CONGESTION, GAS TAXES AND VEHICLE CHOICE

SAM FLANDERS AND MELATI NUNGSARI

University of North Carolina at Chapel Hill

September 2014[1]

Abstract. Road congestion imposes large costs on individuals since long commutes yield significant decreases in productivity and leisure times. Congestion may also have ambiguous impacts on environmental pollution, either increasing it relative to a congestion-free regime through more frequent and longer vehicle usage, or decreasing it due to forgone travel. In this paper, we study the effects of gasoline tax policies on road congestion. To do this, we develop a model of household vehicle choice utilizing individual-level data from the 2009 National Household Travel Survey and combine it with a model of congestion, measured by average road speeds, which utilizes road-level data on traffic congestion collected by state and national-level departments of transportation. We estimate counterfactual regimes in which gas taxes are at different levels in order to answer questions regarding optimal gas taxes for a fixed geographical area.

1. Introduction

Household decisions on vehicle purchases and vehicle use have enormous impact on individuals' daily lives, and the aggregate of these decisions has major economic and environmental consequences. For this reason, there has been a great deal of research into how individuals make travel decisions. A common line of research, pursued by the like of West (2004), Parry and Small (2005), and Feng, Fullerton, and Gan (2005) has been to evaluate how travel decisions relate to environmental impacts. Researchers study how Pigovian gasoline taxes influence vehicle choices and what an optimal gas tax rate might be. However, there is very little analysis of how changes in policy variables like gasoline taxes influence traffic congestion. As with better-studied environmental issues, questions about how gasoline taxes affect congestion and what an optimal gas tax rate might be in light of congestion externalities have great policy relevance. For example, a lower gasoline price may lead to higher vehicle utilization and thus higher congestion, which can lead to less fuel efficiency, attenuating the benefits consumers get from the lower gas prices. Additionally, there is an enormous time cost to congestion. With car travel taking up significant portions of many individuals' waking hours, slower road speeds and correspondingly longer commutes will yield significant decreases in productivity or leisure time, and individuals may forgo some trips they would otherwise take if traffic is too

[1]We'd like to thank Brian McManus, Donna Gilleskie, Gary Biglaiser, and the UNC micro theory workshop group for their indispensable advice and suggestions.

severe, losing out on whatever value they would have obtained from that travel. Congestion may also have ambiguous impacts on vehicle pollution, either increasing it relative to a congestion-free regime through longer travel times and higher gasoline use or decreasing it due to forgone travel. Unfortunately, in the vehicle choice literature, only Parry and Small (2005) attempt to answer some of these questions, and they use a simple macro style calibrated model with no micro data on household behavior or road congestion. Their results are interesting and suggestive, but are far from dispositive on the relationship between gas taxes, household decisions, and congestion.

In this paper, we attempt to answer these questions for the US using rich individual-level data on household vehicle choices supplied by the National Household Travel Survey and road-level data on traffic congestion collected by state and national level departments of transportation. We start with a model of household vehicle choices based heavily on recent simultaneous extensive and intensive margin models of vehicle choice and use such as Spiller (2012) and Bento et al. (2009), and combine this with a simple model of road congestion that, for each segment of roadway we observe, predicts vehicle speed as a function of the density of vehicles on the road, as well as several other factors. We need two models because of the simultaneity between household travel choices and congestion. Households will make decisions of whether and how much to drive based on the time costs of travel, but the aggregate of those household decisions will in fact be the overall level of road utilization that determines congestion. Thus, if we want to predict household behavior in counterfactual policy regimes, we must model both sides of this system to recover accurate estimates. Using these two models in tandem, we can predict household-level vehicle use decisions, which can be aggregated to predict overall the overall density of vehicles on the road. Our road congestion model will then return road speeds corresponding to the given level of vehicle use. These road speeds can then be plugged back into the household choice model, which will yield new vehicle use decisions based on the new prevailing time costs of travel. Iterating this procedure should allow us to asymptotically approach a fixed point where vehicle choices are consistent with the prevailing level of congestion.

Our first goal in this paper will be to perform counterfactual estimates of congestion and household choices for different gas taxes, allowing us to study the relationship between gas taxes, household travel choices, and congestion. We can then compute an optimal gas tax in consideration of the externalities imposed by congestion and vehicle pollution. While the model in this draft does not include pollution, this will be an easy component to add, as there are is a wealth of previous literature incorporating pollution into vehicle choice models. Also, unlike congestion, we don't have a simultaneity issue–we can simply use vehicle-level average emissions data and the model's prediction of vehicle utilization to predict aggregate emissions. We can also derive estimates of the price elasticity of demand for gasoline, using our detailed model of road

congestion to explicitly model local variations in time costs of travel in order to see if these local effects yield an elasticity estimate consistent with the previous literature or not.

