THE JOURNAL OF TRANSPORT AND LAND USE http://jtlu.org
VOL. 7 NO. 2 [2014] PP. 135–51 http://dx.doi.org/10.5198/jtlu.v7i2.468

JTLU

A vehicle ownership and utilization choice model with endogenous residential density

David Brownstone
University of California, Irvine
dbrownst@uci.edu

Hao (Audrey) Fang
eBay Inc.
hafang@ebay.com

Abstract: This paper explores the impact of residential density on households' vehicle type and usage choices using the 2001 National Household Travel Survey (NHTS). Attempts to quantify the effect of urban form on households' vehicle choice and utilization often encounter the problem of sample selectivity. Household characteristics that are unobservable to the researchers might determine simultaneously where to live, what vehicles to choose, and how much to drive them. Unless this simultaneity is modeled, any relationship between residential density and vehicle choice may be biased. This paper extends the Bayesian multivariate ordered probit and tobit model developed in Fang (2008) to treat local residential density as endogenous. The model includes equations for vehicle ownership and usage in terms of number of cars, number of trucks (vans, sports utility vehicles, and pickup trucks), miles traveled by cars, and miles traveled by trucks. We carry out policy simulations that show that an increase in residential density has a negligible effect on car choice and utilization, but slightly reduces truck choice and utilization. The largest impact we find is a -.4 arc elasticity of truck fuel use with respect to density. We also perform an out-of-sample forecast using a holdout sample to test the robustness of the model.

1 Introduction

Attempts to quantify the effect of urban form on households' vehicle choice and utilization often encounter the problem of sample selectivity. That is, household characteristics that are unobservable to the researchers might determine simultaneously where to live, what vehicles to choose, and how much to drive. Unless this simultaneity is modeled, any relationship between residential density and vehicle choice may be biased. In this paper, we study to what extend residential density affects households' vehicle ownership and vehicle miles traveled, using a Bayesian approach that corrects for the endogeneity of the density choice. Moreover, we perform an out-of-sample forecast using the estimates obtained to test the robustness of the model.

The purpose for studying a more precise relationship between residential density and households' vehicle type choice and utilization is to provide a piece of evidence for or against using residential density as a tool to control people's travel behavior, a proposal often explored in urban literature (Cervero and Kockelman 1991, Dunphy and Fisher 1996, Ewing and Cervero 2001, Brownstone and Golob 2009, Kim and Brownstone 2013, Bento et. al. 2005 and 2009).

The paper extends the models developed in Fang (2008) to treat local residential density as endogenous, and it extends the empirical work to cover the United States instead of just California. The model includes equations for vehicle ownership and usage in terms of number of cars, number of trucks, miles traveled by cars, and miles traveled by trucks.[1] Number of cars and trucks are modeled as multivariate ordered probit, and usage of cars and trucks are modeled as multivariate Tobit, both at a disaggregate level. Residential density at the census block level is added to the system as an additional dependent variable. As a whole, we will estimate a simultaneous residential density and vehicle ownership and usage model system. As such, we need additional exogenous covariates in the density equation other than the explanatory variables used in the vehicle ownership and usage equations to identify the system. The extra exogenous variable, or the instrumental variable, we use in this study is the average density for a tract's

[1]Car is defined as automobile, or station wagon; truck refers to van, sports utility vehicle, or pickup truck.

MSA, following Brueckner and Largey (2008). The basic assumption is that the average MSA density is correlated with the density at a more localized level, such as at the census block or tract level, but is uncorrelated with the unobserved factors that influence households' choice of vehicle ownership and utilization. We argue that people's decisions on what types of vehicles to drive and how much to drive are only influenced by immediate areas surrounding where they live, and not by density at the MSA level. Therefore, the average MSA density variable should be excluded from the vehicle ownership and utilization equations, while included in the localized density equation.

The practice of using variables at a more aggregate level as instrumental variables could also be found in Evan, Oates, and Shwab (1992). They found that two thirds of the families who chose to move in the last five years from their current residency moved within the same metropolitan area. The analysis thereafter in this paper is conditional on the metropolitan area people live in, but unconditional on where in the metropolitan area people choose to reside. If the unobserved characteristics also influence a household's decision on which metropolitan area to live, then the average MSA density will no longer be a valid instrument.

Other than addressing the endogeneity issue, this paper differs from Fang (2008) in two other aspects. Fang only uses the California subsample from the 2001 National Household Travel Survey (NHTS), but this paper uses a much larger data set including households across all states in the U.S. The larger data set not only provides more variation in the explanatory variables, but also provides enough observations so that proper out-of-sample forecasting can be executed. To our knowledge, this is the first paper in the literature that performs out-of-sample forecasts as an additional robustness check of the model. Note that we do not use the more recent 2009 NHTS data since this survey did not collect dual odometer readings. Lave (1994) shows that dual odometer readings are needed to accurately measure vehicle miles traveled.

The paper is organized as follows: section 2 describes the model used for estimation and the procedures for the Bayesian estimation; section 3 discusses the data used in the study, and the statistical description of the variables; detailed parameter estimation results and policy simulations are presented in Section 4; in section 5, we perform out-of-sample forecasts to test the robustness of the model; and section 6 concludes.

