ACS and that these areas appear to be relatively stable over time. For instance in both 2010 and 2018, one can observe clusters of counties with substantially lower Data Axle estimates in central California, New Mexico, and southwest Texas. On the other hand, some clusters of counties in Michigan, New England, the Florida panhandle, and Utah have Data Axle estimates substantially higher than the ACS estimates. This relative stability in the differences will be an important feature to be able to produce adjusted Data Axle estimates that better match the ACS estimates.

The Data Axle data also provide an approximation of households at the ZCTA level, as exemplified with Figure 2, which shows the coverage ratio in 2011 and 2018 in four major US metros: New York (upper left), Los Angeles (upper right), Chicago (lower left), and Philadelphia (lower right). The maps show that at this small level of geography (the median ZCTA has about 1,000 households), Data Axle estimates are mostly within 90% to 110% of the ACS estimates, with overall improvements between 2011 and 2018 in these four metropolitan regions. Of note, the share of Data Axle ZCTA estimates within the 90% to 110% range reaches 61% by 2018 with 82% within the 80% to 120% range. To put this in context, about 50% of the ACS ZCTA estimates have margin of errors that are larger than 10%, and 27% have margin of errors that are larger than 20%. Because we are relying on ACS data for comparison, we produced five-year estimates, but given the nature of the data, the Data Axle estimates can be produced annually, creating more temporarily precise recent estimates without loss of precision.

---

[5] In this section we present the five-year ACS data rather than the decennial census data as the benchmark as it is currently the main source of small-area estimates with relatively high temporal variation, allowing us to compare the estimates over time. In addition, the ACS includes a rich set of sociodemographic variables we can use to explore local characteristics associated with over- or undercounts in the Data Axle estimates. In Appendix 2010 Census Estimates we report comparisons of the 2010 Census estimates to the 2010 Data Axle estimates.

Figure 1: County Data Axle to ACS five-year estimates coverage ratio, 2010 and 2018

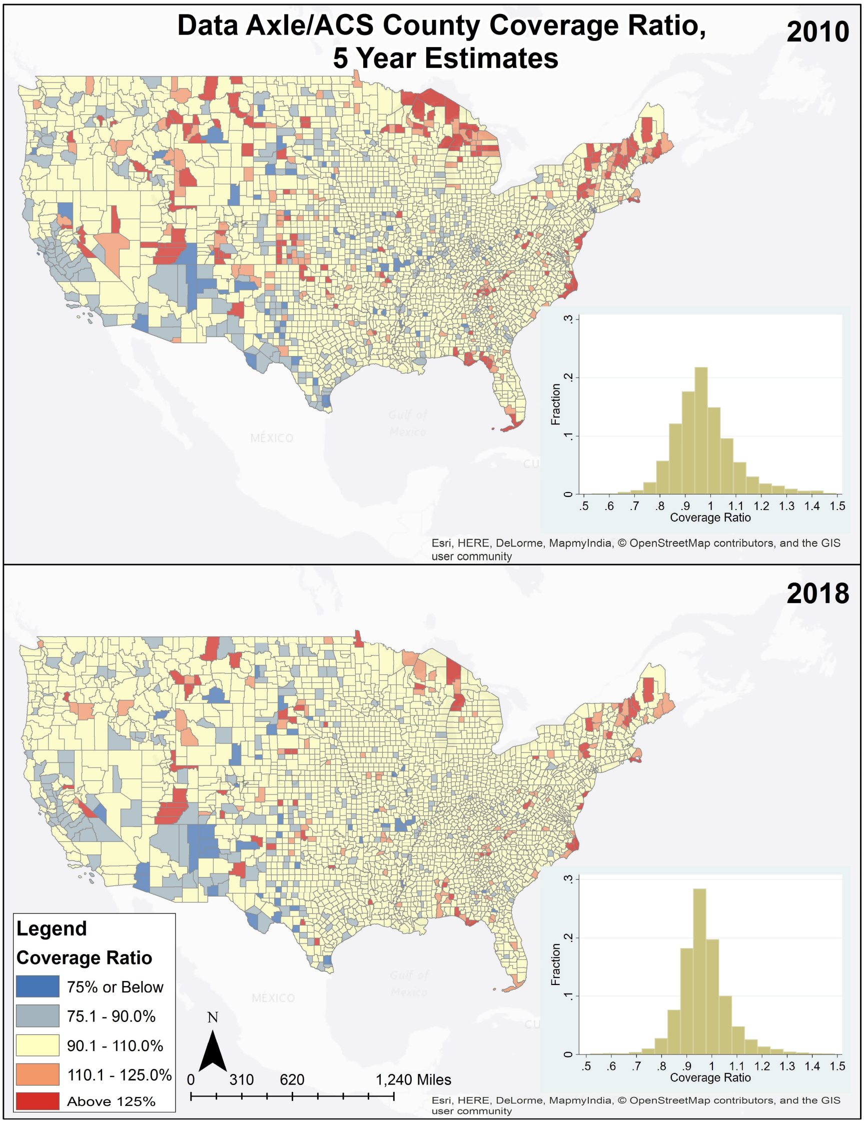
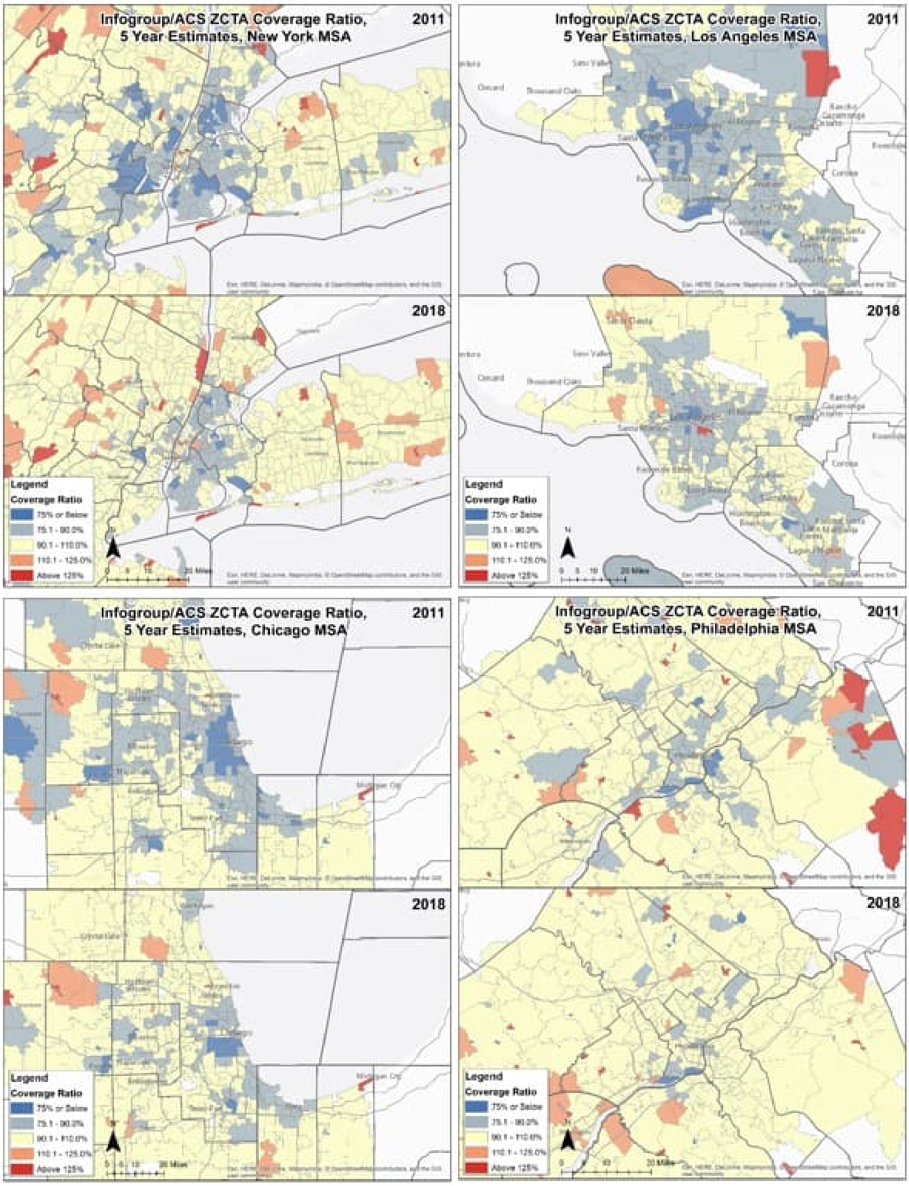


Figure 2: ZCTA Data Axle to ACS five-year estimates coverage ratio, 2011 and 2018



Similar to what we observed nationwide in Figure 1, there also appears to be some level of spatial correlation with certain areas having persistently lower ratios of Data Axle to ACS estimates. Some of the central locations in all four metropolitan regions exhibit some level of undercounting, particularly lower income areas. Conversely, some suburban and exurban areas show consistent overcounts. Another feature is that there appears to be variations in the closeness of the estimates across metropolitan regions with Philadelphia showing overall closer estimates, while estimates in Los Angeles in 2011 show particularly pronounced and widespread undercounts that improve but remain quite widespread by 2018.