The remainder of this paper is organized as follows: Section 2 provides a literature review. Section 3 presents theoretical models for the household decision and road congestion problems. Section 4 describes the data sets used in this analysis. Section 5 shows the methods of estimation. Section 6 presents the results of these estimations and Section 7 concludes.

2. Literature Review

This paper builds on several rich literatures. In terms of estimation, Manski (1975) introduced the semiparametric maximum score estimation technique, allowing estimation of the parameters of a utility function through revealed preference. Estimation is accomplished by sampling sets of choice alternatives, only one of which is picked, and finding the parameters that maximize the likelihood of the observed decisions. Fox (2007) modifies Manski's technique to focus only on pairwise comparisons. That is, he selects an observed choice and a single alternative that was not chosen. Sampling many such pairs, he also finds the parameters that maximize the likelihood. This method is specifically designed to deal with large choice sets. Instead of analyzing all the possible choices, only pairs of choices are analyzed, and typically the entire sample of pairs of choices used in estimation will be an insignificant fraction of the overall choice space. While this method throws out a huge amount of information, Fox shows that it performs very well against standard logit estimators. For durable good joint demand and usage estimation, Dubin and McFadden (1984) provided the basic framework, using a sequential method of first estimating the extensive margin (purchase choice) and then estimating the intensive margin (usage decision) utilizing a nested logit model and a selection correction similar to Heckman.

This paper also follows a long line of research into how individuals and households choose vehicles and make driving decisions. There is a well known literature in IO dealing with vehicle choice and corresponding vehicle market shares. Most importantly, Berry, Levinsohn, and Pakes (BLP) (1995) use a random effects logit model to estimate demand for automobiles using aggregate data. Petrin (2001) modifies BLP to use demographic data as well as market level data.

More directly related to our model are papers that attempt to estimate both vehicle choice and vehicle utilization in terms of miles traveled. West (2004) is an example of a large literature from the 1980's through the 2000's adapting Dubin-McFadden to model vehicle purchase/miles traveled behavior. Common to these older papers are the limitations of Dubin-McFadden such as the imposition of correlation between the intensive and extensive margins. Also, these papers aggregate vehicle characteristics to an extremely coarse level to deal with large choice set problem, and often treat car purchase decisions as independent–agents

are not allowed to take their other vehicles into account when selecting a new vehicle. Another unattractive modeling decision, adopted in lieu of the vehicle independence assumption or in addition to it, is to drop all households with 3 or more cars. Because the number of choices grows explosively with larger bundles, removing all large bundles can dramatically decrease the computational difficulties. However, this means that a large portion of the sample must be dropped.

More recent work addresses many of these issues. Feng, Fullerton, and Gan (2005) Specify a log-linear equation for VMT as a function of various household, vehicle, and personal characteristics, and then use the Hausman (1981) method to recover a corresponding indirect utility function. This allows them to simultaneously estimate both margins of the choice problem without recourse to a two-step model. However, they only include two vehicle categories. Bento et al. (2008) use the same simultaneous equation framework, but use a random-parameter Bayesian estimation technique and a rich set of vehicle choices. However, they impose the independent vehicle choice assumption. Finally, Spiller (2012) utilizes the same simultaneous equation framework, along with a rich set of vehicles and choice at the bundle level rather than iterated independent vehicle choices. She uses the Fox (2007) maximum score technique to deal with the large choice set induced by these modeling decisions. This paper largely forms the basis for the household choice side of our model.

There is very little literature dealing with the relationship between congestion and gas taxes. The only major paper studying this is Parry and Small (2005), which derives optimal gas taxes in consideration of externalities imposed by pollution, car accidents, and congestion. They analyze this question by creating a macro-style theoretical model and plugging in estimates for each coefficient derived from various papers in the literature.

Finally, the road congestion model in this paper can also be situated in an existing literature, but not one in economics. Most of the research on modeling road congestion is found in other fields, such as urban planning, engineering, and physics. While this paper uses some basic theoretical concepts from this literature, such as the macroscopic fundamental diagrams (Geroliminis and Daganzo (2009)), most of the modeling work is not applicable. Typically, researchers use extremely high quality data on a small region such as a district of a city. They will utilize observations of traffic flow on the entire road network–every road in the area–and will make use of the explicit network structure of traffic behavior, modeling where and how bottlenecks occur in the road network. In many cases they'll use agent based modeling, explicitly simulating individual cars traveling the road network. These methods are very complex, have a high computation burden, and require extremely detailed data. When dealing with a small geographic area, these issues are tractable, but in a national level model like ours they are infeasible.