2 Model

The behavior of each household is characterized by five equations:

$$y_i^* = D_i\alpha + X_i\beta + \varepsilon_i \quad (1)$$

$$D_i = Z_i\gamma + \eta_i \quad (2)$$

where y_i^* is a 4 by 1 vector of latent dependent variables for number of cars, number of trucks, mileage on cars, and mileage on trucks; D_i is a measure of density for households i at the census block level, and is endogenous. The relationships between the latent dependent variables and their observed values are:

$$y_j = 0, \text{if } y_j^* \le \alpha_0, j = 1,2$$

$$y_j = 1, \text{if } \alpha_0 < y_j^* \le \alpha_1, j = 1,2$$

$$y_j = 2, \text{otherwise}, j = 1,2$$

$$y_j = y_j^*, \text{if } y_j^* > 0, j = 3,4$$

$$y_j = 0, \text{otherwise}, j = 3,4$$

The two equations of car and truck counts are modeled as bi-variate ordered probit, and the two equations of car and truck miles travelled are modeled as censored Tobit. Parameter identification of the ordered probit specifies the two cut points to be zero and one, and the variances be unrestricted (Nandrum and Chen 1996, Webb and Forster 2008, Fang 2008). Therefore, $\alpha_0 = 0$ and $\alpha_1 = 1$. Note that this ordered probit specification for the car and truck counts can match any unimodal distribution and is more flexible than Poisson or Negative Binomial count models. The miles traveled equations could also be modeled as a Heckit model, but this leads to a much more complicated estimation problem. We have experimented with this Heckit version of our model and get essentially the same results.

X_i is a vector that contains household i's demographics and its neighborhood characteristics; Z_i is a vector of instrument variables that includes X_i. The error terms ε and η are normally distributed with mean zero, and with a 5×5 covariance matrix:

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \sigma_{22} \end{pmatrix} \tag{3}$$

$\Sigma_{11}^{-1/2}\Sigma_{12}\sigma_{22}^{-1}$ gives the correlations between the endogenous density variable and the four dependent variables on vehicle ownership and usage, and measures the degree of endogeneity. We can rewrite Equations 1 and 2 in the following form:

$$\begin{pmatrix} y_i^* \\ D_i \end{pmatrix} = \begin{pmatrix} D_i & X_i & 0 \\ 0 & 0 & Z_i \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \\ \gamma \end{pmatrix} + \begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \tag{4}$$

Equation 4 can be simplified again as the following:

$$Y^* = X\phi + U \tag{5}$$

where $Y^* = (y_i^{*'}, D_i)'$, $X = diag((D_i, X_i^*), Z_i^*)$, $\phi = (\alpha', \beta', \gamma')'$, $U = (\varepsilon', \eta')'$.

Due to the discrete nature of the system, the likelihood function involves integrals of multivariate normal densities. In this paper, we use data augmented Gibbs sampling for limited dependent variable models to avoid direct evaluation of the likelihood function (Albert and Chib 1993, Li 1998, Fang 2008). There are three advantages of the approach used. First, using augmented latent variables avoids evaluation of the multivariate normal distributions and reduces computational costs. Second, it provides exact finite sample inference of the parameters and hence is free from the use of asymptotic approximations. Finally, we can easily take parameter uncertainty into account in deriving posterior and predictive densities for the function of interest (Li 1998).

We assume a normal prior for β: $N(\beta_0, V_0)$, and an Inverse-Wishart for Σ: $IW(\nu, Q)$, where $\beta_0, V_0, \nu,$ and Q are pre-specified prior parameters chosen to make the prior distributions diffuse.

These prior distributions are chosen for computational convenience, but the posterior distributions are not sensitive to any of the prior parameters. The Gibbs sampling procedure is as follows:

Step 1: draw y_i^* conditional on D_i, ϕ, Σ from multivariate truncated normal distribution

$$y_i^* \mid D_i, \phi, \Sigma : \; MVTN(\mu_{1|2}, \sigma_{1|2}) \tag{6}$$

where $\mu_{1|2} = D_i\alpha + X_i\beta + \Sigma_{11}\sigma_{22}^{-1}(D_i - Z_i\gamma)$, and $\sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.

Step 2: draw ϕ conditional on Y_i^*, Σ from multivariate normal distribution

$$\phi \mid Y_i^*, \Sigma : \; MVN(\bar{\phi}, \bar{V}) \tag{7}$$

where $\bar{V} = (V_0^{-1} + \sum_{i=1}^{T} X_i\Sigma^{-1}X_i)^{-1}$, and $\bar{\phi} = \bar{V}(V_0^{-1}\phi_0 + \sum_{i=1}^{T} X_i\Sigma^{-1}Y_i^*)$.

Step 3: draw Σ conditional on Y_i^*, ϕ from Inverse Wishart distribution

$$\Sigma \mid Y_i^*, \phi : \; IW(\nu + T, \sum_{i=1}^{T}(Y_i^* - X_i\phi)(Y_i^* - X_i\phi)' + Q) \tag{8}$$

In this paper, the instrumental variable is the average MSA residential density measured by housing units per square mile. The correlation between the average MSA residential density and the residential density at the census block level is .433.

The model system in equations (1) and (2) could be estimated by maximum likelihood methods, although given the multiple integrals in the likelihood function this would typically be done using simulation methods (see Train, 2003). We have chosen to use Bayesian methods for both computational and statistical reasons. Our Gibbs sampling procedure described above directly samples blocks of parameters and does not use any Metropolis-Hastings steps. It therefore runs very quickly, and sampling the parameters in blocks reduces the correlation between the draws. Maximum likelihood computation may be more difficult because the log-likelihood function is not convex in the correlation parameters (off-diagonal elements in Σ), and this requires manually restarting the optimization from different starting points to help find a global maximum.

Even if the maximum likelihood estimates are correctly calculated, there is still the problem of inference. Most software either uses some numerical approximation to the inverse Hessian of the log likelihood or the "sandwich estimator" favored by Train (2003). Unfortunately these two different methods can give very different estimates, and there is no way to distinguish between them using standard asymptotic theory. Even if the covariance estimates agree, there is still the problem of producing confidence regions for complex functions of the model parameters. Daly, Hess, and de Jong (2012) show that it is quite complicated to get valid confidence intervals for relatively simple functions of the underlying parameters such as willingness to pay measures. Their methods would be very difficult to implement for the policy simulations in Tables 3 and 4 or the predictions in Tables 6 and 7.

