Title

Using census aggregates to proxy for household characteristics: an application to vehicle ownership

Journal

Transportation: Planning - Policy - Research - Practice, 36(2)

ISSN

1572-9435

Authors

Adjemian, Michael
Williams, Jeffrey

Publication Date

2009-03-01

DOI

10.1007/s11116-009-9191-2

Peer reviewed

Transportation (2009) 36:223–241
DOI 10.1007/s11116-009-9191-2

Using census aggregates to proxy for household characteristics: an application to vehicle ownership

Michael Adjemian · Jeffrey Williams

Published online: 27 February 2009

Abstract Traditionally, researchers studying transportation choice have used data either acquired from household surveys or broad, region-wide aggregates. At the disaggregate level, researchers usually do not have access to important variables or observations. This study investigates the potential usefulness of a proxy approach to modeling discrete choice vehicle ownership: substituting narrow area-based aggregate proxies for missing micro-level explanatory variables by accessing large, publicly maintained datasets. We use data from the 2000 Bay Area Travel Survey (BATS) and the contemporaneous U.S. Census file to compare three models of vehicle ownership, drawing area-wide proxies from increasing levels of aggregation. The models with proxies are compared with a parallel model that uses only survey data. The results indicate that the proxy models are preferred in terms of model selection criteria, and predict vehicle ownership as well or better than the survey model. Parameter values produced by the proxy method effectively approximate those returned by household survey models in terms of coefficient sign and significance, particularly when the aggregate variables are representative of their household-level counterparts. The proxy model with the narrowest level of aggregation achieved the best fit, coefficient precision, and percentage of correct prediction.

Keywords Vehicle choice modeling · Ownership choice · Proxy · Census · Survey · Discrete choice

M. Adjemian (✉)
Department of Agricultural and Resource Economics, University of California, Davis, 2409 Charleston Oaks Lane, Decatur, GA 30030, USA
e-mail: mkadjemian@ucdavis.edu

J. Williams
Department of Agricultural and Resource Economics and Giannini Foundation, University of California, Davis, 2144 Social Sciences & Humanities, One Shields Avenue, Davis, CA 95616, USA
e-mail: williams@primal.ucdavis.edu

Introduction

In general, disaggregate travel models use data from household travel surveys. However, problems with that survey collection process are well documented, from consumer resistance and language impediments to response ambiguity and increasing percentages of cell phone-only households. Many large regional surveys do not sample from particular rural and suburban areas, and do not offer a sufficient number of observations to conduct detailed analysis at the sub-regional level. Along with the rising cost of surveying households, these problems have motivated researchers to consider alternative data gathering methods, such as simulating travel surveys (Pointer et al. 2004) or accessing public microdata surveys (Purvis 1994).

One way to obtain data for disaggregate choice modeling while avoiding surveys is to enhance state department of motor vehicles (DMV) micro-level ownership records with socioeconomic and demographic information from the United States census. In that way, characteristics of the local community serve as proxies for information missing about the households, such as income or ethnicity. Although the use of zonal data is widespread in transportation research, using zonal proxies is not common. However, the technique is regularly used in the study of epidemiology, because public health data are often missing important elements due to privacy concerns.

To assess its validity, we simulate the proxy method by supplementing limited information from a recent household auto ownership survey with zonal census aggregates, resembling data that could be constructed using information from the DMV and the Census. Census information is reported at a range of spatial levels, specifically census block group, tract, and zipcode. A census block group, representing about 1,000 people, is the smallest unit for which detailed socioeconomic and demographic information is aggregated. Census tracts, composed of a set of block groups, contain 4,000 people on average. Census tracts are specifically designed to group individuals relatively homogeneous in terms of demographics and economic status (US-Census-Bureau 1994). According to the U.S. Census Bureau, census tracts are intended to be permanent statistical subdivisions, increasing their usefulness in empirical applications. Zip codes, which were introduced by the U.S. Postal Service for operational reasons rather than to represent similar individuals, contain on average 30,000 individuals. Accordingly, we estimate three ownership choice models using these three separate zonal proxy groups and compare our results to a parallel model using only data drawn from a household survey. We evaluate the proxy models on the bases of their predictive ability and their approximation of household model coefficients.

Overview and related work

Traditionally, research into vehicle choice has taken one of two general approaches, differing by the degree of aggregation. In the aggregate approach, the existence of a representative consumer or more flexibly a known distribution of preferences across the population is assumed by the researcher, allowing the use of broad aggregates of demand and supply in estimating the significance of a set of automobile attributes. Such models adequately forecast future car ownership at an aggregate level (Salon 2006). This aggregate approach is considered cost-efficient in terms of computation and data requirements, but it can hamper the ability to distinguish among the presumed explanatory variables in terms of their effect on vehicle choice (Potoglou and Kanaroglou 2008). In the disaggregate approach, the researcher considers the household as the decision-making entity, gathers

micro-level data on its characteristics and recent vehicle choices, estimates the contribution of various components to automobile choice, and then sums over households to project the composition of future vehicle fleets. The disaggregate approach has the advantage of better capturing consumer behavior, and as a result, the relationship between vehicle attributes, household characteristics and ownership choice. Although the data requirements are more exhaustive, disaggregate models are the preferred method of modeling vehicle choice (Bhat and Pulugurta 1998), particularly if policy efficacy is considered.

Annually, each U.S. state's motor vehicles division collects a wide range of information from its registered drivers. For example, the local vehicle stock is implied by registration fees paid annually to the DMV. The typical DMV records intra- and inter-state vehicle transactions. The DMV knows the population of individuals licensed to drive a car, including certain demographic information about them, their type of license and their home addresses. Yet neither the typical DMV nor localities regularly collect detailed information on the makeup of the automobile fleet as well as their operators for the express purpose of studying transportation choices. As a result, those researchers interested in the disaggregate approach to estimating the factors associated with the question of how people determine which automobile to purchase must resort to surveying individual consumers, a process that can be expensive in terms of time and money. Potential difficulties encountered include selection of the appropriate sample, unresponsiveness of targeted subjects, and mistakes or ambiguity in responses. Researchers have also noticed troubling trends in survey response: growing consumer resistance, proportion of homes unavailable to be reached by telephone, and per-home survey costs (Stopher and Greaves 2007). The supposed choice set is often narrowed out of concerns over the cost of customized surveys (Train and Winston 2005).