3. Model

3.1. **Vehicle Choice and Vehicle Miles Traveled.** For the household's problem, we employ a method used in a number of recent papers in the vehicle demand literature–a simultaneous, one-step procedure for estimating the intensive and extensive margins of vehicle use. Instead of specifying separate ad hoc equations for the indirect utility of owning a bundle of vehicles and for VMT for each vehicle in the bundle as in the Dubin-McFadden method, we define an equation for VMT and recover the corresponding indirect utility from that, using the differential equation method of Hausman (1981).

Before we detail the equations, we'll discuss timing. We're modeling vehicle purchase decisions as well as usage old decisions, so ideally we'd like to observe purchase decisions across time and model each decision separately. Unfortunately, there is no good panel data on household travel choices, so we use a cross-sectional survey, and thus we use a single period model where purchase decisions are made over the entire bundle of cars owned by the household, rather than individually over time. Thus the timing of the model is as follows:

(1) Households sell all their vehicles for their (used) blue book value.
(2) Household i observes idiosyncratic preference shocks ϵ_{ik} for each vehicle k, as well as all the other independent variables.
(3) Households choose a bundle of vehicles to purchase, completely aware of the vehicle use decisions and corresponding utilities that will result from each bundle choice.
(4) Households choose their level of use for each vehicle.

Note that steps three and four can equivalently be merged into one step.

Now we must specify an equation giving VMT as a function of various household, vehicle and price variables. To facilitate comparison with the existing literature, we use a log-linear model, adapted from Spiller (2012). For a household i with vehicle bundle J_i, the miles traveled using vehicle j is given by

$$VMT_{ij} = \frac{\beta_{ij}}{\sum_{k \in J_i} \beta_{ik}} exp\left(\sum_{k \in J_i} (\alpha_{ik} - \beta_{ik} \frac{p_i^g}{mpg_k}) + \lambda_i (y_i - \sum_{k \in J_i} p_k^s) \right)$$

where $\alpha_{ik} = \alpha z_{ik}^{\alpha}$, $\beta_{ik} = exp(\beta z_{ik}^{\beta})$, $\lambda_i = exp(\lambda z_{ik}^{\lambda})$. p_i^g is the price of gasoline per gallon in household i's block group. mpg_k is the fuel efficiency of vehicle k in miles per gallon. y_i is household i's income in the counterfactual case where it sells all vehicles. p_k^s is the sale value of the (used) vehicle k. The $z_{ik}'s$ are vectors of household and vehicle characteristics, and interactions thereof. This will include car features like car origin (Domestic, European, Japanese), vehicle type (full size sedan, compact, SUV, etc.), vehicle age, and horsepower/weight, as well as demographics like income, age, race, gender, family size, number of adults, local population density, and geographic region. Additionally, we will include an index of road congestion

(average road speed of nearby traffic monitoring stations weighted by the probability of being at a station that distance from the home census block group, where the probability is estimated from the distribution of trip lengths in the NHTS data). Also included will be ownership of a bicycle, self reported health, and self reported availability of public transit, which all factor into the use of substitutes for car travel.

Since VMT is in fact the Marshallian demand for vehicle use, we can use Roy's Identity to relate this demand to partial derivatives of an indirect utility function V_{ij}. Defining the price per mile driven in vehicle k as $p_{ik}^m = \frac{p_i^g}{mpg_k}$ and income after vehicle ownership decision as $y_i^F = y_i - \sum_{k \in J_i} p_k^s$, we then have

$$VMT_{ij} = \frac{\beta_{ij}}{\sum_{k \in J_i} \beta_{ik}} exp\left(\sum_{k \in J_i} (\alpha_{ik} - \beta_{ik} p_{ik}^m) + \lambda_i y_i^F\right) = \frac{-\partial V_{ij}(p_{ij}^m, y_i^F)/\partial p_{ij}^m}{\partial V_{ij}(p_{ij}^m, y_i^F)/\partial y_i^F}$$

Using the Implicit Function Theorem, we can equate the partial derivatives of the indirect utility function to $\frac{dy_i^F(p_{ij}^m)}{dp_{ij}^m}$, which gives us the following ODE:

$$\frac{dy_i^F(p_{ij}^m)}{dp_{ij}^m} = \frac{\beta_{ij}}{\sum_{k \in J_i} \beta_{ik}} exp\left(\sum_{k \in J_i} (\alpha_{ik} - \beta_{ik} p_{ik}^m) + \lambda_i y_i^F(p_{ij}^m)\right)$$