The Bayesian methods used in this paper have clear prescriptions for inference. Confidence regions are given by highest posterior density regions, and confidence regions for complex functions of parameters and data can easily be calculated by using the draws of parameters from the Gibbs sampling scheme described earlier in this section. It turns out that the highest posterior density regions for the parameters and policy simulations are symmetric and unimodal, so the intervals implied by posterior standard deviations reported in the tables in the rest of this paper are very good approximations to the highest posterior density regions.

Bayesian methods do require a choice of prior distribution, and they may not have good repeated sampling properties. Fortunately the inferences and estimates presented in this paper are not sensitive to different diffuse priors. We carried out some Monte Carlo studies on the model in Fang (2008), and these studies confirmed that the Bayesian procedures were very similar to maximum likelihood and had good repeated sampling properties. The model in Fang (2008) is very similar to the one used here, but it

treats density as exogenous and therefore omits equation (2). It is therefore likely that the methods used in this current paper also have good repeated sampling properties.

3 Data

We use data from the 2001 National Household Travel Survey (NHTS), a cross-section survey of a total of 69,817 households nationwide. Among them, 26,038 are in the national sample, and 43,779 are from nine add-on areas, states or local jurisdictions that purchased additional households in their jurisdiction to be interviewed and included in the NHTS for area-specific studies. This paper only includes households in the national sample, and we do not use the more recent 2009 NHTS because the 2009 survey did not include the two odometer readings required to get accurate utilization measures (Lave, 1994). By merging the household file, vehicle file and person file, we obtain a sample of 25,057 households that contain detailed information on households' demographics, various measures of land use density, vehicle properties including year, make, model, and complete estimates of annual miles traveled. Out of these 25,057 households, we randomly choose 5,863 households for estimation. The rest of the observations will be used for the out-of-sample forecast in Section 5. Households with missing information on various measures of density are dropped from the sample. Throughout the paper, we assume that whatever made people answer the survey is independent of density and vehicle choice, conditional on demographics. Hence the sample used for estimation can be seen as random. All of these data and associated documentation are available from http://nhts.ornl.gov/download.shtml.

Explanatory variables include density and household demographic characteristics. Density is measured by housing units per square mile at the census block level, which is highly correlated with population per square mile and jobs per square mile. To capture local transit networks and non-motorized facilities, an indicator of whether or not the MSA has rail, and the number of bicycles in the households are considered. Demographic variables include total household annual income, the highest education level achieved within a household, household size, number of adults, children's ages, home ownership, and urban/rural indictor of the residence area.

The summary statistics of the variables for the national sample and the sub-sample are listed in Table 1. Only 6 percent of the 25,057 national sample households own 3 or more cars, and only 4 percent of these households own 3 or more trucks. This partially justifies aggregating 2 or more cars and 2 or more trucks into one category. Note that the average variable values largely agree between the national sample and the randomly drawn sub-sample.

Note that there are very few vehicles in the estimation sample whose dual odometer readings show no change. These vehicles are either driven very infrequently or are held primarily as investments or collectibles. We left these vehicles in our samples since the results are unchanged when they are excluded. Unlike Brownstone and Golob (2009) we do not directly model fuel use even though it is relevant for many policies. The measures of fuel use in the NHTS sample are derived from matching vehicle efficiency based on reported year/make/model of each household vehicle. There are substantial missing data in these variables, so directly modeling fuel use reduces the sample size by about 20 percent and potentially biases the sample since the more vehicles in a household the more likely it is for at least one of them to be missing key information.

Table 1: Descriptive Statistics

Variables	National		Subsample	
	Mean	(Std.)	Mean	(Std.)
Observations	25,057		5,863	
Explanatory Variables				
Housing units/sq. mile (block)	1397	(1505)	1452	(1526)
Population/sq. mile (block)	3638	(4657)	3799	(4834)
Employment/sq. mile (tract)	1306	(1472)	1334	(1475)
Housing units/sq. mile (tract)	1217	(1367)	1254	(1388)
Population/sq. mile (tract)	3102	(4051)	3211	(4116)
Number of adults	1.91	(0.70)	1.88	(.71)
Number of children	.65	(1.05)	.65	(1.05)
Highest education achieved high school	30.6%		30.0%	
Highest education achieved bachelor	37.8 %		37.8%	
Youngest child under 6	14.6%		15.4%	
Youngest child between 6 and 15	17.2%		16.4%	
Youngest child between 15 and 21	5.9%		5.4%	
MSA has rail	22.1%		23.6%	
Resides in urban area (tract)	75.3%		77.1%	
Household income is between 20k and 30k	12.4%		12.2%	
Household income is between 30k and 50k	23%		22.0%	
Household income is between 50k and 75k	17.9%		17.4%	
Household income is between 75k and 100k	11%		10.3%	
Household income is greater than 100k	12%		12.7%	
Household owns home	80.1%		78.7%	
Vehicle Choice and Utilization				
Household owns no car	22.1%		22.4%	
Household owns one car	51.7%		51.9%	
Household owns two or more than two cars	26.2%		25.7%	
Household owns no truck	41.2%		43.6%	
Household owns one truck	38.2%		37.6%	
Household owns two or more than two trucks	20.6%		18.8%	
Average car miles per year conditional on owning cars	11,470	10,021	11,362	9,648
Average truck miles per year conditional on owning trucks	12,982	10,669	13,082	11,320

4 Estimation Results

Since we don't want to impose a priori the possible effects of residential density on household's vehicle type choice and utilization, we make the priors relatively noninformative. We set the variance of the normal prior to be large and prior degree of freedom of the Wishart to be small. Specifically, we set β_0 to be a vector of zeros, and V_0 to be a diagonal matrix with 100 on the diagonal, ν to be 10, and Q an identity matrix. We check the effect of the prior by increasing the prior variance of β to reflect the non-informativeness of the prior. Since results obtained from the noninformative priors are virtually the same

with the relatively noninformative prior mentioned above, we conclude that data information is predominant.