4.2 Sociodemographic characteristics associated with over and undercounts

To describe local characteristics that are associated with higher likelihoods of under- and overcounts, we use multinomial logits to examine three outcomes: undercoverage (Data Axle estimates <90% of ACS estimates), overcoverage (>110%), and aligned coverage (between 90% and 110%). Selected covariates are based on characteristics assumed to be related to upward or downward biases resulting from household characteristics and the built environment.[6]

Some of the variables are expected to lead to undercounts in the Data Axle data due to thinner consumer profiles: younger households, those with lower incomes, households living in group quarters, and those who have moved recently. Other characteristics may lead to overcounts as a result of misrepresentations of household residential locations (e.g., owners of vacant rental units, vacation properties assigned to multiple profiles). Other factors should increase the chances of ACS and Data Axle estimates being aligned, including higher incomes, ownership rates, older population, and single-family homes. The size of the ACS household count margin of error is expected to be associated with higher likelihoods of substantial under- or overcounts as these estimates are more imprecise. To account for potential differences in coverage across population groups, we additionally account for race/ethnicity and nativity.

Table 2 reports the odd ratios of the results of the multinomial models for the county and ZCTAs estimates using five-year ACS data from the 2011–2018 period. A larger margin of error in the ACS estimates and higher share of young adults (18 to 24) are associated with a higher likelihood of both under- and overcounts (∪ relationship). A

[6] We adopted the modeling decision to consider the ACS data to represent the true household estimates. However, it is possible that the Data Axle estimates may diverge from the ACS estimates while being closer to the true household estimates. This could happen if Data Axle does a better job at capturing some hard-to-count populations than the census (households with unauthorized persons or those in multifamily units) or if it is more up-to-date than the census master address file in capturing new constructions in rapidly growing areas.

higher unemployment rate is associated with a lower likelihood of under- and overcounts (∩ relationship). This means that the Data Axle estimates are more likely to be substantially above or below the ACS estimates in areas with less precise ACS estimates, a larger share of young adults, and where unemployment is elevated.

Table 2: Area characteristics associated with Data Axle over- and undercounts, county and ZCTA, 2011–2018

Coverage category (ref. =.9-1.1)	County Less than 0.9	 1.1 or above	ZCTA Less than 0.9	 1.1 or above
Household count MOE (%)	1.127	1.077	1.047	1.065
	(13.10)	(7.60)	(83.28)	(122.45)
Log median household income	0.088	0.992	0.333	1.162
	(–19.95)	(–0.05)	(–52.21)	(7.47)
Unemployment (%)	0.955	0.940	0.986	0.993
	(–6.50)	(–6.82)	(–12.46)	(–6.85)
Group quarter (%)	0.912	1.050	0.968	1.014
	(–12.18)	(5.32)	(–26.30)	(12.25)
Own (%)	0.960	1.012	0.981	1.001
	(–9.85)	(2.20)	(–29.74)	(1.07)
Female (%)	0.907	1.083	0.984	1.002
	(–7.35)	(4.90)	(–11.31)	(1.48)
Age 18–24 (%)	1.044	1.031	1.020	1.016
	(5.66)	(2.54)	(14.55)	(11.68)
Under 18 (%)	0.989	1.026	0.991	1.009
	(–1.17)	(2.18)	(–7.44)	(8.15)
65 plus (%)	1.002	1.003	1.001	1.009
	(0.28)	(0.28)	(0.92)	(8.87)
Moved (%)	1.055	0.934	1.010	0.999
	(9.28)	(–8.31)	(10.40)	(–1.12)
Vacant housing units (%)	0.956	1.120	0.990	1.048
	(–13.84)	(39.05)	(–16.10)	(100.19)
Housing stock 3+ units (%)	1.040	1.011	1.015	1.000
	(9.47)	(1.79)	(24.13)	(–0.25)
Black or African American (%)	0.987	0.990	0.999	0.989
	(–7.59)	(–4.53)	(–2.44)	(–24.00)
Hispanic or Latinx (%)	1.023	0.966	1.023	0.994
	(11.63)	(–8.74)	(41.24)	(–9.02)
Asian (%)	1.078	0.918	1.037	0.983
	(7.95)	(–3.72)	(25.95)	(–7.57)
Native American (%)	1.041	0.994	1.026	0.995
	(13.55)	(–1.25)	(38.35)	(–6.49)
Foreign born (%)	0.999	1.047	0.996	0.998
	(–0.20)	(4.26)	(–3.24)	(–1.47)
N	25,129		250,999	
Year FE	Yes		Yes	
County FE	No		No	
pseudo R-sq	0.24		0.20	

Notes: Exponentiated coefficients; *t* statistics in parentheses. All control variables come from the five-year ACS estimates for that time period. MOE = margins of error.

The other variables either exhibit more linear relationships or only affect the chance of either over- or undercounts. The share of residents in group quarters, shares under 18, vacant units, and median household incomes are associated with a lower likelihood of undercount and higher likelihood of overcount. Conversely, a higher share of recent movers is associated with a higher likelihood of undercount and lower likelihood of overcount. For instance, the share of homeowners and the share of female households are associated with a lower likelihood of undercount but not substantially higher likelihood of overcount.

5. Adjusted coverage

Next, we assess the extent to which households based on Data Axle data can be statistically adjusted to account for the systematic errors in coverage identified in the previous section. We focus on whether these adjustments persist throughout the years between each decennial census because we think Data Axle's annual frequency during these years is one of its important advantages.

We generate adjusted households for each year between 2006 and 2018 through the following two-step procedure: First, using only data from 2010, we fit models where Data Axle's households are used as predictors of households from the decennial census. Next, we input Data Axle's households from 2006 through 2018 to the fitted models to generate predictions.

These predictions function as adjusted counts. We take this approach because 2010 is the only year in the time period for which the Census Bureau, via the decennial census, provides a ground-truth annual household estimates for small geographic regions like tracts and ZCTAs. For these smaller regions, the ACS does not yield one-year estimates we can use as outcomes in a model. We evaluate the performance of different adjustment models by measuring how much they improve overcounts based on the raw data in all the other years from 2006 through 2018.

This approach is similar to standard practice for adjusting nontraditional data (Alexander, Polimis, and Zagheni 2020; Zagheni et al. 2018) but differs in two ways. First, whereas previous work evaluates adjustments by projecting out one or two years, we evaluate 'out-of-sample' performance over the entire period of 2006 through 2018 (minus 2010). We believe this is a strong test of how effectively household estimates generated from Data Axle can be adjusted to match survey-based estimates. Because most years are tested out of sample, performance depends heavily on the stability of deviations between Data Axle and the census. Second, our approach involves comparing the performance of multiple model specifications, which we describe next.

Our objective is to fit a model that minimizes deviations from representative survey-based estimates produced by the Census Bureau, which we take here as the ground truth. Across all the models we compare annual census households p^{CE} in geographic region i at time t to be a function of equivalent counts based on Data Axle data p^{DA} and a vector of local demographic characteristics, which we denote $\mathbf{x_i}$ (see Table A-1 in Appendix 6 for list of covariates included). All the models we compare constitute a variation on the following linear structure:

$$\log p_{CEit} = \alpha_0 + \alpha_1 \log p_{DAit} + \beta_0 \mathbf{x_i} + \varepsilon_{it} \quad (1).$$

We build up a linear specification by first running a bivariate regression of p^{CE} on p^{DA}, then successively adding state fixed effects and local characteristics. This model corresponds roughly to others in the literature (Alexander, Polimis, and Zagheni 2020; Zagheni et al. 2018). Next, we add more complexity with a multilevel model that allows α_0 and α_1 to vary at the state level (county level when modeling tracts). Finally, we consider a Lasso penalized regression, where we include the fully interacted set of state dummies and local characteristics as possible parameters (Molina and Garip 2019; Belloni, Chernozhukov, and Hansen 2014; Varian 2014). Appendixes 8 and 9 provide detailed descriptions of how we implement the multilevel and Lasso models.