Solving this ODE yields the following equation,

$$y_i = \frac{1}{\sum_{k \in J_i} \beta_{ik}} exp\left(\sum_{k \in J_i} (\alpha_{ik} - \beta_{ik} p_{ik}^m) + \lambda_i (y_i(p_{ij}^m) - \sum_{k \in J_i} p_k^s)\right)$$

$$V_i = -\frac{1}{\sum_{k \in J_i} \beta_{ik}} exp\left(\sum_{k \in J_i} (\alpha_{ik} - \beta_{ik} p_{ik}^m)\right) - \frac{exp\left(\lambda_i y_i^F\right)}{\lambda_i}$$

Substituting our original variables back in, we have

$$V_i = -\frac{1}{\sum_{k \in J_i} \beta_{ik}} exp\left(\sum_{k \in J_i} (\alpha_{ik} - \beta_{ik} \frac{p_i^g}{mpg_k})\right) - \frac{1}{\lambda_i} exp\left(\lambda_i (y_i - \sum_{k \in J_i} p_k^s)\right)$$

Finally, we can add fixed effects $\theta_k = \theta z_k^\theta$ for the vehicle purchase decision and idiosyncratic error terms $\epsilon_{ik} \sim N(0, \sigma)$. Note that the fixed effects drop drop out of the partial derivatives of V_i, so this new V_i is still consistent with our original VMT equation.

$$V_i = -\frac{1}{\sum_{k \in J_i} \beta_{ik}} exp\left(\sum_{k \in J_i} (\alpha_{ik} - \beta_{ik} \frac{p_i^g}{mpg_k})\right) - \frac{1}{\lambda_i} exp\left(\lambda_i (y_i - \sum_{k \in J_i} p_k^s)\right) + \sum_{k \in J_i} (\theta_k + \epsilon_{ik})$$

3.2. **Road Speed and Congestion.** Given our goal of estimating counterfactual vehicle use under different policy regimes, we want to explicitly model how vehicle use affects traffic congestion, and how this in turn may influence vehicle use. Our model of household choice allows us to predict levels of vehicle use, conditional

on a particular level of congestion. This will tell us how many miles households drive, and, using the given road speeds, we can also estimate the extent of vehicle use in terms of hours of use. The traffic congestion portion of our model must then take this data on VMT or hours of vehicle use and map it to predicted road speeds. Iterating this process, we hope to find an approximate fixed point where the counterfactual household choices consistent with the previous iteration's predicted road speeds predict near identical road speeds for the next iteration.

We proceed in a very simple manner, constructing a simple linear model predicting average vehicle speed in terms of hours of vehicle use, along with several other factors. Before we state this model, we should briefly mention the nature of the observations we're analyzing. In this model we use observations from traffic monitoring stations, which each count the number of cars traveling through a several mile segment of roadway and measure their speed. We begin with hourly observations of the number of cars passing and their average speed, and then aggregate to monthly averages, so each observation is identified by a month-station pair. Then, for traffic monitoring station i during month m, we use the following equation:

$$S_{im} = \alpha_0 + \alpha_1 d_{im} + ... + \alpha_k d_{im}^k + \beta_1 S_{im}^{lim} + \beta_2 z_{im} + \epsilon_{im}$$

where S_{im} is the harmonic mean[2] of vehicle speed at station i in month m, d_{im} is vehicle density, the number of vehicles per lane per mile, S_{im}^{lim} is the speed limit for the segment of road corresponding to station i, ϵ_i is a normally distributed error term, and z_{ik} is a vector of additional covariates, such as latitude interacted with season (spring, summer, etc.), road types such as major arterial, minor arterial, and major collector (the preceding categories ordered from major to minor roadways), population density, number of lanes interacted with density, and interactions between speed limit and density. We'll look at several specifications for z_{ik} comprising various subsets of the aforementioned variables. Theoretically, we expect average speeds should approach the speed limit (or perhaps a bit higher) as traffic density goes to zero ($\alpha_k \approx 1$), and higher levels of density should result in average speeds decreasing from this ideal.

While the model itself is extremely simple, we should take some time to discuss two major assumptions that underlie it. One is the use of the harmonic rather than the arithmetic mean speed. To explain this decision, we need to return to our overall goal for the congestion model. We want our congestion model to take the level of vehicle use as its argument and to return the average speed of vehicle travel, so that we can predict how households will change their driving and vehicle purchase decisions in response to the new ease or difficulty of car travel. Households will re-optimize based on their preferences regarding vehicle speed, so the critical issue is which average is relevant to them. We assume that the primary cost of congestion is the

[2]Recall that, for a set of observations $X = \{x_1, ..., x_n\}$, the harmonic mean is $\mu^h(X) = n\left(\sum_{i=1}^n \frac{1}{x_i}\right)^{-1}$.

time cost of driving, rather than the change in fuel costs. Given this assumption, agents don't have direct preferences over vehicle speed, they only care about speed to the extent it influences drive times. If that's the case, then the arithmetic mean of speed is actually not the average agents care about.