In the Gibbs Sampler, we take 20,000 iterations and burn in the first 2,000 to mitigate start up effects and use the remaining draws to get posterior inferences. We experimented with up to 100,000 iterations and discarding the first 10,000 but there were no differences from the results presented here. We also tried taking every 20th draw to reduce the impact of correlation between draws, but there were no differences suggesting that correlation is not a problem with this model and data. Table 2 lists the estimation results of the model. The five columns stands for the five equations estimated, with log of density at the census block level as dependent variable for the last equation. There is a close relationship between the possibly endogenous variable (the density at the census block level) and the instrument variable (average MSA density). Specifically, a 1 percent increase in the average MSA density is associated with approximately .57 percent increase in the density at the census block level.

The effects of household demographics have expected signs. Household size is positively correlated with number of trucks and truck utilization, and is negatively correlated with number of cars and car utilization. Meanwhile, as the number of adults increases, both numbers of cars and trucks and their utilizations increase. Since the number of children in a household equals household size less number of adults, the above observation shows that when the number of children increases, it is more likely for the family to own trucks. Recall that our definition of trucks includes SUVs and vans, and these vehicle types are useful for transporting children and their friends. Income has a significantly positive impact on vehicle holdings and utilization. Accessibility to public transit, such as rail, makes people choose fewer trucks and drive them less.

After obtaining posterior draws of the parameters, we calculate the marginal effects of density on vehicle choices for each household and present the average effects across households. Table 3 shows the mean and standard deviation of the probability changes for holding zero, one, and two or more cars/trucks with respect to changes in density. When density increases by 50 percent (a very large amount – see Downs, 2004, Chapter 12), the probability of not holding trucks increases by approximately 2.67 percentage points (equivalent to an arc elasticity of .053), and the probabilities of holding one truck and two trucks decrease by around 1.07 and 1.60 percentage points respectively (equivalent to arc elasticities .021 and .032). These changes are around two times bigger than those obtained in Fang (2008), in which only California data are used and endogeneity left uncorrected. In that study, when density increases by half, the probability of not holding trucks increases by approximately 1.2 percentage point, and the probabilities of holding one truck and two trucks decrease by around .75 and .46 percentage point respectively. Qualitatively, however, the two sets of results largely agree - residential density has a modest and statistically significant impact on truck ownership. If we further increase residential density to the extent that it doubles, the reduction in truck ownership deepens by modest 4.56 percentage points.

Residential density affects households' choice of cars with a much smaller scale and in a less significant way. When density increases by 50 percent, the probability of holding zero cars decreases by .47 percentage points, that of holding one car increases by .05 percentage points, while the probability of holding two or more cars increases by .42 percentage points.

Table 3 shows that the demand for car ownership is inelastic with respect to residential density, but the demand for truck ownership is relatively more elastic. The intuition is that the demand for vehicles is largely influenced by income, the life cycle of the family, number of children, and many factors other than residential density. As will be shown later, however, vehicle utilization is more susceptible to residential density variation. When we add the effects of vehicle ownership change and utilization reduction together, we found that residential density has a fairly large impact on energy consumption.

Note that the difference in density between Philadelphia (2561 units/sq. mile) and Los Angeles (3322) is about 25 percent, the difference between Phoenix (2317) and New York (3792) is about 50 percent, and the difference in density between Atlanta (1180) and Phoenix is about 100 percent.

Table 2: Coefficient Estimates

Variable	Coefficient				
	number of cars	number of trucks	annual avg car miles (in 1,000s)	annual avg truck miles (in 1,000s)	Log of block density
log(block density)	0.0375	-0.1969	0.0342	-3.2304	-
	(0.0433)	(0.0455)	(0.4929)	(0.6602)	-
Number of bikes	-0.0273	0.1093	-0.1293	1.2140	0.0138
	(0.0130)	(0.0130)	(0.1480)	(0.2097)	(0.0127)
Household size	-0.1204	0.0980	-1.1827	2.1654	-0.0317
	(0.0270)	(0.0274)	(0.3115)	(0.4278)	(0.0262)
Number of adults	0.3239	0.1671	3.5415	1.5293	-0.0113
	(0.0346)	(0.0358)	(0.4002)	(0.5610)	(0.0336)
Urban	-0.0039	0.1747	-1.0355	3.1176	2.4098
	(0.1250)	(0.1298)	(1.4218)	(1.9134)	(0.0385)
Income between 20k and 30k	0.1255	0.3805	1.1598	5.5918	-0.0032
	(0.0561)	(0.0614)	(0.6343)	(0.9864)	(0.0532)
Income between 30k and 50k	0.1554	0.5828	2.4567	8.7760	-0.0686
	(0.0501)	(0.0556)	(0.5693)	(0.8782)	(0.0483)
Income between 50k and 75k	0.1347	0.7135	2.7229	11.8910	-0.1108
	(0.0553)	(0.0603)	(0.6334)	(0.9540)	(0.0539)
Income between 75k and 100k	0.3262	0.6780	4.2178	11.4340	-0.1700
	(0.0655)	(0.0697)	(0.7414)	(1.1015)	(0.0641)
Income greater than 100k	0.2539	0.7526	3.9113	12.8280	-0.3294
	(0.0660)	(0.0700)	(0.7490)	(1.1065)	(0.0646)
Income data missing	0.2381	0.2795	0.6552	3.7614	-0.1050
	(0.0650)	(0.0731)	(0.7459)	(1.1589)	(0.0631)
Owns home	0.0675	0.3937	-0.4018	3.3768	-0.3576
	(0.0423)	(0.0458)	(0.4828)	(0.7257)	(0.0372)
MSA has rail	0.0598	-0.1962	0.2095	-2.0256	-0.0203
	(0.0421)	(0.0449)	(0.4758)	(0.7046)	(0.0413)
Highest education: high school	0.1008	-0.0022	1.1975	0.6450	0.0217
	(0.0385)	(0.0402)	(0.4415)	(0.6449)	(0.0375)
Highest education: Bachelor	0.2265	-0.1654	2.5117	-1.1363	0.1622
	(0.0421)	(0.0441)	(0.4815)	(0.7033)	(0.0403)
Youngest child under 6	0.1033	0.1264	2.4547	2.1375	-0.0254
	(0.0711)	(0.0730)	(0.8176)	(1.1478)	(0.0695)
Youngest child between 6 and 15	0.1197	0.0873	2.1364	1.3270	-0.0418
	(0.0634)	(0.0649)	(0.7299)	(1.0186)	(0.0619)
Youngest child between 15 and 21	0.0779	-0.1235	2.0036	0.4597	-0.0193
	(0.0683)	(0.0717)	(0.7839)	(1.1416)	(0.0685)
log(average MSA Density)	-	-	-	-	0.5743
	-	-	-	-	(0.0244)