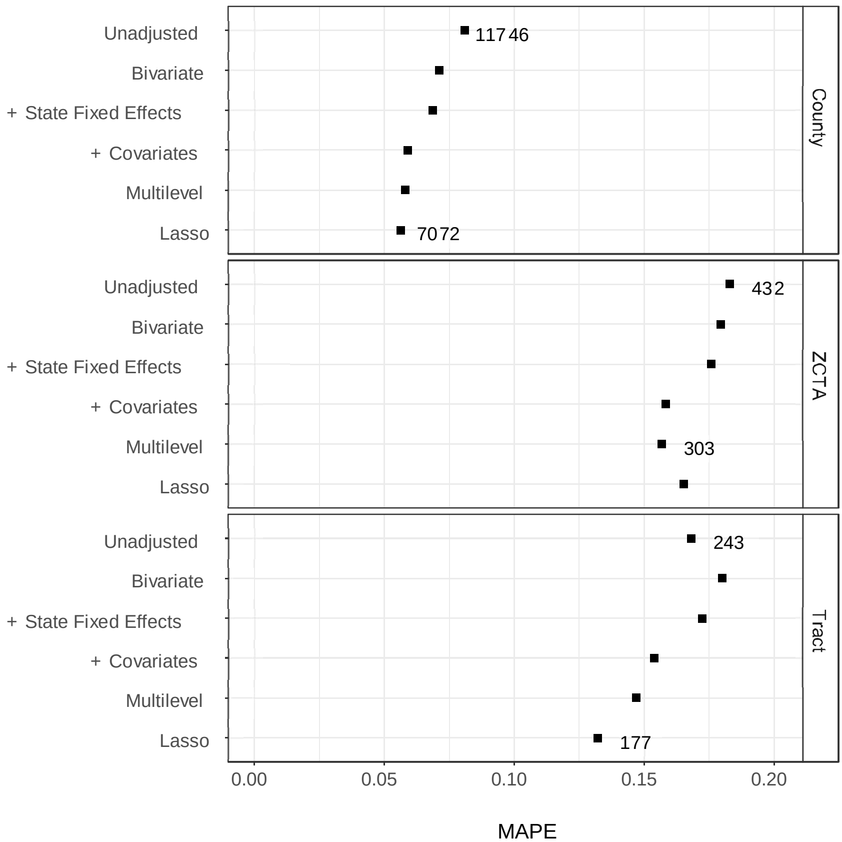
We assess the performance of the adjusted estimates by comparing them to the best available equivalents produced by the Census Bureau in each year for which we generated estimates. For counties, we directly compare adjusted estimates to annual household estimates available from the ACS for the 773 largest counties in the country. For tracts and ZCTAs, we compute a five-year rolling average of Data Axle–based estimates and compare these against the ACS five-year estimates. To quantify errors we follow previous work (Zagheni and Weber, and Gummadi 2017) in computing mean absolute proportional error (MAPE):

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{|p_i - \hat{p}_i|}{p_i},$$

where p_i is the estimate derived from census surveys, $\hat{p}_i$ is the adjusted estimate using Data Axle, and N is the total number of geographic regions.

Figure 3 presents the results of fitting models in 2010 using decennial data, projecting it to all other years, then computing the MAPE for each year and taking the average over all years. The first row of each panel shows the MAPE comparing unadjusted households from Data Axle to the ACS equivalent. Counties have the smallest raw proportional error, with an overall MAPE of 8% (or 12,015 households in absolute terms) of the ACS. ZCTAs and tracts have higher proportional errors of 17% (439 households) and 18% (245 households), respectively. The Lasso model performs best at reducing errors at the county level, reducing MAPE to 5.6% (7,062 households). The multilevel model performs best for ZCTAs, reducing the MAPE to 15.7% (303 households). Finally, the Lasso yields the most improvement for tracts, reducing the MAPE to 13.2% (177 households).

Figure 3: Average errors throughout the entire period 2006–2018 using raw unadjusted Data Axle counts and adjusted counts based on various calibration models

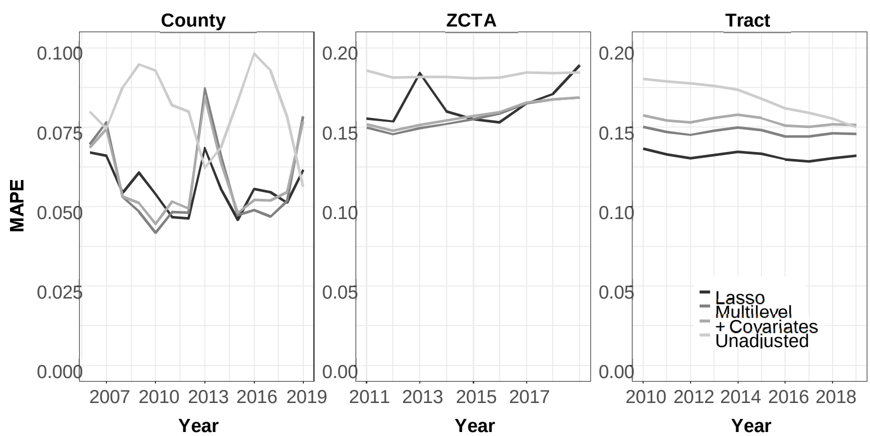


Notes: 'Unadjusted' refers to raw counts. The remaining models are described in the main text and in Appendix 7. The numbers beside the points indicate to the mean absolute errors for the unadjusted and best-performing models for each geography. They are meant to give a sense of the absolute magnitude of the deviations. All errors are based on predictions made out of sample in time, meaning that models were fit using household measures from the 2010 Decennial Census data only and evaluated based on their alignment with the ACS in every other year.

The magnitude of the errors, particularly after adjustment, are of similar magnitude to the uncertainty in the ACS estimates themselves. The Census Bureau describes uncertainty with margins of error, which represent 90% confidence intervals around ACS household estimates. For households at the tract level, the average ratio of margins of error to the count itself is 7%, roughly 50% smaller than the errors from the best-performing adjustment. For ZCTAs, the average ratio of margins of error to estimates is 13%, roughly 10% smaller than the prediction errors from the best-calibrated model. In other words, the deviations between Data Axle and the ACS are similar in size to the uncertainty inherent in ACS estimates, and adjusted estimates compare particularly favorably. We view this as an encouraging sign for the potential utility of consumer trace data for demographic estimation.

Figure 4 shows how the deviations between Data Axle and ACS vary over time and by calibration model. For tracts and ZCTAs, the results represented by a particular year are based on that year and the previous four. For example, the 2010 point for tracts is based on comparing the ACS five-year household estimates based on 2006–2010 data against an average of Data Axle household estimates over the same period. For tracts, errors remain fairly stable over time, with the Lasso outperforming the others in every year. For ZCTAs, adjusted errors grow over time, although calibration remains beneficial. Counties have more variation, with the Lasso performing worse than the multilevel and covariate models in 2009 and 2010 while outperforming them in 2013.

Figure 4: Prediction errors for different adjusted household estimates by year



Notes: For counties, large counties are compared each year. For ZCTAs and tracts, five-year averages of adjusted estimates are compared against the ACS five-year averages for all units. All errors are based on predictions made out of sample in time, meaning that models were fit using household measures from the 2010 Decennial Census data only and evaluated based on their alignment with the ACS data in every other year.

The time trends also expose the effect of the 2013 ‘data shock’ described in section 3. At the county level, households based on raw Data Axle data abruptly improve as a predictor of the census in 2013 before getting worse again in subsequent years. As expected given that the calibrations are time invariant, all the calibrated counts perform worse in 2013, with the Lasso performing least poorly. For ZCTAs, only the Lasso appears affected by the shock, and the change is not apparent for tracts. The shock is likely less apparent for smaller geographies because the extra households in the data are spread over tens of thousands of units and further smoothed by five-year averaging.

It is notable that the Lasso model appears least sensitive to the 2013–2014 data shock at the county level but most sensitive at the ZCTA level. We expect this is due to differences in the complexity of the models. As described in Appendix 7, the most complex models aside from the Lasso are the multilevel models, which each have 124 parameters when fit for tracts and ZCTAs. The Lasso model used for counties has only 36 parameters while the ZCTA Lasso model has 394.

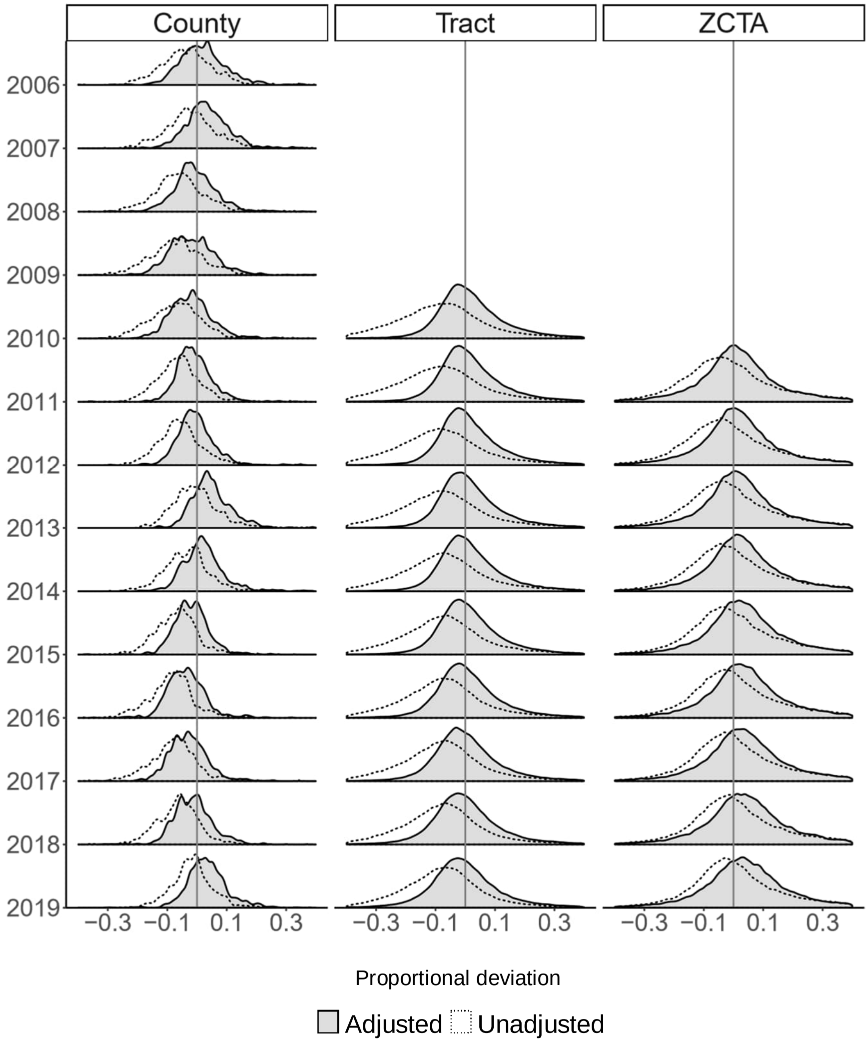
The parameters in these models act as shifters that adjust Data Axle–based household estimates away from the raw counts. Thus, it seems reasonable that the Lasso performs poorly in response to the shock at the ZCTA level, where it is highly parameterized and performs relatively well when more conservatively parameterized at the county level.