Another possible source of socioeconomic and demographic data is the U.S. decennial census, which is freely available. Information such as age, ethnicity, education, poverty status, and income level are collected, aggregated and reported at multiple spatial scales, from statewide averages to groupings by residential block. Additionally, the census collects other information directly relevant to the study of vehicle choice, such as average travel to work time and the percentage of residents that use various modes of transit.

By using a proxy approach, merging zonal census-level information with data on household vehicle ownership or purchases could result in important, novel insights into the vehicle choice decision, expanding not only the viable topics available for research but also the size of the population considered. Neighborhood characteristics, which have been found significant in automobile choice (Zegras 2007), could be obtained easily. Furthermore, a vehicle choice model combining aggregated characteristics with disaggregate choice data may help improve on the traditional role of inexpensively generated aggregate models to forecast car ownership, while still allowing the researcher freedom to consider policy analysis as allowed by disaggregate models.

Although the use of zonal aggregates to proxy for missing micro-level data is not common in vehicle choice research, it has been applied for nearly a century in the study of health outcomes (Krieger et al. 2003). Whereas residential address and individual health status, such as disease cases or discrete self-reported health scale measures, are often available in public health records, important possible predictors of disease susceptibility such as educational attainment, occupational status, and income often go uncollected (Berjon et al. 2005). To investigate the role of socioeconomic status in predicting health outcomes, researchers frequently match individual addresses to one or more spatial levels for which census information is collected, whether at the level of zip codes, tracts, or block groups. These area-wide aggregates are then used to substitute for missing data in order to

control for important explanatory variables that would otherwise not be included, with final analysis conducted on the resulting unified data set (Gowrisankaran and Town 1999). The U.S. does not report census data at the household level.

Use of proxy aggregates in place of micro-level variables does not suffer from the ecological fallacy by the classic definition (Krieger 1992), but it does bias the results of statistical analysis. In the study of the effect of socioeconomic status on health outcomes, this point has lead some researchers to advise caution in the interpretation of the results it yields (Geronimus et al. 1996). Others have found that the proxy method has unique advantages, and that the degree of bias diminishes when researchers derive census information from smaller geographic units of aggregation (Soobader et al. 2001). According to Subramanian et al. (2006), area-based proxies provide conservative approximations of the micro-level variables they are intended to represent.

Statistical considerations

Before we can estimate and compare the results of our analyses, there are a few important points to consider regarding the statistical validity of using aggregate-level variables to proxy for household information. These concerns apply to the value and significance attached to parameter values estimated using the proxy method. If the researcher is solely interested in predicting the level of household automobile ownership, the following concerns will not affect the validity of the results.

To control for household information unavailable in DMV data, values of missing explanatory variables are imputed from area-wide aggregates. According to Wickens (1972), the use of even a poor proxy is preferred to simply omitting the information. Nevertheless, the use of aggregate proxies likely biases household-level parameters. Geronimus et al. (1996) show that two sources of bias can affect the proxy estimates. First, since an aggregate variable is only partially correlated with its individual counterpart, its use introduces a measurement error related to the degree of correlation. Second, aggregate variables may themselves represent information useful in the micro-level analysis. Perhaps households consider both their own characteristics as well as those of their neighborhood in making the vehicle choice decision: beyond individual income, local income level may be an important factor. In that case, the magnitude of estimated coefficients will increase. Of course, if neighborhood contextual effects themselves factor into household vehicle choice, the absence of local aggregates from traditional models leads to a similar set of problems. Additionally, Geronimus et al. demonstrate that estimates of the remaining explanatory variables are also affected by the introduction of proxy aggregates, with bias resulting from their correlation with the proxy aggregates and the missing original variables, and the relationship between the missing variables and the outcome.

Because aggregate explanatory variables tend to be relatively smooth with respect to their household counterparts, they may not exhibit enough variance to provide meaningful estimates. Indeed, if the spatial scale of aggregation were large enough so as to envelop all the observations, the census proxies would be identical for every individual. This obstacle should diminish as the geographic unit of aggregation narrows relative to the size of the studied region. Also, aggregate variables may exhibit a high degree of multicollinearity, making it difficult to estimate parameter values. Moreover, according to Moulton (1990), failing to account for potential error correlations at the aggregate level could bias standard errors downwards. As a result, if spatial aggregates are geographically correlated, the statistical significance attached to results found via the proxy method may be spurious.

Data

The first source of data for this analysis is the 2000 San Francisco Bay Area Travel Survey (BATS), commissioned by the Bay Area Metropolitan Transportation Commission (MTC). Data on household vehicle fleet ownership and socioeconomic/demographic information were collected from 15,064 households in the nine-county region, during the period February 2000 to March 2001. Although the residential addresses of survey participants are not reported, BATS includes the pertinent matched census block group, tract, and zip code. The survey achieved a 99.9% success rate in geocoding the home addresses of surveyed households. Inclusion of geocoded locations in the BATS dataset allows comparison between individual and area-wide aggregate choice models.

Census information is imported from the year 2000 United States Census Summary File 3 for the census tracts, block groups, and zip codes listed in the BATS dataset. For each spatial level of aggregation, Census data included population size, racial composition, average age, employment figures, and median income. Census data for each aggregation level were appended to the individual information provided by BATS according to the geocoded block group, tract, and zip code indicated in the dataset. Because this study compares the results of parallel models, the 2,911 BATS households that did not report income, age, employment status, or ethnicity are excluded from the analysis.

Methodology

Automobile ownership affects a household's transportation mode, its destination of interest in leisure activities, and the number of trips it makes (Nobile et al. 1997). Disaggregate models focused on predicting the number of cars chosen by a household are used to provide inputs into transport projection models (De Jong et al. 2004). As in the study of scaled health outcomes, since a given household chooses its ownership level from a known set of available options, discrete choice models are generally used to estimate relevant vehicle choice parameters. Among these, Bhat and Pulugurta (1998) found that unordered-response mechanisms, such as the multinomial logit model (MNL), fit ownership data better than do ordered-response models. In order to verify this, we estimated an ordered logit (ORL) version of each ownership model and rejected it on the basis of Akaike's Information Criterion (AIC).