Notes: The base groups are households with income below 20k, do not own home, are high school dropout, have no children, and live in rural area. Posterior standard deviations are reported in parentheses;

Table 3: Changes in vehicle choice when block density increases

Percent changes in density	Probability changes for truck choice		
	Δ P(tnum=0)	Δ P(tnum=1)	Δ P(tnum ≥ 2)
10 %	.0063	-.0024	-.0038
	(.0014)	(.0005)	(.0009)
25 %	.0147	-.0058	-.0089
	(.0032)	(.0012)	(.0020)
50 %	.0267	-.0107	-.0159
	(.0058)	(.0023)	(.0035)
100%	.0456	-.0190	-.0265
	(.0099)	(.0042)	(.0058)
Percent changes in density	Probability changes for car choice		
	Δ P(cnum=0)	Δ P(cnum=1)	Δ P(cnum ≥ 2)
10 %	-.0011	.0001	.001
	(.0013)	(.0002)	(.0011)
25 %	-.0026	.0003	.0023
	(.0030)	(.0004)	(.0026)
50 %	-.0047	.0005	.0042
	(.0054)	(.0007)	(.0048)
100%	-.0080	.0008	.0072
	(.0092)	(.0010)	(.0083)

Notes: posterior standard deviations are reported in parentheses

Table 4 shows that changes in density do not seem to affect car utilization. Annual average miles driven in cars by a household would only increase by around 14 miles when housing units per square mile increases by 50 percent. Even when the housing density doubles, the annual average car utilization would merely increase by about 24 miles. On the contrary, annual average miles of trucks respond more sharply to density changes. When housing units per square mile increases by 50 percent, utilization of truck would decrease by approximately 610 miles, with a standard deviation of about 118 miles. This effect is in the same scale as that found in Fang (2008), in which a 50 percent increase in density will reduce truck utilization by about 562 miles. Doubling the residential density would reduce annual average truck miles by about 1004 miles, which is a 13.6-percent reduction in truck utilization and equivalent to a .136 arc elasticity.

Table 4: Changes in vehicle miles when density increases

	Δ car miles	% Δ car miles	Δ truck miles	% Δ truck miles
10 %	3.23	.04	-149.63	-2.03
	(46.29)	(.53)	(29.76)	(.40)
25 %	7.63	.08	-344.34	-4.67
	(108.34)	(1.23)	(67.61)	(.92)
50 %	14.02	.16	-610.5	-8.27
	(196.79)	(2.23)	(117.66)	(1.59)
100%	24.37	.28	-1003.6	-13.6
	(336.14)	(3.82)	(187.23)	(2.54)

Notes: posterior standard deviations are reported in parentheses

We can also obtain an approximation of residential density's marginal effect on energy consumption using vehicle fuel efficiency data and density's marginal effect on vehicle type choice and utilization. In our sample, average fuel efficiency of cars is 21.8 miles per gallon, and average fuel efficiency of trucks is 16.6 miles per gallon. The 5863 households in our sample drive a total of 74 million car miles and 61 million truck miles per year, equivalent to a total consumption of 3.4 million gallons by car usage and 3.7 million gallons by truck usage. When density doubles, we redistribute cars and trucks among the 5863 households using probability changes presented in Table 3. Because we classify number of vehicles equal or larger than two as one group, the redistribution of cars/trucks among families with cars/trucks exceeding quantity one is done based on the assumption that the percentage of two, three, etc., vehicles in the group remain constant before and after the density change. This assumption is conservative because one would expect the vehicle number distributed more towards smaller numbers when density increases. By holding constant the vehicle distribution for households with two or more vehicles, we provide conservative (lower magnitude) estimate of the marginal effect of density increase. Average car/truck miles after the density increase can be easily calculated using the percentage changes in vehicle miles presented in Table 4. With the new distribution of cars and trucks among the households in the sample, and new average car/truck miles, we calculate the total energy consumption by the 5863 households after the density doubling to be 3.4 million gallons by car usage and 2.2 million gallons by truck usage. The energy usage of cars barely changes at all by increasing about 1.8 percent, and the energy usage of trucks decreases by about 40.7 percent (corresponding to an arc elasticity of .41). This amounts to a substantial reduction of 1.4 million gallons, or 20 percent, of total gasoline consumption by vehicle usage.

Table 5 shows the correlation matrix of the structural error matrix Σ. We find that the unobserved characteristics affecting number of cars held and number of trucks held have a negative correlation of -.40. The correlation between miles driven by cars and miles driven by trucks is -.15. This indicates a substitution effect between cars and trucks, not only type-wise but also usage-wise. The unobserved characteristics that make people to live in dense areas also tend to make people choose more trucks, and drive more truck miles. The correlation, controlled for observed characteristics, between density and the number of trucks is .09 with a standard deviation of .051, and that between density and average truck miles is .1 with a standard deviation of .044. Hence we conclude that controlling for the endogeneity of the density variable is necessary in the estimation.