Figure 5 shows kernel density plots of the proportional deviations for each geography and year. The dashed line shows the density of deviations before any adjustment, while the gray shaded region shows the deviations for the best-performing model at each geographic level. For tracts and ZCTAs, adjustment shifts the distribution positively so that the mode and the bulk of the density come close to zero. For counties, the adjusted estimates appear to improve things in most years. The exception is 2013, where deviations from the unadjusted estimates become centered around zero and the adjusted estimates become upwardly biased.

We also performed an out-of-sample evaluation, in which we fit the same set of calibration models to 80% of tracts in the 2010 Decennial Census and evaluated their performance on the other 20%. We present the results of this evaluation in Appendix 10, which are generally consistent with the out-of-time evaluation results presented above.

Furthermore, it is often desirable for small-area estimates to satisfy the benchmarking property, that when aggregated they should align with estimates for larger areas (Bell, Datta, and Ghosh 2013). In Appendix 5, we show that tract-level adjusted estimates from Data Axle largely exhibit this property. We sum-adjusted tract-level household counts to the county level and show that these sums align more closely to the ACS data than adjusted county-level estimates presented above.

Figure 5: Density plots for the distribution of deviations between Data Axle and census population estimates



Notes: The deviations for the best-performing model at each geographic level are shaded in gray. The dashed line shows deviations for the raw data. The deviations are computed by subtracting census from Data Axle households. All errors are based on predictions made out of sample in time, meaning that models were fit using households from the 2010 Decennial Census data only and evaluated based on their alignment with the ACS data in every other year.

We also explored the parameters that contribute most to each model. In Appendix 12 we present estimates for $alpha_0$, $alpha_1$, and the top 10 largest estimated coefficients of the five models at each geographic level. The coefficients from the Lasso models at ZCTA and tract levels suggest that certain characteristics may be important in certain states. For instance, in Hawaii and Washington, DC, tracts with high native proportions are undercounted in Data Axle, while tracts with high native proportions are overcounted in Delaware and Maryland.

The results of this adjustment exercise suggest that fitting calibration models to consumer trace data can help reduce the nonrepresentative aspects of the resulting estimates. Calibration appears particularly beneficial when constructing estimates for census tracts (i.e., neighborhoods).

6. Nowcasting

Finally, we provide an example of how using households from Data Axle can improve demographic research. We consider the task of nowcasting small-area estimates. Our goal is to predict tract-level households as measured in the 2020 Decennial Census (cross-walked back to 2010 tract geography), using data from the 2009–2019 ACS five-year tables and annual adjusted counts from Data Axle, as derived in the previous section. This maps to an applied situation in which the ACS five-year data have not yet been released, or a researcher requires a year-specific estimate that does not rely on data collected over a five-year period. We ask to what extent Data Axle data can aid in generating accurate small-area estimation nowcasts.

Previous examples of combining survey and non-survey data for nowcasting have used a Bayesian framework to combine observations from multiple sources into a single time series while accounting for the different variances associated with each source (Alexander, Polimis, and Zagheni 2020; Sakshaug et al. 2019). This approach is beneficial in cases where the two sources of data have limited temporal overlap and must be combined into a single time series. In our case, there exists a complete overlap between the time series of ACS five-year estimates and the time series of Data Axle estimates. We can leverage the repeated co-occurrence of the two sources to fit models that account for the relationship between the two. Below, we describe different ways we attempt to perform nowcasting.

6.1 Baselines

To set a baseline, we perform three simple nowcasts of 2020 tract-level households. First, we take a 'naïve' approach of using the 2015–2019 ACS five-year counts as a direct estimate of 2020 households. Second, we take the mean households for each tract over all years from 2006 through 2019. Third, we take the 2020 adjusted household estimates from Data Axle as a direct estimate itself.

6.2 Time series linear models

While there exist many methods for generating small-area estimation forecasts, we first consider straightforward extrapolative methods. Extrapolative methods have performed well at small-area forecasting (Rayer 2008; Wilson 2015; Hauer 2017). Moreover, more complex approaches like component forecasting or machine learning are unlikely to perform well given that our setting involves short time series and only a one-period nowcasting window (Wilson et al. 2021).

Accordingly, we fit a time series linear model (TSLM), where each tract's household estimate is modeled as a linear function of time (Hyndman and Athanasopoulos 2018). We fit this model to the ACS five-year time series, then refit it to include adjusted estimates from Data Axle as an exogenous covariate.

6.3 Autoregressive models

Autoregressive (AR) and AR-integrated moving-average models are less frequently used for demographic forecasting (Walters and Chai 2008). Nonetheless, a basic AR model makes intuitive sense for this application. Data Axle estimates come from a nonrepresentative dataset, while the ACS estimates are outdated. By fitting a model that contains the previous year's ACS household (the AR term) and the current year's Data Axle estimate (the exogenous predictor term) as predictors of the current year's estimate, we are essentially learning how to optimally weight data from two imperfect sources. Again, we fit an AR model to the ACS time series, then refit it to include adjusted estimates from Data Axle as an exogenous predictor.

6.4 Modeling the difference between the ACS and decennial census

An additional challenge emerges from the systematic differences between the five-year ACS and the decennial census. Since the population of the United States on the whole is increasing over time, we expect that nationwide estimates based on the five-year ACS will, on average, underestimate population levels in the year of its release.

To account for the differences between these two data Census Bureau data sources, we introduce a second stage to our nowcasting strategy. We fit an additional linear model to capture the average difference between the five-year ACS household estimates released in a decennial year and the decennial census itself.

This second-stage model essentially serves as a 'level shifter' that corrects for the expected systematic undercount caused by the ACS's reliance on data collected in earlier years. We fit this second stage using households from the 2010 Decennial Census as the dependent variable and from the 2006–2010 five-year ACS as the independent variable.

Table 3 formally summarizes the models applied throughout our nowcasting procedure. We proceed as follows: First, we fit different forecasting models to the five-year ACS time series. Next, we use each model to generate predictions of household counts in each tract in 2020. We use these forecasts directly as 'one-stage' estimates of the 2020 Decennial Census. Then, to correct for expected downward bias, we use the one-stage nowcasts to generate predictions from another model fit to capture differences between contemporaneous household estimates from the five-year ACS and the decennial census.

We implement the full procedure in R. Ideally, we would fit both stages within the same Bayesian procedure to correctly propagate errors (Alexander, Polimis, and Zagheni 2020). However, the scale of our task makes Bayesian inference computationally infeasible. We therefore elected to use more computationally efficient modeling packages for each stage separately. We used the fable package to implement forecasting (O'Hara, Hyndman, and Wang 2021) and the base lm() function to fit the linear model.

Table 3: Summary of all models applied for nowcasting the tract-level households from the 2020 Decennial Census

A) Nowcasting models

Baselines		
	Naïve	$A_{i,t} = A_{i,t-1}$
	Mean	$A_{i,t} = \frac{1}{T-1}\sum_{j \neq t} A_{i,j}$
	2020 Data Axle	$A_{i,t} = D_{i,t}$
AR(1)		
	AR(1)	$A_{i,t} \sim N(\alpha_0 + \alpha_1 A_{i,t-1}, \epsilon_{i,t})$
	AR(1) + Data Axle	$A_{i,t} \sim N(\alpha_0 + \alpha_1 A_{i,t-1} + \alpha_2 D_{i,t}, \epsilon_{i,t})$
TSLM		
	TSLM	$A_{i,t} \sim N(\alpha_0 + \alpha_1 t, \epsilon_{i,t})$
	TSLM + Data Axle	$A_{i,t} \sim N(\alpha_0 + \alpha_1 t + \alpha_2 D_{i,t}, \epsilon_{i,t})$

B) Stages

	One-stage	$Yi,t = Ai,t$
	Two-stage	$Yi,t \sim N(\beta 0 + \beta 1 Ai,t,\varphi i,t)$

Notes: Panel A presents various models for ACS five-year tract-level households ($A_{i,t}$) in terms of ACS measures from previous periods and Data Axle measures from the same period ($D_{i,t}$). Panel B presents two ways of relating ACS five-year estimates to counts from the decennial census ($Y_{i,t}$). The one-stage approach assumes the two are equivalent, while the two-stage approach requires estimating the difference between the sources.