Model of household ownership choice

We model only the demand side of the auto market, and assume that the supply of cars is perfectly elastic. The explanatory variables are chosen both on the basis of their likelihood to play a role in ownership choice according to the literature, and their being comparable across approaches. As the most commonly used model to explain systems with several discrete outcomes, the MNL model applies random utility theory to the auto ownership decision (Ben-Akiva and Lerman 1985). A household chooses the number of cars in order to maximize its own utility; for each level of car ownership (j), let the utility for an individual household be given by

$$U_j = V_j + \varepsilon_j, j = 0, 1, 2, 3, 4 \quad (1)$$

where V_j represents the deterministic portion of utility, and ε_j denotes a random component. The household subscript i is suppressed for simplicity. Deterministic utility is then defined as being composed of a vector of attributes multiplied by parameters. For each household,

$$V_j = \beta_j' x, \tag{2}$$

and

$$V_j^I = \beta_j^{I'} x_*^I;\ V_j^A = \beta_j^{A'} x_*^A,$$

for

$$x_*^I = \left[x^C x^I\right],\ x_*^A = \left[x^C x^A\right]$$

Define V_j^I as the deterministic component in the case of alternative j where the vector of household characteristics (x^I), such as logarithm of income, and householder age are employed in estimation. Likewise, V_j^A represents the deterministic component of utility when area-wide aggregates (x^A) are used as proxies in the definition of household utility; for instance, the logarithm of median income, average age, and neighborhood racial composition represented in percentages for the census block group. For both coefficient vectors in E.2, x^C represents household data on the number of licensed individuals in the home, since this information was provided by BATS and is available in the state DMV records one could make use of in the census approach.

Given that BATS data on households that own five or more cars make up only about 1% of the sample, we model Bay Area residents as having five possible vehicle ownership choices: zero, one, two, three, or "four or more" cars. The household chooses an ownership alternative—the dependent variable (y)—in the choice set in order to maximize its utility. Under the maintained assumption of independent random error, ownership utilities across individuals are uncorrelated. We assume that ε is identically, type I extreme value distributed for all households, and can thus represent the model with an MNL framework. Given the multinomial outcome variable and the specified choice probabilities, the maximum likelihood function is implicitly defined. Coefficient values are estimated in order to maximize the likelihood of observing the original sample.

One potential drawback of the multinomial logit framework is that it assumes the Independence of Irrelevant Alternatives (IIA). That is, the ratio of the probabilities of any two alternatives must remain independent to changes in other alternatives. In effect, cross-alternative correlation is not allowed. For example, if the option of owning four or more cars was suddenly removed from the choice set, households in that group would not necessarily be predisposed to own three cars, but instead would distribute to the other choice alternatives at the ratio of the original probabilities. Modeling alternatives that have the advantage of avoiding the IIA problem include ordered GEV and mixed logit (Train 2003). Mixed logit models allow flexibility in substitution patterns, and also account for coefficients that vary in the population. Although we did not use a mixed logit model in this paper, the effect of using aggregate proxies in mixed logit models has not been thoroughly studied, and is an area we intend to research in the future.

Empirical results and discussion

Descriptive statistics

Table 1 displays means and standard deviations for the variables we used in multinomial logit estimation of the factors associated with vehicle ownership choice. We removed BATS records that neglected to provide information for any of the variables in the table. After listwise deletion, the sample of surveyed households amounted to 12,153

Table 1 Descriptive statistics

Variable	Mean	SD	Min	Max
No. household vehicles	1.84	0.90	0	4
No. licensed individuals in household	1.76	0.71	0	6
Household size (people per household)	2.33	1.27	1	11
Avg. household size (Block group)	2.59	0.64	0.78	19.91
Avg. household size (Tract)	2.59	0.56	1.13	19.91
Avg. household size (Zipcode)	2.60	0.46	0.62	4.97
Age of householder	47.44	14.10	18	91
Avg. age (Block group)	42.20	6.01	12.25	62.39
Avg. age (Tract)	42.29	5.13	13.88	54.12
Avg. age (Zipcode)	42.37	4.28	21.38	59.79
Logarithm of 1999 income	11.08	0.69	8.52	12.01
Logarithm of median 1999 income (Block group)	11.11	0.40	7.82	12.21
Logarithm of median 1999 income (Tract)	11.11	0.36	7.82	12.21
Logarithm of median 1999 Income (Zipcode)	11.08	0.31	9.40	12.21
Householder is hispanic indicator	0.03	0.18	0	1
Percentage of hispanic residents (Block group)	0.14	0.13	0	0.95
Percentage of hispanic residents (Tract)	0.15	0.12	0	0.85
Percentage of hispanic residents (Zipcode)	0.16	0.11	0	0.77
Householder is black indicator	0.04	0.20	0	1
Percentage of black residents (Block group)	0.05	0.09	0	0.88
Percentage of black residents (Tract)	0.05	0.09	0	0.81
Percentage of black residents (Zipcode)	0.06	0.09	0	0.59
Householder is Asian indicator	0.07	0.25	0	1
Percentage of Asian residents (Block group)	0.16	0.15	0	0.94
Percentage of Asian residents (Tract)	0.16	0.14	0	0.93
Percentage of Asian residents (Zipcode)	0.16	0.13	0	0.66
Householder is unemployed indicator	0.24	0.43	0	1
Percentage of unemployed residents (Block group)	0.29	0.09	0.02	0.96
Percentage of unemployed residents (Tract)	0.29	0.07	0.13	0.89
Percentage of unemployed residents (Zipcode)	0.29	0.05	0	0.71
Median year of residential construction (Block group)	1965	15	1939	1999
Percent of residents with a bachelor's degree (Block group)	0.19	0.09	0	0.51
Population density (Block group Residents per square mile)	9,177	11,159	0.57	172,898

observations, representing 3,626, 1,344, and 261 different census block groups, tracts and zip codes, respectively. The mean auto ownership for the households under study was about two (with the truncation at four).

The first two items in Table 1 are elements available to both survey and area-based analyses. That is, only having access to DMV driver and ownership registration records would not preclude the researcher from discovering how many licensed drivers reside in a household as well as its vehicle portfolio. After that, sections are constructed so that the individual household characteristic is listed above the relevant census measure. For example, resident age precedes the average age of the block group population. Finally, the

last three variables listed in the table are used only in the full model that combines both household and census-level information.