Table 5: Correlation Matrix of Structural Errors (Σ)

	number of cars	number of trucks	avg car mile	avg truck mile	density
number of cars	1.00	-	-	-	-
number of trucks	-.40	1.00	-	-	-
	(.014)				
avrg car mile	.53	-.29	1.00	-	-
	(.011)	(.015)			
avrg truck mile	-.31	.59	-.15	1.00	-
	(.015)	(.011)	(.015)		
density	-.016	.09	-.04	.1	1.00
	(.049)	(.051)	(.046)	(.044)	

Notes: Highest posterior standard deviations are reported below each correlation

5. Prediction

As a robustness check, we carry out the out-of-sample forecast of vehicle choice and utilization for random observations from the rest of the national sample. Generally, the Bayesian predictive probability distribution function of the future observable dependent variable $\mathbf{y}^p$ can be expressed as the following,

$$f(\mathbf{y}^p \mid \mathbf{y}) = \iint f(\mathbf{y}^p \mid \mathbf{y}, \beta, \Sigma)\, f(\beta, \Sigma \mid \mathbf{y})\, d\beta\, d\Sigma \qquad (9)$$

where $\mathbf{y}$ is the in-sample data used for estimation, and $f(\beta, \Sigma \mid \mathbf{y})$ is the posterior distribution of the parameters. Since Equation 9 cannot be solved analytically, one may use the following strategy (Koop 2003) in the same fashion of a Markov Chain Monte Carlo to obtain draws of $\mathbf{y}^p$ that can be considered to be from the predictive probability distribution:

Step 1: Get draws of β^s, Σ^s from the posterior $f(\beta, \Sigma \mid \mathbf{y})$. In this case, they are simply draws from the Gibbs Sampler from the in-sample estimation.

Step 2: Draw $\mathbf{y}^{ps}$ from a multivariate Normal distribution of $MVN(X\beta^s, \Sigma^s)$.

With sequence of random draws of $\mathbf{y}^{ps}$, we can obtain the mean and standard deviation of its predictive distribution. One complication with the prediction in this paper is that the dependent variables are not continuous, but limited. Therefore, additional steps are needed to obtain the quantitative probabilistic predictions for vehicle ownership. For example, if we would like to predict the probability of having zero car for a particular household, we obtain the probability that the latent utility towards having zero car, $y_1^{p*} < 0$, from the following:

$$\text{Prob}(y_1^{p*} < 0 \mid y) = \int_{-\infty}^{0} f(y_1^{p*} \mid y)\, dy_1^{p*}$$

(substitute in Equation 9) $$\int_{-\infty}^{0} \left(\iint f(y_1^{p*} \mid y, \beta, \Sigma)\, f(\beta, \Sigma \mid y)\, d\beta\, d\Sigma \right) dy_1^{p*}$$

(Fubini´s Theorem) = $$\iint\left(\int_{-\infty}^{0} f(y_1^{p*} \mid y,\beta,\Sigma)dy_1^{p*}\right) f(\beta,\Sigma \mid y)d\beta d\Sigma$$

= $$\iint \text{Prob}(y_1^{p*} < 0 \mid y,\beta,\Sigma)\, f(\beta,\Sigma \mid y)d\beta d\Sigma$$

The steps needed to calculate the above probability are:

Step 1: Get draws of β^s, Σ^s from the posterior $f(\beta,\Sigma \mid \mathbf{y})$.

Step 2: Calculate $P^s = \Phi(\frac{-X_1\beta^s}{\sigma_{11}^s})$.

Step 3: Averaging across all the probability draws, $\text{Prob}(y_1^{p*} < 0 \mid y) \approx \frac{1}{N}\sum_{s=1}^{N} P^s$.

Calculation for the other predictive probabilities follows the same procedure. A number of random samples are taken to perform the prediction, and the forecast results from which all follow the same pattern. Table 6 lists the actual and predicted number of households that hold zero, one, and two or more cars/trucks for a random sample of 101 and a random sample of 4991 observations. The prediction for zero car, one car, one truck, and two and more trucks are in the ball-park of the actual values, taking standard deviations into account. But the model consistently underestimates the number of households for holding two or more cars and overestimates the number of households not holding trucks.

Table 6: Predicted number of households

	c=0	c=1	c ≥ 2	t=0	t=1	t ≥ 2
			Random sample of 101 obs.			
Predicted number of households	26	54	21	50	35	16
(standard deviation)	(.6)	(.7)	(.5)	(.6)	(.7)	(.6)
True number of households	24	49	28	49	33	19
			Random sample of 4991 obs.			
Predicted number of households	1301	2677	1013	2413	1774.6	804
(standard deviation)	(28.8)	(33.8)	(25.5)	(29.7)	(34.9)	(25.9)
True number of households	1060	2601	1330	2165	1884	942

Forecasts for vehicle miles perform much better than those for vehicle type choice aforementioned, as are shown in Table 7. The predicted average miles are more accurate for a random sample of 4,991 households than for that of 101 households, presumably due to simulation errors, as reflected by the difference in standard deviations. For a sample of 101 households, the predicted car utilization is 9,155 miles, 16 miles less than the true value, and the predicted truck utilization is 7,592 miles, less than two standard deviations away from the true value. For a random sample of 4,991 households, the predicted average miles driven by cars is 9,114, 21 miles less than the actual value observed; the predicted average miles driven by truck is 7,649, 445 miles higher than the actual value.