Table 4 presents the results of this exercise. The baseline of using the 2015–2019 ACS with the two-stage adjustment are off the decennial by 117.5 households on average and tend to undercount by an average of 46.2 households. Adding the second stage to the modeling procedure reduces the bias of the naïve model by 44%. Using adjusted estimates from Data Axle to directly estimate households leads to greater errors (median absolute deviation (MAD) = 194.28) but less bias (–25.68) than the naïve baseline.

The only model that outperforms the baseline in terms of both absolute deviations and bias is the AR(1) model with adjusted estimates from Data Axle included as an exogenous predictor (MAD = 114.9, bias = 41.5). Adding estimates from Data Axle as exogenous predictors reduces bias in the two-stage models by roughly 20% for the AR(1) specification (from –52 to –41.5) and by 8% for the TSLM specification, which suggests that the inclusion of these data meaningfully improves household nowcasting.

Table 4: Error metrics comparing alternative approaches to nowcasting tract-level households in 2020

	Forecasting model	One-stage		Two-stage	
		MAD	Bias	MAD	Bias
Baselines					
	Naïve	127.32	–82.50	117.49	–46.20
	Mean	158.65	–129.72	143.06	–94.57
	Data Axle	194.05	–28.26		
AR(1)					
	AR(1)	123.79	–88.19	114.49	–52.02
	AR(1) + Data Axle	114.49	–76.47	114.86	–40.01
TSLM					
	TSLM	136.06	–78.70	128.36	–42.30
	TSLM + Data Axle	134.06	–74.88	126.71	–38.39

Notes: Comparisons are made to counts from 2020 Decennial Census, cross-walked back to 2010 block geographies. One-stage metrics compare forecasts produced from each nowcasting model directly to the 2020 Decennial Census. Two-stage metrics compare forecasts passed through a fitted level-shifting model that accounts for the systematic difference between the five-year ACS and the decennial census. *N* = 69,594 tracts.

7. Discussion

Surveys such as the decennial census and the ACS are the major source of small-area demographic estimates in the United States. The decennial census provides as close to a full population count as exists every ten years, and the ACS provides more timely demographic estimates based on a large but still restricted sample. Such rigorously developed and administered data collection efforts are essential for demographic research.

Alternative sources of small-area demographic estimates have emerged in recent years that rely on large-scale data sources, such as social network user data, credit records, or consumer trace data. These nontraditional data sources have the potential to contribute to demographic research alongside the census by providing data with more flexibility, at more granular scales, and at higher frequency. The consumer trace data we study here can yield small-area estimates on an annual frequency, and because they are based on address data they can be used to produce estimates with flexible boundaries. We anticipate increased interest in consumer trace data among demographers in the coming years because the cost of these data is now relatively low and the availability and the usability of the 2020 Decennial Census for small-area demographic estimation remains uncertain given the proposed differential privacy disclosure rule.

However, it is important to recognize that consumer trace data are not designed to produce demographic estimates. The construction process of these data remains mostly unknown to researchers, and they come with few guarantees about the rigor and quality of their processed data. This paper asks whether consumer trace data can be used for constructing small-area demographic estimates and what issues may arise in using them for demographic research. The results indicate that raw counts of households based on these data are highly correlated with the ACS estimates for the same period. For counties for which annual ACS estimates are available, the mean absolute proportional error is less than 10% in all years, and for five-year estimates at the ZCTA and tract level it is less than 20%. When looking at factors associated with divergence between the ACS estimates and estimates based on consumer data, we find that the consumer data are more likely to not be within 10% of the ACS estimates in areas with less precise ACS estimates, as indicated by a large margin or error, as well as in areas with a larger share of young adults and more unemployment. Given that these characteristics are likely to be persistent over time, they can be used as input in adjusting the consumer data–based estimates.

We also show that calibration based on linear models of local characteristics can further reduce deviations between consumer trace data from the survey-based estimates and the ACS. The goal of this calibration exercise is to show that the data can be modeled, not to identify the best possible model. The best-performing models we applied reduce errors by 32%, 39%, and 15% for counties, tracts, and ZCTAs, respectively. The remaining deviations are only slightly larger than margins of error around the ACS estimates.

Our calibration exercise illustrates the importance of fitting and comparing multiple models and the value of using parameter-selection procedures that learn from the data. Because nontraditional sources lack transparency, it can be difficult to know a priori how they should be modeled. In this paper we found that Lasso penalized regressions performed best at two of the three geographic scales we analyzed. At all three geographic levels, the Lasso outperformed linear models based on local characteristics, like those used previously for this task (Alexander, Polimis, and Zagheni 2020; Zagheni et al. 2018).

Finally, we have also evaluated the utility of Data Axle data for nowcasting tract-level households. We consider a range of standard time series specifications for predicting counts from the 2020 Decennial Census using older ACS data and current Data Axle data. A single period autoregression specification that includes adjusted Data Axle estimates as an exogenous predictor achieves marginally lower errors than the baseline. The bulk of the advantage gained from using Data Axle data is in prediction bias. Using the ACS five-year estimates alone to nowcast households yields substantial undercounts.

In comparison, Data Axle counts have larger prediction errors but are less biased. Using different modeling assumptions, one can shift the relative weights between the two sources to obtain an optimal balance between bias and error for a particular application.

While we conclude that in general the data match up well with representative surveys, some of our findings also validate concerns about the 'black box' nature of many nontraditional data sources (Salganik 2019). Most notably, we observed a substantial undocumented spike in raw households in 2013 due to the addition of new sources to the database. This spike damaged the efficacy of the calibration models we fit, which all assume that deviations between the data and the census are constant over time. Fortunately, although the spike is obvious at the county level, its effect on estimates is not as apparent for small areas.

We view the results presented here as broadly encouraging for the potential of future demographic research leveraging consumer trace data. Several potential applications of the data remain to be explored. These include measuring migration flows between small areas and leveraging the point-located nature of the data to develop demographic estimates for non-census geographic regions, like a voting precinct, school district, or the radius around a point of interest.

In conclusion, the consumer data examined in this paper appears usable as a source of small-area demographic estimates, being able to produce annual estimates down to the tract level that are closely aligned with ACS estimates. Additional adjustments based on local characteristics further improve the accuracy of these estimates, supporting the use of adjustments for systematic differences that remain stable over time.

References

Alburez-Gutierrez, D., Zagheni, E., Aref, S., Gil-Clavel, S., Grow, A., and Negraia, D.V. (2019). Demography in the digital era: New data sources for population research. In: Arbia, G., Peluso, S., Pini, A., and Rivenllini, G. (eds.). *Book of short papers*. New York: Pearson: SIS2019. doi:10.31235/osf.io/24jp7.

Alexander, C.H. (2002). A discussion of the quality of estimates from the American Community Survey for small population groups. Draft prepared for discussion at the Fall Census Bureau Advisory Committees Meeting.

Alexander, M., Polimis, K., and Zagheni, E. (2020). Combining social media and survey data to nowcast migrant stocks in the United States. arXiv preprint arXiv:2003.02895. doi:10.1007/s11113-020-09599-3.

Asquith, B.J., Mast, E., and Reed, D. (2021). Local effects of large new apartment buildings in low-income areas. *The Review of Economics and Statistics* 1–46. doi:10.1162/rest_a_01055.

Bell, W.R., Datta, G.S., and Ghosh, M. (2013). Benchmarking small area estimators. *Biometrika* 100(1): 189–202. doi:10.1093/biomet/ass063.

Belloni, A., Chernozhukov, V., and Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives* 28(2): 29–50. doi:10.1257/jep.28.2.29.

Billari, F.C. and Zagheni, E. (2017). Big data and population processes: A revolution? In: Petrucci, A. and Verde, R. (eds.). *Statistics and Data Science: New challenges, new generations*. Proceedings of the Conference of the Italian Statistical Society, 28–30 June 2017, Florence (Italy). Florence: Firenze University Press: 167–178. doi:10.31235/osf.io/f9vzp.

Boarnet, M.G., Bostic, R.W., Burinskiy, E., Rodnyansky, S., and Prohofsky, A. (2018). Gentrification near rail transit areas: A micro-data analysis of moves into Los Angeles Metro Rail station areas. UC Davis: National Center for Sustainable Transportation. https://escholarship.org/uc/item/4p4584w8.

Bryan, T. (2004). Population estimates. In: Siegel, J.S. and Swanson, D.A. (eds.). The methods and materials of demography Amsterdam: Elsevier: 523–560. doi:10.1016/B978-012641955-9/50054-0.

Bürkner, P.C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software* 80(1). doi:10.18637/jss.v080.i01.

Cesare, N., Lee, H., McCormick, T., Spiro, E., and Zagheni, E. (2018). Promises and pitfalls of using digital traces for demographic research. *Demography* 55(5): 1979–1999. doi:10.1007/s13524-018-0715-2.