As expected, the standard deviation and range of each area-based measure decreases as the geographical region of interest grows in size. The aggregate averages are smoothed when the size of the relevant census area increases, from block group, to tract, to zipcode. Although the standard deviation and range of aggregate values is lower than that of their individual counterparts, they contain enough variation to make estimation meaningful. For example, while the dichotomous indicator for an unemployed head of household takes on values of zero or one by definition, the proportion of unemployed individuals in a block group ranges from 2 to 96%.

Household survey data for income, age of the householder and household size all display similar means to those derived from the census data. For instance, the mean of log income earned in 1999 for the sampled individuals in BATS matched the median log census zipcode income. This indicates that the BATS sample is representative of the Bay Area geographic region with respect to these factors. Beyond that, they are reported in like units to their household counterparts.

On the other hand, all race and unemployment aggregates are reported as percentages, while their household counterparts are dichotomous. Besides the indicator for a black head of household, the household BATS race variables do not match the area aggregates very well. Only 3% of the BATS sample reported a Hispanic heritage, while Hispanics represented 15% of census tracts. For Asians, this disparity was 7% and 16%, respectively. These inconsistencies were not due to such households being less likely to report information critical to this study and being subject to listwise deletion. Instead, Asian and Hispanic households were undersampled by BATS relative to the population as reported by US Census Summary File 3. Census unemployment aggregates are calculated with respect to the entire adult population of a geographic region, not necessarily those of working age, and thus report higher means than their individual counterparts. Additionally, the census aggregates for these variables are reported in different units, percentage of residents versus a 0–1 dummy, and represent perhaps entirely different characteristics than do the household level data.

Full model

We estimated a MNL model using both household information and area aggregates to determine whether neighborhood characteristics themselves factor into the vehicle ownership decision. Following Geronimus et al. (1996), this model yields some intuition about the direction of the bias that we can expect from employing aggregate proxies. Recall that if a given household considers average neighborhood income when selecting its ownership level even after controlling for household income, then the parameter for the income proxy may be biased upwards. Because they represent the lowest aggregation level for which census information is available, we incorporated block group averages in the model.

As a result of using the MNL, an unordered discrete choice model, the parameters in Table 2 convey the sign of the effect on the probability of selecting that ownership alternative relative to not owning a car, given an increase in the explanatory variable. For example, the significant coefficient for logged income of 0.86 relates that an increase in income is associated with a household being more likely to own a single car than none at all. The table shows that block group median income, average household size and some aggregate race variables are significant at 1% level in the combined model. Average resident age and the remaining aggregate race controls were significant at the 5% level. Broadly, these results indicate that households weigh both their own characteristics as well as those

Table 2 Ownership choice parameters for the full model (Base category: Zero car households)

Variable	Auto ownership level			
	1 Auto	2 Autos	3 Autos	4 Autos
Logarithm of 1999 income	0.86** (10.60)	1.91** (18.90)	2.16** (18.00)	2.46** (14.40)
Log of 1999 median income[a]	1.03** (4.72)	1.27** (4.96)	1.49** (5.15)	1.74** (4.79)
Household size	−0.43** (−6.73)	−0.13 (−1.85)	−0.12 (−1.59)	−0.04 (−0.51)
Average household size[a]	−0.04 (−0.50)	0.33** (3.19)	0.51** (4.26)	0.66** (3.89)
No. licensed individuals in household	2.17** (9.52)	4.68** (18.90)	6.30** (23.40)	7.16** (25.70)
Age of the householder	0.01 (1.30)	0.02** (3.37)	0.03** (5.17)	0.04** (6.01)
Average age of residents[a]	0.01 (0.92)	0.03* (2.19)	0.04* (2.15)	0.03 (1.34)
Hispanic householder	−0.08 (−0.31)	−0.31 (−0.89)	−0.22 (−0.54)	−0.58 (−1.14)
Pct. of hispanic residents[a]	−1.71** (−3.42)	−2.12** (−3.48)	−2.16** (−3.09)	−2.90** (−3.39)
Black householder	0.03 (0.10)	0.32 (0.84)	0.41 (1.01)	0.55 (1.24)
Pct. of black residents[a]	−0.93 (−1.65)	−1.06 (−1.47)	−1.66* (−2.00)	−2.27* (−1.96)
Asian householder	0.2 (0.81)	0.04 (1.13)	−0.19 (−0.62)	−0.7 (−1.93)
Pct. of Asian residents[a]	−0.64 (−1.46)	−1.66** (−3.36)	−1.63** (−2.94)	−1.36* (−2.10)
Unemployed householder	−0.23 (−1.45)	−0.29 (−1.63)	−0.77** (−4.03)	−0.76** (−3.39)
Pct. of unemployed residents[a]	0.73 (0.92)	0.08 (0.08)	−0.14 (−0.13)	−0.93 (−0.74)
Median year of homes[a]	0.02** (4.84)	0.03** (5.18)	0.02** (3.93)	0.02* (2.30)
Pct. of residents with a bachelor's degree[a]	−6.03** (−6.19)	−8.60** (−7.56)	−10.96** (−8.55)	−13.31** (−8.87)
Population density[a]	−0.00003** (−7.28)	−0.00007** (−8.05)	−0.00009** (−6.41)	−0.00014** (−8.29)
Constant	−60.59** (−6.73)	−91.29** (−8.64)	−93.96** (−7.81)	−87.71** (−6.36)

Table 2 continued

Variable	Auto ownership level			
	1 Auto	2 Autos	3 Autos	4 Autos
Log likelihood at zero			−15677	
Log likelihood at convergence			−10067	
Pseudo R-squared			0.358	
Average probability of correct prediction[b]			65.4%	
Average probability of correct prediction[c]			70.0%	
MNL akaike information criterion			20284	
ORL akaike information criterion			20507	

t-statistics in parentheses

[a] Census proxies drawn from block group level

[b] The predicted probability of selecting the correct choice was at least 0.5

[c] The predicted probability of selecting the correct choice was higher than all other alternatives, but not necessarily at least 0.5

* significant at 5%; ** significant at 1%

of their immediate social network in their ownership choice. For example the coefficients for household and neighborhood median income are significant and positive, suggesting that individuals may seek to "keep up with the Joneses" with respect to auto ownership. The age and race parameters suggest that cultural factors may also affect choice level.