Table 7: Predicted average miles driven for households in the sample

	average miles by cars	average miles by trucks
	Random sample of 101 obs.	
Forecast	9155.6	7592.4
(standard deviation)	(927.6)	(1018.7)
True	9171.9	5882.2
	Random sample of 4991 obs.	
Forecast	9113.6	7649.3
(standard deviation)	(178.9)	(210.6)
True	9135	7204.4

It is difficult to interpret the results of the out of sample predictions discussed above. Ideally we would like the posterior forecast intervals to always contain the true values, but failure to reach this ideal does not necessarily imply that the model is performing worse than other models used for this type of work. Until other models are subjected to these out of sample forecasting exercises it will be difficult to judge the results.

6. Conclusion

This paper extends the model in Fang (2008) to include the possibility of unobserved factors that affect both vehicle choice and density choice - an endogeneity problem that might bias the estimation results. We control for part of this by using disaggregate data and detailed household characteristics. More importantly, we utilize an instrument variable, average MSA density, in the estimation to correct for the endogeneity. We apply this model to the 2001 NHTS survey data, and we find statistically significant error correlations indicating endogeneity bias. However, the magnitude of this bias is small and our results are qualitatively and quantitatively similar to Fang (2008) who assumed zero error correlations. This finding of essentially no error correlation between density and vehicle miles corroborates similar findings using a different model structure in Brownstone and Golob (2009). However if detailed household characteristics are not controlled, then these models have larger error correlations and substantial endogeneity bias.

The results show that even a very large increase in residential density has a negligible effect on car choice and utilization, but slightly reduces truck choice and utilization. Since trucks are considerably less efficient than cars due to differences in fuel economy regulations in the U.S., fuel consumption is reduced by a larger amount. The changes in residential density used in our policy simulations are very large, and it is very unlikely that these changes will occur except in isolated new developments. The Bayesian confidence intervals are quite narrow, so these results are precisely estimated. To further test the robustness of the model, we perform forecasting on a number of random samples from the population. We find that the predicted values are largely consistent with the true values, more so for vehicle utilization than vehicle choice, confirming the robustness of the model used.

The model used here only looks at the choice of cars and trucks, but U.S. fuel economy standards imply that this split is responsible for most of the differences in fuel economy. Fang (2008) extended the model to split trucks and cars into large and small subcategories, but the qualitative and quantitative results were not changed. The New York MSA is frequently an outlier in studies of vehicle use due to its high density and high share of transit use. The appendix re-estimates our model excluding the New York MSA, and we find that our results are essentially unchanged. This suggests that the socio-demographic

variables included in our model effectively capture the differences between New York and the rest of the country.

7 Acknowledgments

The authors gratefully acknowledge financial support from the University of California, Irvine School of Social Sciences and the University of California Transportation Center. Two anonymous referees and Kara Kockelman provided many useful comments on an earlier draft, and Phillip Li provided excellent research assistance. The authors bear sole responsibility for any errors.

References

Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* 88: 669–679. doi:10.1080/01621459.1993.10476321

Bento, A.M., Cropper, M.L., Mobarak, A.M., Vinha, K., 2005. The effect of urban spatial structure on travel demand in the United States. *Review of Economics and Statistics* 87(3): 466-478. doi:10.1162/0034653054638292

Bento, A.M., Lawrence H. Goulder, Mark R. Jacobsen, and Roger H. von Haefen. 2009. "Distributional and Efficiency Impacts of Increased US Gasoline Taxes." *American Economic Review* 99(3): 667–99. doi:10.1257/aer.99.3.667

Brownstone, D., Golob, T.F., 2009. The impact of residential density on vehicle usage and energy consumption. *Journal of Urban Economics* 65(1): 91-98. doi:10.1016/j.jue.2008.09.002

Brueckner, J., Largey, A., 2008. Social interaction and urban sprawl. *Journal of Urban Economics* 64(1): 18-34. doi:10.1016/j.jue.2007.08.002

Cervero, R., Kockelman, K., 1997. Travel demand and the 3Ds: density, diversity and design. *Transportation Research D* 2(3): 199-219. doi:10.1016/S1361-9209(97)00009-6

Daly, A., Hess S. and de Jong, G., 2012. Calculating errors for measures derived from choice modelling estimates, *Transportation Research B:* 46(2), 333-341. doi:10.1016/j.trb.2011.10.008

Downs, A. 2004. *Still stuck in traffic: coping with peak-hour traffic congestion*, The Brookings Institution, Washington, D.C.

Dunphy, R., Fisher, K., 1996. Transportation, congestion, and density: new insights. *Transportation Research Record* 1552(1): 89-96. doi:10.3141/1552-12

Evans, W., Oates, W., Schwab, R., 1992. Measuring peer group effects: A study of teenage behavior. *Journal of Political Economy* 100(5): 966-991. doi:10.1086/261848

Ewing, R., Cervero, R., 2001. Travel and the built environment. *Transportation Research Record*, 1780(1): 87-114. doi:10.3141/1780-10

Fang, A., 2008. A discrete-continuous model of households' vehicle choice and usage, with an application to the effects of residential density. *Transportation Research B* 42(9): 736-758. doi:10.1016/j.trb.2008.01.004

Kim, J., Brownstone, D., 2013. The impact of residential density on vehicle usage and fuel consumption: Evidence from national samples. *Energy Economics* 40: 196-206. doi:10.1016/j.eneco.2013.06.012

Koop, G., 2003. *Bayesian Econometrics*. John Wiley & Sons.

Lave, C., 1994. State and national VMT estimates: it ain't necessarily so. University of California Transportation Working Paper accessed at http://escholarship.org/uc/item/5527j8dj.

Li, K., 1998. Bayesian inference in a simultaneous equation model with limited dependent variables. *Journal of Econometrics* 85(2): 387-400. doi:10.1016/S0304-4076(97)00106-1

Nandram, B., Chen, M., 1996. Reparameterizing the generalized linear model to accelerate Gibbs sampler convergence. *Journal of Statistical Computation and Simulation* 54(1-3): 129-144. doi:10.1080/00949659608811724

Train, K.E., 2003. *Discrete choice methods with simulation*. Cambridge, UK: Cambridge University Press.