An example will help to illustrate this: suppose an individual is driving 140 miles to some destination. This segment of roadway is monitored by two traffic stations, one that watches the first 70 miles and one that watches the second 70 miles. If they travel at 10 miles per hour for the first 70 miles, and 70 miles an hour for the last 70 miles, it will take eight hours, and station one will report a vehicle speed of 10 mph while station two reports 70 mph. If we simply take the arithmetic mean of speeds from these two stations, we get 40 mph. However, we could imagine an alternative trip where the actual speed is 40 mph on both legs. This trip would only take 3.5 hours. Clearly, the time cost of the first trip is far greater than the second trip, but if we use the arithmetic mean they both yield the same average speed. Thus, this arithmetic mean doesn't actually allow us to recover the corresponding time cost of the trip, which is what agents actually have preferences over. On the other hand, if we use the harmonic mean speed, the first trip yields an average speed of $\frac{2}{1/70+1/10} = 17.5$ mph. Dividing the trip distance by this speed, we can recover the true time cost of eight hours. Thus, the harmonic mean is the correct mean for averaging speeds over several legs of a trip to find the average speed traveled over the entire course of the trip, and the corresponding trip time associated with that speed. This is because the harmonic mean takes the average over the inverse of speed. Note, given a fixed distance of travel, speed and time of travel have a precisely inverse relationship: $distance = speed * time$, so for some given distance d $time = d/speed$. Thus, the harmonic mean averages a scalar multiple of travel time. The same logic applies if we imagine individuals facing 10 mph traffic on half of the days of the year and 70 mph travel on the other half–the time cost on the slow days relative to a speed of 40 mph corresponds to dramatically more time cost than the 70 mph days make up for, something that is not captured in the arithmetic mean.

Thus, when we have data where road speeds are sometimes very slow and sometimes very fast, the mean we use should capture the extreme cost of those very slow periods, as the harmonic mean does. Note that this model doesn't account for strategic behavior by drivers. A more complex model might take into consideration that drivers can change (some of) their trip times to take advantage of low traffic periods. This would be useful and interesting, but we do not pursue such a model here.

The other major assumption we make in this model is that density is the relevant independent variable. In fact, we don't directly observe density in the data. What we observe is flow, the count of cars that pass the monitoring station, or cars per lane per hour. The most obvious way to write this model, then, would be to make speed a function of flow. However, we believe there are several reasons why it is better to use density, which we define as flow/speed or cars per lane per mile. Before we detail them, however, let's establish a

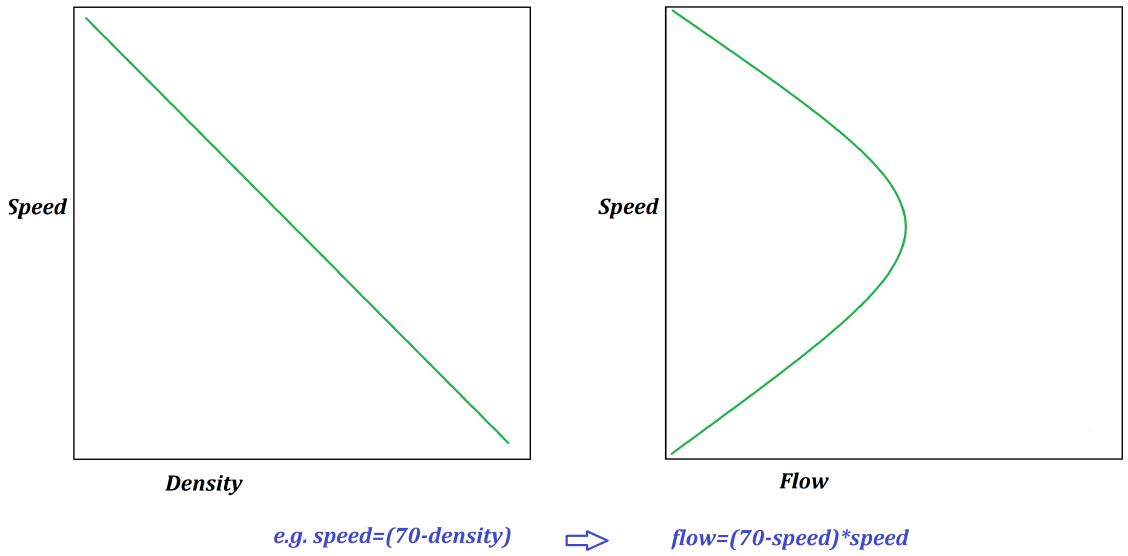

FIGURE 3.1. Macroscopic fundamental diagrams: the stylized relationship between speed, flow, and density in a simple road network.