Additionally, we found that some neighborhood characteristics unavailable in traditional household surveys yet easily constructed from census data were significant in the model, concurring with Zegras' result (2007). Households in block groups with newer homes were more likely to select any level of car ownership over zero. The reverse was true for homes in densely populated and highly educated areas. Consequently, vehicle choice modelers should consider accounting for such neighborhood characteristics even if they have access to household survey data.

We confirmed Bhat and Pulugurta's (1998) result for the BATS data: MNL was preferred to ORL on the basis of AIC. Presumably, this result signifies that the difference between ownership alternatives is not limited to simply the number of cars. One possible explanation is that the utility placed on different ownership levels may be related to vehicle heterogeneity in multi-car homes. Additionally, we found that the full model predicted at least a 0.5 probability of selecting the actual level of household ownership for over 65% of the BATS sample. When we relaxed the criteria and instead labeled the alternative with the highest probability of selection (although not necessarily at least 0.5) as the predicted alternative, the model achieved a successful prediction rate of 70%.

Proxy models

According to Geronimus et al. (1996), the bias resulting from the use of census proxies depends on two factors. First, the extent to which census aggregates represent their micro level counterparts. Second, whether aggregates themselves belong in the household choice model. Keeping these aspects in mind, we evaluate the usefulness of the census approach by comparing its parameter values, marginal effects, and predictive ability with those produced using only household survey data.

Table 3 displays estimates for the household model alongside those for the simulated proxy models. For the latter, census aggregates from block groups, tracts and zipcodes substitute median income, average household size, average resident age, proportionate race makeup and unemployment for their household level counterparts. The household variable for the number of licensed individuals in the home is shared by the census approach, since this information is also found in DMV records. The base category is defined as a zero car home. All estimates are robust to errors clustered at the tract level.

The MNL results in Table 3 show that the more representative proxy variables well approximate their household level counterparts: coefficients for aggregate income, household size, and age each display similar results to the household model in terms of sign and statistical significance. This is also true for the "number of licensed individuals" variable, the only one shared by both approaches. For example, the household coefficient of 2 for a single car home indicates that the addition of a licensed individual increases the likelihood of owning a car, relative to none at all. The proxy models estimate this coefficient as 2.18, 2.25, and 2.34 using block group, tract, and zipcode data, respectively. For each census level, estimated coefficients for the alternative specific constant—indicating whether a household will choose a given ownership level relative to not owning a car, everything else being equal—match the survey approach very well.

MNL parameters do not specify the magnitude of the effect on probability, since one of the consequences of nonlinear regression is that marginal effects differ by point of

Table 3 Parameter estimates for parallel ownership choice models (Base category: Zero car households)

Variable	Household				Proxy models			
					Block group			
	1 Auto	2 Autos	3 Autos	4 Autos	1 Auto	2 Autos	3 Autos	4 Autos
Logarithm of 1999 income/Log of 1999 median income	0.78** (10.60)	1.73** (17.90)	1.94** (17.10)	2.29** (14.90)	1.11** (4.97)	1.75** (6.67)	1.90** (6.66)	2.16** (6.33)
Household size/Average household size	0.21** (−3.68)	0.24** (3.67)	0.30** (4.18)	0.39** (4.77)	0.18 (1.92)	0.62** (5.23)	0.91** (6.88)	1.11** (7.14)
No. licensed individuals in household	2.00** (8.85)	4.31** (17.10)	5.88** (21.40)	6.70** (23.50)	2.18** (8.43)	5.28** (19.20)	6.89** (23.60)	7.80** (26.00)
Age of the householder/Average age of residents	0.02** (5.96)	0.04** (9.18)	0.06** (11.10)	0.07** (11.10)	0.09** (7.96)	0.15** (10.70)	0.17** (10.90)	0.17** (9.50)
Hispanic householder/Pct. of hispanic residents	−0.29(−1.43)	−0.54* (−1.97)	−0.55(−1.59)	−1.07* (−2.43)	−1.78** (−3.98)	−2.31** (−4.08)	−2.46** (−3.90)	−3.44** (−4.46)
Black householder/Pct. of black residents	0.03(0.12)	0.37(1.18)	0.46(1.36)	0.5(1.31)	−1.06* (−2.27)	−1.61** (−2.58)	−2.33** (−3.05)	−3.36** (−3.13)
Asian householder/Pct. of Asian residents	−0.05(−0.24)	−0.39(−1.58)	−0.65* (−2.41)	−1.17** (−3.54)	−0.98* (−2.31)	−2.34** (−4.71)	−2084** (−5.22)	−3.42** (−5.45)
Unemployed householder/Pct. of unemployed residents	−0.2(−1.43)	−0.29(−1.93)	−0.79** (−4.71)	−0.80** (−3.91)	1.29(1.94)	1.24(1.46)	1.45(1.55)	1.46(1.31)
Constant	−8.89** (−11.0)	−24.48** (−22.6)	−31.97** (−24.9)	−4.27** (−22.9)	−15.84** (−6.29)	−30.76** (−10.6)	−38.48** (−12.2)	−45.48** (−11.8)
Log likelihood at zero	−15677				−15677			
Log likelihood at convergence	−10730				−10715			
Pseudo R-squared	0.316				0.317			
Average probability of correct prediction[a]	64.4%				64.8%			
Average probability of correct prediction[b]	69.4%				69.7%			
SSD[c]	0				1.11			
MNL Akaike information criterion	21,529				21,500			
ORL Akaike information criterion	21,598				21,668			