Webb, E.L., Forster, J.J., 2008. Bayesian model determination for multivariate ordinal and binary data. *Computational Statistics and Data Analysis*, 52(5), 2632-2649. doi:10.1016/j.csda.2007.09.008

Appendix: Estimation of Tables 3, 4, and 5 excluding the New York MSA:

Table 8: Coefficient Estimates

Variable	Coefficient				
	number of cars	number of trucks	annual avg car miles (in 1,000s)	annual avg truck miles (in 1,000s)	Log of block density
log(block density)	0.0492	-0.2039	0.2201	-3.1961	-
	(0.0445)	(0.0480)	(0.5072)	(0.6641)	-
Number of bikes	-0.0303	0.1085	-0.1281	1.1545	0.0151
	(0.0133)	(0.0135)	(0.1530)	(0.2072)	(0.0129)
Household size	-0.1351	0.1015	-1.2631	1.9227	-0.0235
	(0.0277)	(0.0283)	(0.3201)	(0.4275)	(0.0270)
Number of adults	0.3463	0.1595	3.6115	1.6839	-0.0221
	(0.0363)	(0.0371)	(0.4163)	(0.5654)	(0.0351)
Urban	-0.0473	0.2114	-1.6378	3.2418	2.4392
	(0.1300)	(0.1396)	(1.4905)	(1.9448)	(0.0391)
Income between 20k and 30k	0.1107	0.3987	1.1120	5.7851	0.0024
	(0.0575)	(0.0629)	(0.6628)	(0.9794)	(0.0555)
Income between 30k and 50k	0.1339	0.5919	2.4171	8.7623	-0.0674
	(0.0517)	(0.0561)	(0.5955)	(0.8643)	(0.0496)
Income between 50k and 75k	0.1079	0.7342	2.6215	11.6860	-0.1051
	(0.0569)	(0.0615)	(0.6524)	(0.9417)	(0.0558)
Income between 75k and 100k	0.3102	0.6788	4.1733	11.3900	-0.1832
	(0.0677)	(0.0721)	(0.7723)	(1.1044)	(0.0661)
Income greater than 100k	0.2307	0.7533	3.9036	12.6610	-0.2888
	(0.0683)	(0.0722)	(0.7830)	(1.0999)	(0.0666)
Income data missing	0.2240	0.2695	0.6353	3.4862	-0.1083
	(0.0678)	(0.0747)	(0.7795)	(1.1687)	(0.0647)
Owns home	0.0550	0.4076	-0.4725	3.3606	-0.3448
	(0.0427)	(0.0473)	(0.4946)	(0.7259)	(0.0380)
MSA has rail	0.0627	-0.1910	0.5114	-2.1291	0.0101
	(0.0447)	(0.0487)	(0.5135)	(0.7320)	(0.0434)
Highest education: high school	0.1128	-0.0117	1.2880	0.5475	0.0274
	(0.0397)	(0.0415)	(0.4586)	(0.6454)	(0.0384)
Highest education: Bachelor	0.2219	-0.1589	2.4910	-1.0394	0.1605
	(0.0431)	(0.0450)	(0.4969)	(0.6940)	(0.0420)
Youngest child under 6	0.1368	0.1083	2.4931	2.3481	-0.0342
	(0.0734)	(0.0755)	(0.8450)	(1.1509)	(0.0713)
Youngest child between 6 and 15	0.1501	0.0765	2.3366	1.4450	-0.0427
	(0.0651)	(0.0671)	(0.7497)	(1.0135)	(0.0636)
Youngest child between 15 and 21	0.0959	-0.1383	2.2486	0.3411	-0.0269
	(0.0701)	(0.0736)	(0.8226)	(1.1326)	(0.0704)
log(average MSA Density)	-		-	-	(0.5690)
	-	-	-	-	0.0246

Notes: The base groups are households with income below 20k, do not own homes, are high school dropouts, have no children, and live in a rural area. Posterior standard deviations are reported in parentheses;

Table 9: Changes in vehicle choice when block density increases

Percent changes in density	Probability changes for truck choice		
	Δ P(tnum=0)	Δ P(tnum=1)	Δ P(tnum ≥ 2)
10 %	.0065	-.0024	-.004
	(.0014)	(.0005)	(.0009)
25 %	.0152	-.0058	-.0093
	(.0034)	(.0012)	(.0021)
50 %	.0276	-.0109	-.0167
	(.0061)	(.0024)	(.0038)
100%	.0471	-.0193	-.0278
	(.0104)	(.0043)	(.0062)

Percent changes in density	Probability changes for car choice		
	Δ P(cnum=0)	Δ P(cnum=1)	Δ P(cnum ≥ 2)
10 %	-.0014	.0002	.0013
	(.0013)	(.0002)	(.0011)
25 %	-.0034	.0005	.0030
	(.0031)	(.0004)	(.0027)
50 %	-.0062	.0008	.0054
	(.0056)	(.0007)	(.0050)
100%	-.0105	.0011	.0094
	(.0094)	(.0010)	(.0085)

Notes: posterior standard deviations are reported in parentheses

Table 10: Changes in vehicle miles when density increases

	Δ car miles	% Δ car miles	Δ truck miles	% Δ truck miles
10 %	20.64	.23	-153.01	-2.07
	(47.52)	(.54)	(30.66)	(.42)
25 %	48.40	.55	-352.15	-4.77
	(111.29)	(1.26)	(69.61)	(.94)
50 %	88.10	.10	-624.33	-8.46
	(202.31)	(2.30)	(121.06)	(1.64)
100%	151	1.71	-1026.1	-13.90
	(345.97)	(3.91)	(192.69)	(2.61)

Notes: posterior standard deviations are reported in parentheses