simple model for thinking about traffic. In principle, traffic flow along road networks can be quite complex and difficult to visualize. However, we can get a great deal of intuition from one of the simplest kinds of road network, a circular race track. This track has a length, say one mile, and n cars are driving on it. For simplicity, let's say there is room for cars to pass, but only one lane. These cars drive at an average speed of v miles per hour. The traffic monitoring station is located at the finish line. Whenever cars complete a lap, they are counted towards flow for the hour of measurement (and their speed is measured as well). Then the average speed recorded is v and the density is n. We can then see that the flow, or count of cars passing the finish line per hour, must be nv. With this framework in mind, we can discuss the issues with using flow to predict speed. First, there is a simultaneity issue–while speed is indeed a function of flow, flow is also a function of speed. On the other hand, density influences speed by forcing cars to navigate around one another and stop at bottlenecks, but speed doesn't directly determine density. Thus, while density must be computed by dividing flow by speed–that is, dividing the flow by the (disaggregate elements of the) dependent variable itself, it actually removes the linear dependence of flow on speed. The other issue with using flow is that, theoretically, speed isn't a well defined function of flow, but is a well defined function of density. As mentioned before, we would predict that speed will decrease monotonically with density. This is a standard assumption in the traffic modeling literature (which is mostly outside of economics), and they call the relationships between flow, density, and speed that generally hold in a simple road network like the one described above macroscopic fundamental diagrams (Geroliminis and Daganzo (2009)), two of which are shown in Figure 3.1 .

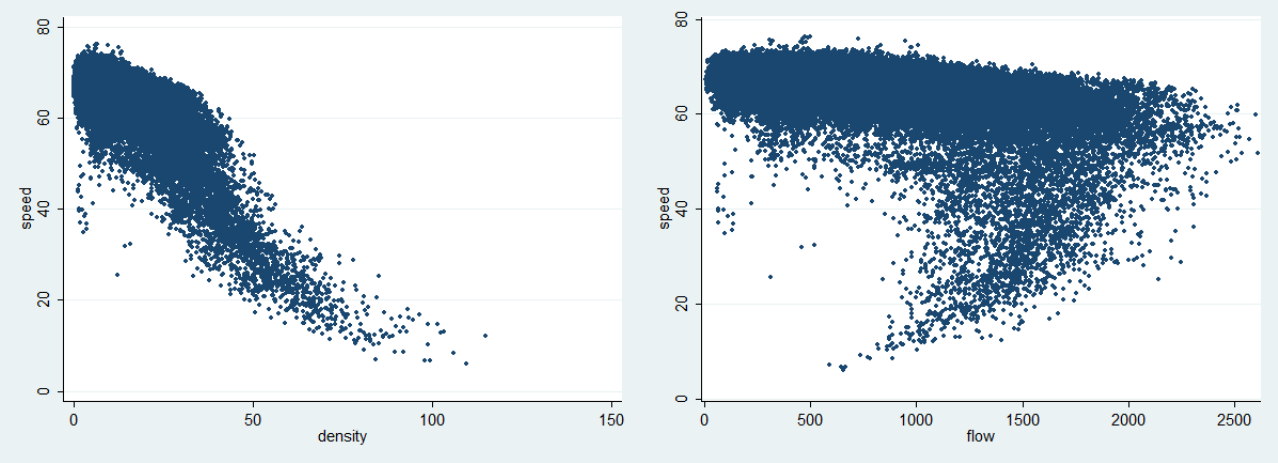

FIGURE 3.2. Scatter plots of a) speed and density and b) speed and flow. Hourly data from 4-lane highways in California District 3.

The right-hand diagram shows a backward bending relationship between flow and speed. Intuitively, if the density of cars is extremely low, speed will be high, but flow, but product of speed and density, will be low. If there is a moderate level of density, speed will still be fairly high, and flow will be maximized. However, if there is a high level of density, traffic will grind to a halt, and flow will again be low. In fact, we can observe this relationship in the data, as seen in Figure 3.2.