Table 3 continued

Variable	Proxy models							
	Tract				Zipcode			
	1 Auto	2 Autos	3 Autos	4 Autos	1 Auto	2 Autos	3 Autos	4 Autos
Logarithm of 1999 income/Log of 1999 median income	1.05** (3.96)	1.64** (5.21)	1.79** (5.11)	1.76** (4.35)	0.73* (2.07)	1.31** (3.15)	1.37** (3.02)	1.22** (2.39)
Household size/Average household size	0.33** (3.45)	0.76** (6.10)	0.88** (5.75)	1.25** (6.75)	0.75** (3.52)	1.19** (4.10)	1.41** (4.51)	1.81** (5.13)
No. licensed individuals in household	2.28** (8.57)	5.38** (19.30)	6.99** (23.70)	7.90** (26.00)	2.34** (9.16)	5.51** (20.40)	7.13** (24.70)	8.03** (27.10)
Age of the householder/Average age of residents	0.10** (7.28)	0.18** (10.10)	0.21** (10.40)	0.22** (8.81)	0.12** (7.19)	0.21** (9.05)	0.25** (9.49)	0.25** (8.17)
Hispanic householder/Pct. of hispanic residents	−2.02** (−4.00)	−2.31** (−3.54)	−2.00** (−2.61)	−3.38** (−3.71)	−2.86** (−3.77)	−2.95** (−2.89)	−2.96** (−2.62)	−4.37** (−3.45)
Black householder/Pct. of black residents	−0.75(−1.43)	−1.1(−1.57)	−1.71* (−1.99)	−2.58* (−2.33)	−0.75(−1.30)	−1.11(−1.45)	−1.6(−1.70)	−3.14(−2.66)
Asian householder/Pct. of Asian residents	−1.15* (−2.56)	−2.36** (−4.35)	−2.50** (−4.17)	−3.20** (−4.70)	−1.59** (−2.92)	−2.87** (−4.26)	−3.17** (−4.25)	−3.88** (−4.63)
Unemployed householder/Pct. of unemployed residents	0.43(0.57)	−0.45(−0.52)	−0.34(−0.33)	−1.9(−1.55)	−0.48(−0.35)	−1.94(−1.19)	−1.26(−0.70)	−1.83(−0.95)
Constant	−18.87** (−5.36)	−30.78** (−8.96)	−38.87** (−10.2)	−42.42** (−9.45)	−13.77** (−3.45)	−29.07** (−6.23)	−36.83** (−7.22)	−39.31** (−6.85)
Log likelihood at zero	−15677				−15677			
Log likelihood at convergence	−10719				−10726			
Pseudo R-squared	0.316				0.316			
Average probability of correct prediction[a]	64.7%				64.8%			
Average probability of correct prediction[b]	69.7%				69.9%			
SSD[c]	1.50				2.94			

Table 3 continued

Variable	Proxy models							
	Tract				Zipcode			
	1 Auto	2 Autos	3 Autos	4 Autos	1 Auto	2 Autos	3 Autos	4 Autos
MNL Akaike information criterion	21,509				21,522			
ORL Akaike information criterion	21,660				21,691			

t-statistics in parentheses

[a] The predicted probability of selecting the correct choice was at least 0.5

[b] The predicted probability of selecting the correct choice was higher than all other alternatives, but not necessarily at least 0.5

[c] The sum of squared deviations between the proxy and household models is calculated for the representative variables: income, household size, number of licensed individuals, and age of householder

[*] significant at 5%; [**] significant at 1%

Table 4 Marginal effects of the modeled explanatory variables on ownership choice (Evaluated at the mean)

Variable	Household				
	No Autos	1 Auto	2 Autos	3 Autos	4 Autos
Logarithm of 1999 income/Log of 1999 median income	0.01** (−4.43)	−0.19** (−16.7)	0.13** (12.60)	0.05** (8.38)	0.01** (7.35)
Household size/Average household size	0.000 (−1.91)	−0.09** (−12.1)	0.07** (9.92)	0.02** (6.25)	0.001** (5.08)
No. licensed individuals in household	−0.06** (−11.3)	−0.35** (−48.2)	0.15** (16.20)	0.20** (26.60)	0.06** (12.90)
Age of the householder/Average age of residents	−0.0002** (−3.75)	−0.004** (−8.72)	0.002** (4.18)	0.002** (7.74)	0.0004** (6.70)
Hispanic householder/Pct. of hispanic residents	0.000 (1.55)	0.050 (1.28)	−0.040 (−1.06)	−0.010 (−0.42)	−0.01* (−2.28)
Black householder/Pct. of black residents	0.000 (−1.13)	−0.06** (−2.59)	0.040 (1.62)	0.020 (1.15)	0.000 (0.82)
Asian householder/Pct. of Asian residents	0.00 (1.21)	0.08** (3.24)	(0.04) (−1.72)	−0.03** (−3.46)	−0.01** (−5.00)
Unemployed householder/Pct. of unemployed residents	0.000 (1.82)	0.03* (2.24)	0.020 (1.24)	−0.05** (−7.05)	−0.01** (−3.87)
Variable	Proxy models				
	Tract				
	No Autos	1 Auto	2 Autos	3 Autos	4 Autos
Logarithm of 1999 income/Log of 1999 median income	−0.004** (−3.29)	−0.11** (−4.51)	0.08** (3.32)	0.03* (2.10)	0 (1.15)
Household size/Average household size	−0.002** (−3.36)	−0.09** (−5.91)	0.05** (3.56)	0.02* (2.52)	0.01** (4.03)
No. licensed individuals in household	−0.05** (−12.6)	−0.40** (−95.1)	0.18** (27.90)	0.20** (36.00)	0.07** (20.70)
Age of the householder/Average age of residents	−0.0005** (−3.79)	−0.02** (−9.69)	0.01** (6.39)	0.01** (6.27)	0.001** (3.53)
Hispanic householder/Pct. of hispanic residents	0.01** (2.70)	0.05 (0.69)	−0.06 (−0.84)	0.02 (0.57)	−0.02 (−1.95)
Black householder/Pct. of black residents	0 (1.57)	0.09 (1.21)	0 (−0.0100)	−0.07 (−1.49)	−0.02 (−1.86)
Asian householder/Pct. of Asian residents	0.01** (2.93)	0.23** (4.97)	−0.17** (−3.92)	−0.05 (−1.92)	−0.02** (−2.71)
Unemployed householder/Pct. of unemployed residents	0 (0.29)	0.17* (2.11)	−0.13 (−1.71)	−0.01 (−0.22)	−0.03 (−1.93)