Chetty, R., Friedman, J.N., Hendren, N., Stepner, M., and The Opportunity Insights Team (2020). The economic impacts of COVID-19: Evidence from a new public database built using private sector data. No. w27431. National Bureau of Economic Research. doi:10.3386/w27431.

Damiano, A. and Frenier, C. (2020). Build baby build? Housing submarkets and the effects of new construction on existing rents. (Center for Urban and Regional Affairs Working Paper) University of Minnesota. Retrieved from https://www.tonydamiano.com/project/new-con/bbb-wp.pdf.

Dever, J.A., Rafferty, A., and Valliant, R. (2008). Internet surveys: Can statistical adjustments eliminate coverage bias? *Survey Research Methods* 2(2).

DeWaard, J., Johnson, J., and Whitaker, S. (2019). Internal migration in the United States: A comprehensive comparative assessment of the Consumer Credit Panel. Demographic Research 41(33): 953–1006. doi:10.4054/DemRes.2019.41.33.

Diamond, R., McQuade, T., and Qian, F. (2019). The effects of rent control expansion on tenants, landlords, and inequality: Evidence from San Francisco. *American Economic Review* 109(9): 3365–3394. doi:10.1257/aer.20181289.

Dragan, K., Ellen, I.G., and Glied, S. (2020). Does gentrification displace poor children and their families? New evidence from Medicaid data in New York City. *Regional Science and Urban Economics* 83: 103481. doi:10.1016/j.regsciurbeco.2019.103481.

Folch, D.C., Arribas-Bel, D., Koschinsky, J., and Spielman, S.E. (2016). Spatial variation in the quality of American Community Survey estimates. *Demography* 53(5): 1535–1554. doi:10.1007/s13524-016-0499-1.

Freemark, Y. (2020). Upzoning Chicago: Impacts of a zoning reform on property values and housing construction. *Urban Affairs Review* 56(3): 758789. doi:10.1177/1078087418824672.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., and Qian, J. (2021). Package 'glmnet'. *CRAN R Repositary*.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., and Rubin, D.B. (2013). *Bayesian data analysis*. New York: Chapman and Hall/CRC. doi:10.1201/b16018.

Greenlee, A.J. (2019). Assessing the intersection of neighborhood change and residential mobility pathways for the Chicago metropolitan area (2006–2015). *Housing Policy Debate* 29(1) (2019): 186–212. doi:10.1080/10511482.2018.1476898.

Hauer, M.E. (2017). Migration induced by sea-level rise could reshape the US population landscape. *Nature Climate Change* 7(5): 321–325. doi:10.1038/nclimate3271.

Hauer, M.E. and Santos-Lozada, A.R. (2021). Differential privacy in the 2020 census will distort COVID-19 rates. *Socius* 7. doi:10.1177/2378023121994014.

Hyndman, R.J. and Athanasopoulos, G. (2018). *Forecasting: principles and practice*. OTexts.

Kennel, T.L. and Li, M. (2009). Content and coverage quality of a commercial address list as a national sampling frame for household surveys. *Proceeding of the 2009 Joint Statistical Meetings, American Statistical Association, Section on Survey Research Methods, Washington, DC.* Alexandria, VA: American Statistical Association. 2364–2378.

Lee, D and Van der Klaauw, W. (2010). An introduction to the FRBNY consumer credit panel. *FRB of New York Staff Report* 479. doi:10.2139/ssrn.1719116.

Lewandowski, D., Kurowicka, D., and Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* 100(9): 1989–2001. doi:10.1016/j.jmva.2009.04.008.

Li, X. (2021). Do new housing units in your backyard raise your rents. *Journal of Economic Geography*. doi:10.1093/jeg/lbab034.

Ling, C., Heck, J.E., Cockburn, M., Liew, Z., Marcotte, E., and Ritz, B. (2019). Residential mobility in early childhood and the impact on misclassification in pesticide exposures. *Environmental Research* 173: 212–220. doi:10.1016/j.envres.2019.03.039.

López-Vizcaíno, E., Lombardía, M.J., and Morales, D. (2013). Multinomial-based small area estimation of labour force indicators. *Statistical Modelling* 13(2): 153–178. doi:10.1177/1471082X13478873.

Molina, M. and Garip, F. (2019). Machine learning for sociology. *Annual Review of Sociology* 45: 27–45. doi:10.1146/annurev-soc-073117-041106.

O'Hara-Wild, M., Hyndman, R., and Wang, E. (2021). fable: Forecasting Models for Tidy Time Series. R package version 0.3.1. https://CRAN.R-project.org/package=fable.

Pasek, J., Jang, S.M., Cobb, C.L., Dennis, J.M., and Disogra, C. (2014). Can marketing data aid survey research? Examining accuracy and completeness in consumer-file data. *The Public Opinion Quarterly* 78(4): 889–916. doi:10.1093/poq/nfu043.

Penner, A.M. and Dodge, K.A. (2019). Using administrative data for social science and policy. *The Russell Sage Foundation Journal of the Social Sciences* 5(3): 1–18. doi:10.7758/rsf.2019.5.3.01.

Pennington, K. (2020). Does building new housing cause displacement? The supply and demand effects of construction in San Francisco. SSRN preprint. doi:10.2139/ssrn.3867764.

Pfeffermann, D. (2013). New important developments in small area estimation. *Statistical Science* 28(1): 40–68. doi:10.1214/12-STS395.

Phillips, D.C. (2020). Measuring housing stability with consumer reference data. *Demography* 57(4): 1323–1344. doi:10.1007/s13524-020-00893-5.

Rao, J.NK. (2003). *Small-area estimation*. Hoboken, NJ: John Wiley & Sons. doi:10.1002/0471722189.

Rayer, S. (2008). Population forecast errors: A primer for planners. *Journal of Planning Education and Research* 27(4): 417–430. doi:10.1177/0739456X07313925.

Ruggles, S., Fitch, C., Magnuson, D., and Schroeder, J. (2019). Differential privacy and census data: Implications for social and economic research. In: Johnson, W.R. and Markel, K. (eds.). *AEA papers and proceedings vol. 109*. Nashville, TN: AEA: 403–408. doi:10.1257/pandp.20191107.

Sakshaug, J.W., Wiśniowski, A., Perez Ruiz, D.A., and Blom, A.G. (2019). Supplementing small probability samples with nonprobability samples: A Bayesian approach. *Journal of Official Statistics* 35(3): 653–681. doi:10.2478/jos-2019-0027.

Salganik, M.J. (2017). *Bit by bit: Social research in the digital age*. Princeton, NJ: Princeton University Press. Open review edition.

Salganik, M.J. (2019). *Bit by bit: Social research in the digital age*. Princeton, NJ: Princeton University Press.

Schnake-Mahl, A.S., Jahn, J.L., Subramanian, S.V., Waters, M.C., and Arcaya, M. (2020). Gentrification, neighborhood change, and population health: A systematic review. *Journal of Urban Health* 97(1): 1–25. doi:10.1007/s11524-019-00400-1.

U.S. Census Bureau (2015). United States Public Use Microdata Sample (PUMS): 2010 Census of Population and Housing, Technical Documentation.

Varian, H.R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives* 28(2): 3–28. doi:10.1257/jep.28.2.3.

Walters, A. and Chai, Q. (2008). Investigating the use of holt-winters time series model for forecasting population at the state and sub-state levels. *Journal of the Demographics and Workforce Section* 21: 7–8.

Williams, N.E., Thomas, T.A., Dunbar, M., Eagle, N., and Dobra, A. (2015). Measures of human mobility using mobile phone records enhanced with GIS data. *PloS One* 10(7). doi:10.1371/journal.pone.0133630.

Wilson, T. (2015). New evaluations of simple models for small area population forecasts. *Population, Space and Place* 21(4): 335–353. doi:10.1002/psp.1847.

Wilson, T., Grossman, I., Alexander, M., Rees, P., and Temple, J. (2021). Methods for small area population forecasts: State-of-the-art and research needs. *SocArXiv Papers*. doi:10.31235/osf.io/sp6me.

Winkler, R.L., Butler, J.L., Curtis, K.J., and Egan-Robertson, D. (2021). Differential privacy and the accuracy of county-level net migration estimates. *Population Research and Policy Review* 41: 417–435. doi:10.1007/s11113-021-09664-5.

Zagheni, E., Garimella, V.R.K., Weber, I., and State, B. (2014). Inferring international and internal migration patterns from twitter data. In: *WWW '14 Companion: Proceedings of the 23rd International Conference on World Wide Web April 2014.* New York: Association for Computing Machinery: 439–444. doi:10.1145/2567948.2576930.

Zagheni, E., Polimis, K., Alexander, M., Weber, I., and Billari, F.C. (2018). Combining social media data and traditional surveys to Nowcast migration stocks. In: *Annual Meeting of the Population Association of America*.

Zagheni, E. and Weber, I. (2012). You are where you e-mail: Using e-mail data to estimate international migration rates. *Proceedings of the 4th annual ACM web science conference*. In: *WebSci '12: Proceedings of the 4th Annual ACM Web Science Conference June 2012* New York: Association for Computing Machinery: 348–351. doi:10.1145/2380718.2380764.

Zagheni, E., Weber, I., and Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review* 43(4): 721–734.