Thus, for any given level of flow, there will typically be a high and a low speed consistent with that flow. Trying to predict average speed in terms of flow will then give some convex combination of the high and low speeds, which won't correspond to either "correct" speed. This problem could be avoided by using quantile based estimation instead, but even then we'd have problems with simulating counterfactuals. In Figure 3.2 b), we can see that there are many more observations in the top half of the distribution than in the bottom half. We'd then predict that a flow of, say, 1000, is likely to correspond to a high speed like 70 mph, with a very small chance of corresponding to a low speed like 17 mph. However, we intend to run counterfactuals where gas taxes change significantly, and if gas taxes are greatly reduced we would likely see observations shifting from the densely populated top of the distribution to the sparsely populated bottom, and the distribution of speeds consistent with a flow of 1000 would now be heavily weighted towards extremely low speeds. This would not be reflected in our analysis, however, which would remain unchanged. By contrast, the relationship between density and flow is clearly monotonic and very smooth, so counterfactual increases in vehicle use should correctly map to decreases in speed when predicting speed by density. The fact that the support of the distribution in 3.1 a) would shift down and to the right would not compromise the predictions of the model.

4. Data Description

4.1. **Vehicle Choice and Vehicle Miles Traveled.** This analysis combines a variety of data sources. For the household's vehicle decision, the primary data source is the 2009 National Household Travel Survey (NHTS). The NHTS is a national survey of travel habits conducted by the US Federal Highway Administration (FHWA). The NHTS is conducted at five to ten year intervals–previous surveys were conducted in 1995 and 2001, and another is being conducted for 2015. Unfortunately, it is a cross-sectional survey with no longitudinal component–no effort is made to maintain individuals across survey years.

This survey collects a great deal of information about respondents. In particular, the questions relevant to our project include household characteristics such as household income and the demographic characteristics of the household's census block group and tract. The data includes personal characteristics such as age, sex, race, work status, driving status, health, travel disability, frequency of use of public transit, talking, and biking, and subjective views on the availability of public transit. Respondents also supply the make, model and year of their vehicles, how long they've been owned, and the annual miles driven for each vehicle. Respondents also are assigned a day to record all their travels, including the time and distance of each leg of each trip they take, the mode of transportation, and the purpose of the trip. While this provides a very brief look into the travel habits of any one person, the data taken as a whole provides a great deal of information on the nature of day to day travel behavior.

The 2009 NHTS survey was conducted from March 2008 through May 2009. Daily travel logs were collected for every day of the year. The NHTS survey covers the civilian, noninstitutionalized population of the US. These were collected from households with landline telephones. The initial stage of the NHTS was conducted by computer assisted telephone survey, preceded by an advance letter with a five dollar cash incentive where mailing was possible. FHWA mailed letters to over 400,000 households. During this stage, respondents answered all survey questions except the travel diary section and were assigned a "travel day" to record all travel on. Over 300,000 household members completed this portion of the interview. Several days later, respondents received a travel diary, and 72% of household members who completed the first stage also filled out the travel diary. In total, 196,619 households were recruited for the NHTS and the final data set includes 150,147 useable households, where at least 50% of adults completed the survey. This sample is composed of of a national sample containing 26,000 households, state level add-on samples ordered by state departments of transportation totaling roughly 107,500 households, and local add-ons ordered by several municipalities totaling about 14,500 households.

Table 1: Summary Statistics for NHTS Sample

Variable	Obs	Mean	Std. Err.	Min	Max
	Household Variables				
Population/sq. mi.	150,145	3,614	5,140	50	30000
Heavy rail in MSA	150,145	0.17	0.38	0	1
Number of respondents	150,145	2.06	1.06	1	13
100% response	150,145	0.87	0.34	0	1
Number of HH vehicles	150,145	2.05	1.16	0	27
Number of Adults in HH	150,145	1.89	0.69	1	10
	Individual Variables				
Medical difficulty with travel	270,221	0.12	0.33	0	1
Number of bike trips in the last week	308,366	0.38	1.73	0	99
Number of walking trips in the last week	305,816	4.31	6.79	0	99
Age	308,899	49.43	22.28	5	92
Male	308,899	0.46	0.50	0	1
Time to school (min)	21,925	15.47	11.49	1	95
Time to work (min)	116,096	23.74	21.07	1	660
Used Public Transit on travel diary day	262,868	0.02	0.14	0	1
Full time job	308,617	0.34	0.48	0	1
Part time job	308,617	0.10	0.30	0	1
Multiple jobs	308,617	0.00	0.06	0	1
Number of workers in HH	308,899	1.17	0.96	0	6
Individual yearly VMT	197,552	12,233	12,656	0	200000
	Vehicle Variables				
Vehicle yearly VMT	301,431	10,803	9,951	0	200000
Vehicle lifetime VMT	233,049	79,600	66,029	-7	999999
Months of ownership of vehicle	292,271	63.25	59.66	0.0333333	720

Notes: Final sample not yet determined.