Table 4 continued

Variable	Proxy models				
	Block group				
	No Autos	1 Auto	2 Autos	3 Autos	4 Autos
Logarithm of 1999 income/Log of 1999 median income	−0.005** (−3.57)	−0.12** (−5.90)	0.09** (4.16)	0.03** (2.72)	0.01** (2.95)
Household size/Average household size	−0.002** (−3.37)	−0.09** (−7.42)	0.05** (3.54)	0.04** (6.04)	0.01** (5.07)
No. licensed individuals in household	−0.04** (−12.0)	−0.40** (−92.5)	0.18** (27.70)	0.20** (35.40)	0.07** (20.10)
Age of the householder/Average age of residents	−0.0004** (−3.79)	−0.01** (−9.21)	0.01** (6.53)	0.004** (5.54)	0.0006** (3.32)
Hispanic householder/Pct. of hispanic residents	0.01** (3.00)	0.11 (1.82)	−0.06 (−1.10)	−0.03 (−0.92)	−0.02* (−2.43)
Black householder/Pct. of black residents	0.005* (2.35)	0.13 (1.94)	−0.02 (−0.35)	−0.08* (−2.00)	−0.03* (−2.22)
Asian householder/Pct. of Asian residents	0.01** (3.14)	0.27** (6.60)	−0.17** (−4.35)	−0.09** (−4.01)	−0.02** (−3.70)
Unemployed householder/Pct. of unemployed residents	0 (−1.48)	0 (0.06)	−0.02 (−0.32)	0.02 (0.47)	0 (0.27)
Variable	Proxy models				
	Zipcode				
	No Autos	1 Auto	2 Autos	3 Autos	4 Autos
Logarithm of 1999 income/Log of 1999 median income	−0.003* (−2.48)	−0.11** (−3.51)	0.09** (2.83)	0.02 (1.28)	0 (0.22)
Household size/Average household size	−0.003** (−2.97)	−0.09** (−3.40)	0.05 (1.73)	0.03* (2.54)	0.01** (3.74)
No. licensed individuals in household	−0.05** (−13.7)	−0.40** (−98.8)	0.18** (27.60)	0.20** (36.50)	0.07** (21.10)
Age of the householder/Average age of residents	−0.0005** (−3.89)	−0.02** (−8.14)	0.01** (5.13)	0.01** (5.65)	0.001** (3.51)
Hispanic householder/Pct. of hispanic residents	0.01* (2.46)	0.02 (0.22)	0 (−0.042)	0 (−0.038)	−0.02* (−2.10)
Black householder/Pct. of black residents	0 (1.47)	0.09 (1.16)	0 (−0.053)	−0.06 (−1.11)	−0.03* (−2.38)
Asian householder/Pct. of Asian residents	0.01** (2.94)	0.25** (4.26)	−0.17** (−3.03)	−0.07* (−2.06)	−0.02** (−2.67)
Unemployed householder/Pct. of unemployed residents	0 (0.96)	0.25 (1.55)	−0.28 (−1.76)	0.03 (0.30)	−0.01 (−0.31)

t-statistics in parentheses

*significant at 5%; **significant at 1%

evaluation. Instead, the marginal effects calculated at the mean of the covariate vector are shown in Table 4, and represent the partial derivative of the outcome with respect to the regressors. In other words, the table shows the effect on the probability of selecting an alternative given a one unit increase in the explanatory variable. For instance, if an additional individual joins a single car household, the marginal effect of −0.09 in Table 4 indicates that the probability that a given household chooses to own one car decreases by 9%.

Like in Table 3, the marginal effects for the representative variables displayed in Table 4 are comparable across approaches on the basis of sign and significance. Once again, the magnitude of the proxies is similar to the corresponding household measures. For instance, using survey data alone, we find that a one unit increase in the logarithm of income is associated with a 13% chance that a household is more likely to own two cars. The complementary estimates generated by the median income proxy are 9%, 8%, and 9% for the three models.

On the other hand, when the census aggregates do not represent their household level counterparts very well, the estimates they generate often do not match those using solely micro-level data. The BATS survey was not representative of the San Francisco Bay area with respect to race or employment status. Additionally, the census aggregates for these variables are reported in different units, percentage of residents versus a 0–1 dummy, and represent perhaps entirely different characteristics than do the household level data. From that standpoint, it is not surprising that the proxy models do not approximate the household estimates with respect to these variables.

On the basis of the above results, the census approach approximates the parameter values and marginal effects of the BATS survey model reasonably well when its data are representative and presented in like units to the household-level variables. Krieger (1992) noted that proxies drawn from census block groups matched individual results better than tract data. It is intuitive that block group aggregates should better resemble the characteristics of their residents, since they are about 1/3 the size of census tracts. In our study of ownership choice, this proves to be the case. As shown in Table 3, block group proxies minimized the sum of squared deviations (SSD) from the representative household parameter values when compared with tract and zipcode census data.

In order to calculate the SSD, divide the variables so that

$$x = [x_R x_N] \tag{3}$$

where x_R is composed of these representative variables, while x_N are the remaining characteristics. Then, the SSD for every proxy model p is calculated by summing over all representative variables l

$$SSD_p = \sum_l \left(\beta_R^A - \beta_R^I\right)^2 \tag{4}$$

where β_R^A is the coefficient on a representative proxy variable, and β_R^I its household level complement SSD is minimized for the block group model, meaning that its representative parameters best approximate those produced from the survey data, followed by tract and zipcode parameters. This result verifies that the proxy variables taken from lower levels of census aggregation better match their BATS household counterparts.

Finally, the proxy models fit the observed ownership outcomes very well, as related by the likelihood and predictive attributes displayed in Table 3. In fact, each of the aggregate models would be preferred to the survey model on the basis of traditional criteria for model selection. The block group model exhibited the highest log likelihood at convergence, and the correspondingly lowest AIC value. Consequently, because all models began with the same log

likelihood at zero, the block group model displays the highest pseudo R-squared value, 31.7%. On the basis of average predictive ability, the block group proxies are again preferred to every other estimated model. The block group model assigned at least a probability of 0.5 to the correct ownership choice 64.8% of the time, while assigning a plurality of probability to the correct choice for 69.7% of all observations. Using individual data alone, the comparable correct predictions were 64.4% and 69.7%, respectively. These results lend some confidence to the proxy method if a researcher does not have access to household survey data, particularly if the research objective is to predict household automobile ownership.