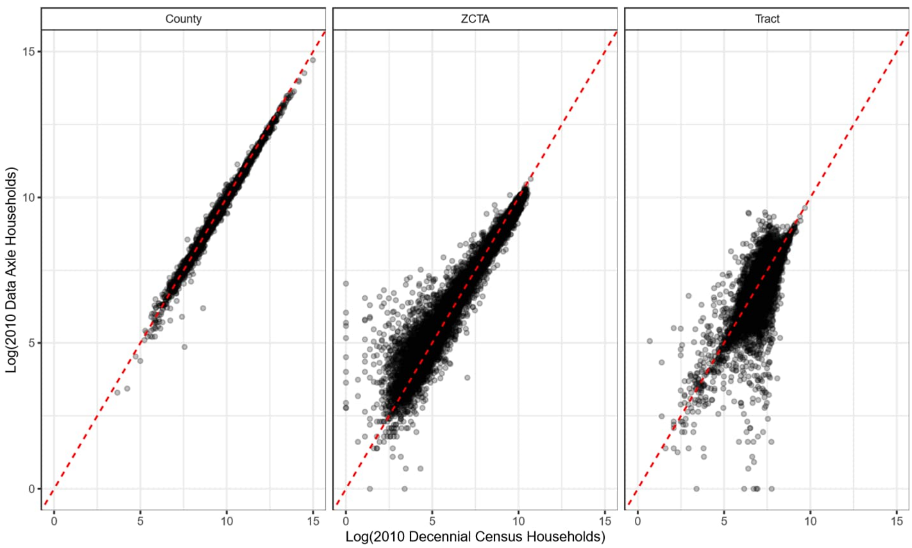
Appendices

Appendix 1: 2010 Census estimates

Given that the decennial census estimates combine both the lowest variance and lowest bias, we provide comparison of the 2010 Data Axle estimates to those of the 2010 Decennial Census. A limit of the 2010 Data Axle data is that all observations for Arizona and Delaware have been lost.

As shown on the scatterplot below, the fit of Data Axle to census estimates is extremely good at the county level, while there are more variations at the ZCTA and tract levels.

Figure A-1: Scatterplots comparing raw households from 2010 Data Axle to the 2010 Decennial Census at all three geographic levels

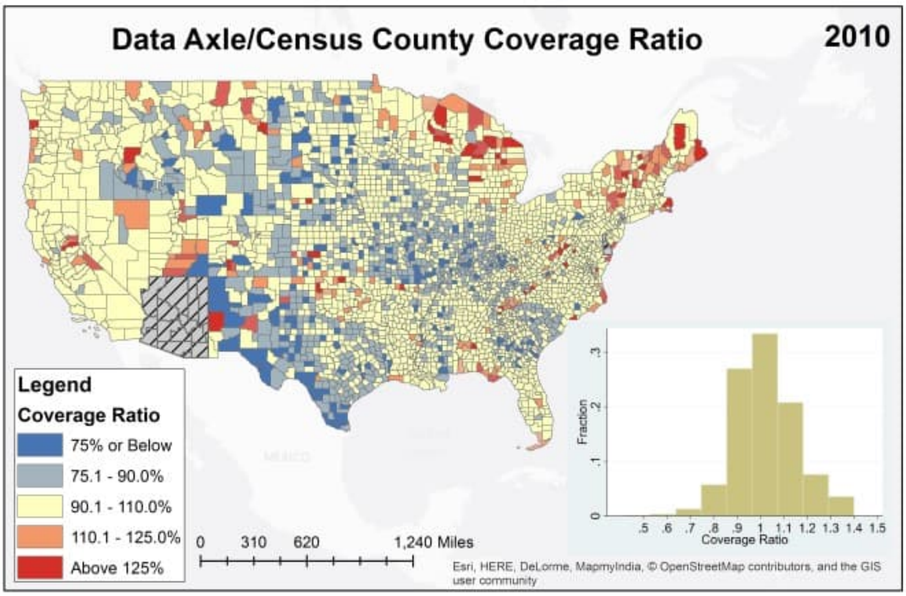


Note: Data for Arizona and Delaware are missing.

When mapping coverage at the county level nationwide, the data show most Data Axle estimates falling within 10% of the decennial census estimate. However, there are a substantial number of counties in New Mexico, Texas, and the Great Plains with substantial undercounts. These appear to be mostly rural counties. On the other hand, a number of counties in the Midwest and Northeast exhibit substantial Data Axle overcounts. In both cases there appears to be spatial correlations with Data Axle household estimates in adjacent counties being substantially under or over the census counts.

Overall, the comparison of the raw Data Axle to the decennial census indicates that Data Axle estimates at various geographical levels are tightly aligned with decennial census estimates, but some areas exhibit substantial deviations that need to be addressed to produce more reliable estimates.

Figure A-2: Maps of coverage ratio of raw households from 2010 Data Axle to the 2010 Decennial Census at county level



Note: Data for Arizona and Delaware are missing.

Appendix 2: Data preparation

We began by counting the number of census household equivalents in the Data Axle data, summing up to the county, ZCTA, and tract across the country. In some instances, Data Axle counts were missing for certain geographies in certain years. Wherever this occurs, we impute the Data Axle count as the mean of the previous year and the following year. Imputations were made for 18 county-year units, 484 ZCTA-year units, and 1,709 tract-year units. This imputation is most notable for the states of Arizona and Delaware, which are missing from the Data Axle data in 2010. To test the robustness of the imputation for these two entire states, we refit the model at the county level with them excluded and achieved a similar fit. After imputation, we merged these counts with census count datasets to prepare the data for modeling. We describe these merges below.

Appendix 3: Counties

We first joined county-level Data Axle counts to decennial census count data. The resulting dataset contains households from both datasets for all 3,142 counties in the dataset. This excludes counties in Puerto Rico, for which Data Axle does not collect data. From the model fit on this 2010 data, we project forward and backward to all 3,142 counties in the dataset. When we compare our adjusted estimates to large counties, we use the subset of 773 counties with estimates included in the ACS one-year data.

Appendix 4: ZCTAs

We joined ZCTA-level Data Axle counts to decennial data, which contains estimates for 32,803 ZCTAs. Of these, 32,369 have matches in the Data Axle data (98.7%). We further removed 3,142 ZCTAs with very small populations (less than 100 people). After fitting the model, we projected these ZCTAs for 2006–2018. We then computed five-year running averages and compared these to the five-year ACS. Our comparisons include the above ZCTAs from 2007–2011 and 2014–2018.

Appendix 5: Tracts

We joined tract-level Data Axle counts to the decennial data, which contains estimates for 72,067 tracts. Of these, 71,093 are in the Data Axle data (98.6%). We projected these tracts for 2006–2018, then computed five-year running averages and compared these to the five-year ACS. Our comparisons include the above ZCTAs from 2006–2010 and 2014–2018.

Appendix 6: Covariates

In our adjustment model specification, we define a vector x_i of covariates, which we include in all our models. These are the subset of variables explored in sociodemographic characteristics section[5] that are available in the 2010 Census. The full list of included variables, along with their corresponding labels in the code, are listed in Table A-1.

Table A-1: List of covariates included adjusted count models

Variable	Definition
pown	Percent homeowners
pfemale	Percent female
page 1824	Percent population age 18–24
page under18	Percent population age under 18
page 65plus	Percent population age over 65
pnw	Percent non-white
pblack	Percent black
phisp	Percent Hispanic
pasian	Percent Asian
pnative	Percent native
pvacant	Percent vacant
phu othervacant	Percent vacant, other
phu seasonal	Percent vacant, seasonal
pgq	Percent population in group quarters
ruca 1	Rural/urban classification (10 levels)
chh density	Households per square mile (standardized)

Appendix 7: Models

As described in the main text, we fit and compare five different calibration models:

1. Bivariate: $\log p^{CE}_{it} = \alpha_0 + \alpha_1 \log p^{DA}_{it}$.
2. + Fixed effects: $\log p^{CE}_{it} = \alpha_0 + \alpha_1 \log p^{DA}_{it} + \beta^0 \mathbf{s_i}$, where $\mathbf{s_i}$ contains 50 state dummies.
3. + Covariates: $\log p^{CE}_{it} = \alpha_0 + \alpha_1 \log p^{DA}_{it} + \beta^0 \mathbf{s_i} + \delta^0 \mathbf{x_i}$, where $\mathbf{s_i}$ contains 50 state dummies and $\mathbf{x_i}$ contains local characteristics.
4. Multilevel: $\log p^{CE}_{itg} = \alpha_{0g} + \alpha_{1g} \log p^{DA}_{itg} + \beta' \mathbf{s_i} + \delta_0 \mathbf{x_i}$, where slopes and intercepts can vary within states *g*.
5. Lasso: Fully interacted set of covariates and state dummies selected to minimize $\sum_{i=1}^{n} (p^{CE}_{itg} - \hat{p}^{CE}_{itg})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$ (Molina and Garip 2019).

Models 1, 2, and 3 are fit using basic linear specifications using R's base lm function. Models 4 and 5 are described below.