Conclusions

In this paper, we have evaluated the performance of census proxies for household variables in the estimation of a micro-level discrete choice model of auto ownership. The models using census proxies display impressive fit of the data. We find that the proxy method better approximates its household survey counterparts when the aggregates are representative. Among the models using proxies, we determined that the block group data are preferred to tract and zipcode aggregates. At least in the case of the BATS 2000 survey, our results show that vehicle choice estimates derived from the census proxy approach can serve as reasonable approximations for micro-level parameters, particularly in terms of exhibiting the proper sign and significance.

Not only are aggregate proxies less expensive to acquire and analyze, they may in fact offer an improvement over the traditional survey-driven approach. The block group ownership model actually performs better than the model using survey data alone in terms of likelihood and predictive ability. Evidently, an individual considers not only his status, but also the condition of the neighborhood around him in deciding how many vehicles to own. In the health studies literature, these are referred to as contextual effects (Diez-Roux 2003). Even researchers that have access to household survey data should contemplate controlling for local aggregates.

If the researcher's objective is to predict household vehicle choice, our results demonstrate the promise of using proxy data. Although we have shown that models using some aggregate proxies can predict original ownership at least as well as the survey model, an important caveat is that prediction of future vehicle ownership is dependent upon the availability of aggregate data. If tract-level demographic or education estimates are nonexistent for a future period, the researcher must constrain the predictive model to existing data for practical applications.

Acknowledgments We thank Dr. Aaron Smith (associate professor of Agricultural and Resource Economics, U.C. Davis) for insightful comments in reviewing drafts of the manuscript. Feedback that we received at departmental seminars was very helpful in focusing the ideas we present in this paper. The authors are grateful to three anonymous referees for their valuable perspective and observations. All errors are the author's.

References

Ben-Akiva, M., Lerman, S.: Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Cambridge, MA (1985)

Berjon, F.D., Borrell, C., et al.: The usefulness of area-based socioeconomic measures to monitor social inequalities in health in southern Europe. Eur. J. Pub. Health **16**(1), 54–61 (2005)
Bhat, C., Pulugurta, V.: A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. Transp. Res. B: Methodol. **32**(1), 61–75 (1998)
De Jong, G.C., Fox, J., et al.: Comparison of car ownership models. Transp. Rev. **24**(4), 379–408 (2004)
Diez-Roux, A.V.: A glossary for multilevel analysis. Epidemiol. Bull. **24**(3), 12–13 (2003)
Geronimus, A.T., Bound, J., et al.: On the validity of using census geocode characteristics to proxy individual socioeconomic characteristics. J. Am. Stat. Assoc. **91**(434), 529–537 (1996)
Gowrisankaran, G., Town, R.J.: Estimating the quality of care in hospitals using instrumental variables. J. Health Econ. **18**, 747–767 (1999)
Krieger, N.: Overcoming the absence of socioeconomic data in medical records: validation and application of a census-based methodology. Am. J. Public Health **82**(5), 703–710 (1992)
Krieger, N., Chen, J.T., et al.: Race/Ethnicity, gender, and monitoring socioeconomic gradients in health: a comparison of area-based socioeconomic measures—the public health disparities geocoding project. Am. J. Public Health **93**(10), 1655–1671 (2003)
Moulton, B.R.: An illustration of a pitfall in estimating the effects of aggregate variables on micro units. Rev. Econ. Stat. **72**(2), 334–338 (1990)
Nobile, A., Bhat, C.R., et al.: A random effects multinomial probit model of car ownership choice. Case Stud. Bayesian Stat. **III**, 419–434 (1997)
Pointer, G., Stopher, P., et al.: Monte Carlo simulation of household travel survey data for Sydney, Australia: Bayesian updating using different local sample sizes. Transp. Res. Record: J. Transp. Res. Board **1870**, 102–108 (2004)
Potoglou, D., Kanaroglou, P.S.: Modeling car ownership in urban areas: a case study of Hamilton, Canada. J. Transp. Geogr. **16**(1), 42–54 (2008)
Purvis, C.L.: Using 1990 census public use microdata sample to estimate demographic and automobile ownership models. Transp. Res. Record **1443**, 21–29 (1994)
Salon, D.: Cars and the city? A model of the determinants of auto ownership and use for commuting in New York City with endogenous choice of residential location. The Transportation Research Board Annual Meeting (2006)
Soobader, M., LeClere, F.B., et al.: Using aggregate geographic data to proxy individual socioeconomic status: does size matter? Am. J. Public Health **91**(4), 632–636 (2001)
Stopher, P.R., Greaves, S.P.: Household travel surveys: where are we going? Transp. Res. Part A **41**, 367–381 (2007)
Subramanian, S.V., Chen, J.T. et al.: Comparing individual- and area-based socioeconomic measures for the surveillance of health disparities: a multilevel analysis of Massachusetts Births, 1989–1991. Am. J. Epidemiol. **164**(9), 823–834 (2006)
Train, K.: Discrete Choice Methods with Simulation. Cambridge University Press, New York (2003)
Train, K.E., Winston, C.: Vehicle Choice Behavior and the Declining Market Share of U.S. Automakers. Brookings Business Paper, Washington (2005)
US-Census-Bureau: Geographic Areas Reference Manual (1994). Retrieved Jan 1 2008, from http://www.census.gov/geo/www/garm.html
Wickens, M.R.: A note on the use of proxy variables. Econometrica **40**(4), 759–761 (1972)
Zegras, C.: The built environment and motor vehicle ownership and use: evidence from Santiago de Chile. Transportation Research Board 86th Annual Meeting. Washington DC, United States: 14 (2007)

Author Biographies

Michael K. Adjemian is a PhD candidate in the Department of Agricultural and Resource Economics at UC Davis. His research interests include choice modeling, transportation behavior, spatial econometrics, and finance.

Jeffrey Williams is the D. Barton DeLoach Professor in the Department of Agricultural and Resource Economics. Most of his research has concerned commodity futures markets, but recently has moved towards vehicles and air pollution, as the result of his serving on the Inspection and Maintenance Review Committee, which oversees California's Smog Check Program.