Appendix 8: Multilevel models

We fit the multilevel models within a Bayesian framework using Stan via the brms package in R (Bürkner 2017). At all three geographic levels, we fit multilevel models with random slopes and intercepts. We use vectorized priors to improve the efficiency of the sampler. This means we define the same prior for all the same parameter types across these models.

In settling on our preferred specification, we follow recommended practices in selecting weakly informative priors (Gelman et al. 2013). We specify the priors for all intercepts and coefficients as a normal distribution with mean 0 and standard deviation of 5. All standard deviations are modeled with student's T distributions, with 3 degrees of freedom, mean 0, and standard deviation of 5. Finally, we model the correlations between varying slopes and intercepts using an LKJ-correlation prior with parameter ζ set to 2, which slightly favors a correlation of 0 (Lewandowski, Kurowicka, and Joe 2009).

All posterior distributions were sampled using the NUTS sampler from Stan. We varied the number of samples, chains, warm-up draws, and the thinning parameter to balance adequate sampling with computational efficiency. Ultimately, we obtained Rhat values equaling 1.0, along with bulk and tail effective sample sizes exceeding 2,000 for all parameters estimated in all Bayesian models. We also visually inspected the distributions and trace plots to ensure no problematic sampling anomalies emerged.

Appendix 9: Lasso models

Lasso models and other forms of penalized regression seek to achieve a balance between under- and overfitting in prediction problems. They address underfitting by exploring the possible importance of many parameters and address overfitting by algorithmically removing parameters to minimize some objective function. The Lasso objective function written in bullet 5 in Appendix 7 selects coefficients to achieve a balance between minimizing the sum of squared residuals and minimizing the absolute sum of all the coefficients in the model. This second objective is called the penalty term. The relative weight of the penalty term is controlled by λ, a tuning parameter (Molina and Garip 2019; Belloni, Chernozhukov, and Hansen 2014).

We apply Lasso models using the R package glmnet. We include 1,665 possible parameters obtained by fully interacting all covariates and dummy variables in the data. We use 100-fold cross-validation to test model performance across a range of possible λ values. We then follow the package authors' recommendation in choosing the largest λ value that achieves within-sample errors within one standard error of the minimum (Friedman et al. 2021). This results in the inclusion of 36, 394, and 881 parameters for counties, ZCTAs, and tracts, respectively.

Appendix 10: Out-of-sample performance of calibration models

In section 5 of the main text, we evaluated the performance of calibration models out of time during the years between the 2010 and 2020 Census by comparing calibrated estimates from Data Axle against the ACS across every tract in the country. Here, we conduct a second out-of-sample evaluation, where we fit each calibration model to 80% of the tracts in the 2010 Decennial Census and evaluate it against the other 20%.

Table A-2 presents the results of this evaluation. In general, the models performing best in the main text also perform the best here. One exception is at the county level, where the fixed effects model with covariates and multilevel model both outperform the Lasso. For smaller geographies, we again observe that the Lasso model obtains slightly lower errors than the other models for ZCTAs and exhibits larger improvements for tracts.

Table A-2: Out-of-sample MAPE for estimating households at each geographic level

Model	County	ZCTA	Tract
Unadjusted	0.101	0.153	0.322
Bivariate	0.096	0.143	0.355
+ State FEs	0.080	0.131	0.350
+ Covariates	0.066	0.111	0.282
Multilevel	0.064	0.113	0.273
Lasso	0.080	0.104	0.175

Appendix 11: Testing the benchmarking property

It is generally desirable for small-area estimates to align with larger area estimates when aggregated. Here we present an assessment of whether the adjusted estimates exhibit this benchmarking property. We considered whether the adjusted tract-level estimates we produced, when aggregated to the county level, could accurately estimate the number of households in large counties. We computed the MAPE and MAD for these aggregates using the same procedure as the main text, averaging over all county-years. We present the results in Table A-3.

Table A-3: MAPE and MAD for estimating households for large counties, averaged over all years

Model	MAPE	MAD
Unadjusted	0.080	11,889
Bivariate	0.066	7,979
+ State FEs	0.061	6,812
+ Covariates	0.051	5,845
Multilevel	0.043	5,569
Lasso	0.047	5,680

The results indicate that adjusted tract-level estimates aggregated up to counties are more accurate estimates of county-level households than adjusted county-level estimates. Nearly every tract-level adjustment model has a lower MAPE than the best county-level model.

We draw two conclusions from these results. First, tract-level household estimates from Data Axle largely exhibit the benchmarking property. Second, adjustment models fit at the tract level capture important within-county heterogeneity in the deviations between the ACS and Data Axle estimates.

Appendix 12: Largest coefficients from bias adjustment models

Table A-4: Coefficients from each county-level adjustment model

Bivariate		+ State FEs		+ Covariates		Multilevel		Lasso	
Var.	Coef.	Var.	Coef.	Var.	Coef.	Var.	Coef.	Var.	Coef.
α0	–0.02	α0	–0.203	α0	1.05	α0	0.98	α0	0.498
α1	1.007	α1	1.021	α1	0.983	α1	0.99	α1	0.964
		State 15	0.334	pnw	0.779	State 21	0.89	page 1824:page under 18	1.131
		State 4	0.173	page 1824	0.514	State 29	0.816	page under18:page 65plus	–0.703
		State 29	0.163	phu othervacant	0.349	pfemale	–0.781	pown:pvacant	–0.548
		State 35	0.163	State 4	0.178	pgq	–0.663	phisp:pnative	0.379
		State 48	0.161	State 29	0.156	page under18	–0.633	State 40:phu othervacant	–0.327
		State 6	0.132	State 48	0.115	State 20	–0.586	State 29:page 65plus	0.246
		State 46	0.131	State 45	0.114	page 1824	0.537	State 21:phu othervacant	0.228
		State 21	0.13	phu seasonal	0.104	pvacant	–0.503	State 48:phu seasonal	0.2
		State 16	0.126	State 16	0.101	State 15	0.389	State 36:phu seasonal	–0.157
		State 28	0.111	State 21	0.1	page 65plus	–0.363	pown	–0.148

Notes: All α_0 and α_1 are presented, along with the top 10 largest coefficients from each model. For the Bayesian multilevel model, α_1 represents the mean across all states.

Table A-5: Coefficients from each ZCTA-level adjustment model

Bivariate		+ State FEs		+ Covariates		Multilevel		Lasso	
Var.	Coef.	Var.	Coef.	Var.	Coef.	Var.	Coef.	Var.	Coef.
α0	–0.201	α0	–0.259	α0	0.453	α0	0.516	α0	0.216
α1	1.031	α1	1.032	α1	1.006	α1	1.01	α1	1.007
		State 15	0.281	phu othervacant	0.836	pvacant	–0.806	State 15:pnative	54.276
		State 35	0.235	page 1824	0.498	State 4	0.756	State 42:pnative	–8.046
		State 4	0.221	pnw	0.352	phu othervacant	0.72	State 50:pasian	–7.375
		State 6	0.166	State 4	0.192	pgq	–0.615	pasian:phu_seasonal	–4.082
		State 48	0.157	phu seasonal	0.171	State 10	–0.605	State 10:phu othervacant	–3.981
		State 29	0.147	State 29	0.149	State 8	0.577	State 16:pblack	3.588
		State 46	0.129	State 48	0.125	b page 1824	0.525	phu othervacant:phu seasonal	3.352
		State 45	0.111	State 35	0.121	State 35	0.501	State 50:pblack	2.784
		State 18	0.11	State 45	0.118	State 15	0.456	State 28:pasian	2.52
		State 21	0.107	pnative	0.111	b page under18	–0.43	pasian:phu othervacant	2.178

Notes: All α0 and α1 are presented, along with the top 10 largest coefficients from each model. For the Bayesian multilevel model, α1 represents the mean across all states.

Table A-6: Coefficients from each tract-level adjustment model

Bivariate		+ State FEs		+ Covariates		Multilevel		Lasso	
Var.	Coef.	Var.	Coef.	Var.	Coef.	Var.	Coef.	Var.	Coef.
α0	1.831	α0	1.679	α0	2.19	α0	2.325	α0	1.253
α1	0.762	α1	0.775	α1	0.766	α1	0.74	α1	0.877
		State 11	0.424	page 1824	0.435	State 11	4.248	State 15:pnative	28.882
		State 35	0.19	pnative	0.341	State 53	2.745	State 30:pasian	15.616
		State 53	0.185	State 11	0.302	State 38	2.717	State 11:pnative	8.828
		State 15	0.167	State 35	0.128	State 35	2.344	State 38:pblack	7.765
		State 6	0.154	pnw	0.118	State 27	2.063	State 24:pnative	–5.281
		State 41	0.141	State 53	0.101	State 54	1.676	State 54:pnative	4.869
		State 16	0.134	State 16	0.099	State 2	1.648	State 10:pnative	–4.417
		State 48	0.128	State 28	0.09	State 19	1.574	pasian:pnative	4.372
		State 4	0.123	State 13	0.085	State 26	–1.499	pasian:ruca 19	4.176
		State 38	0.119	State 48	0.081	State 30	1.441	page under18:phu seasonal	3.945

Notes: All α_0 and α_1 are presented, along with the top 10 largest coefficients from each model. For the Bayesian multilevel model, α_1 represents the mean across all states.