Table 1 shows summary statistics for the NHTS sample, omitting categorical variables. A confidential supplement to the data set al.so identifies households up to census block group, generally allowing us to observe the location of households to within a few miles or even closer.

In addition to the NHTS data, we will need data on gasoline prices, vehicle characteristics, and used vehicle prices for 2009. For gas data, we'll use The American Chamber of Commerce Researchers Association's (ACCRA) quarterly gasoline prices by MSA. For 2009 vehicle prices, we plan to use the National Automobile Dealers Association's (NADA) used car prices. While current prices are easily obtainable, we are currently

trying to find out whether we can obtain a data file of 2009 prices or if we'll need to buy copies of the 2009 regional blue books and enter prices manually. Finally, for detailed data we could use Ward's Automotive Yearbook. However, this data source is somewhat expensive to use, so we're looking into less expensive vehicle datasets first.

4.2. **Road Speed and Congestion.** In addition to data on household behavior, we also employ road use data from state and federal departments of transportation. Specifically, we utiliz rich road use datasets from CalTrans' Performance Measurement System (PEMS) and the NYDOT. Throughout their respective states, the New York and California DOT's maintain traffic monitoring stations. These stations each observe a section of roadway and count every vehicle that passes. They also record the speed of each vehicle, as well as the lane it is in. We use hourly data on average speed and vehicle counts, along with data on the station, such as its geographic location, the type of road, the number of lanes, and whether the area is rural or urban. Because we have well over 100 million station-time observations, in all further analysis we aggregate this data by month so that stations can have at most 12 observations per year.

While this data is quite detailed, there are some caveats. In 2009, most Caltrans detectors were in metropolitan areas, and were primarily in use on interstates and other major roads. They do not provide data on all segments of all roads. In our dataset, there were 5,842 detectors at the beginning of the year and 6,711 by December. In December, 433 were in rural areas and only 637 had a speed limit less than 60 miles per hour. These stations were all permanent, so, except for those that were constructed in 2009, all stations provided data year-round. In total, we have 73,946 station-month observations in California. In New York, there were fewer stations, and many were temporary stations that only measured traffic for a few days of the year. In total, there were 2,812 stations yielding 4,834 station-month observations. Here, there were more observations of minor roads, with 4,061 station-month observations at roads below 60 miles an hour and 2,105 station-month observations of rural roads. Table 1 provides a summary of this data.

A few notes on how these variables were computed: given hourly data on vehicle counts and average speeds, we collapse the observations to monthly averages. For speeds, we weight observations by flow (vehicle counts). The logic is that we're interested in the effect of traffic on driving behavior, so if more cars are using a road at a particular time, more of them will experience that level of road speed/traffic, and this level of traffic will have a greater effect on the average speed drivers travel at than the road speed when few people are driving. Additionally, we use the harmonic mean of speed rather than the arithmetic mean.

Finally, because much of the station data fails to include a posted speed limit, we utilize the first percentile of the distribution of hourly average speeds as an alternative measure of maximum speed. Our goal is to measure traffic congestion in terms of the reduction in speed relative to an idealized situation with no other

Table 2: Road Use Summary Statistics

Variable	Obs	Mean	Std. Dev.	Min	Max
Average speed	78780	59.09	7.83	13.08	84.99
Harmonic mean of speed	78780	56.15	10.54	5.00	84.91
1st percentile of speed	78780	67.97	6.06	30.03	91.80
Posted speed limit	76611	68.91	4.42	30	70
Flow*	78780	747.11	336.83	3.92	7083.54
Density**	78780	12.87	6.43	0.11	153.23
Number of lanes	78780	2.93	1.23	1	4

Note: data from NY and CA only

*Average number of vehicles per lane passing the station in one hour

**Flow/speed--the average number of cars occupying one mile of one lane

cars on the road. In order to do that, we need to know what that maximum speed would actually be. The speed limit is the obvious choice here, but in lieu of data on posted limits we might simply look at the fastest speeds that are actually observed on the road. Since most roads are largely deserted in the very early morning, we have observations of traffic free vehicle use, and by looking at the top one percent of speeds we'll only be taking those high speed observations into account. An obvious objection here is that the first percentile might just be capturing extreme speeding by outliers. Note, however, that observations here are not individual cars but hourly averages of car speeds, so the issue of outliers should be mitigated somewhat. In fact, the first percentile of speeds seems to agree fairly well with posted speed limits in the data, except in the case of roads posted at 60 miles an hour, where the first percentile is closer to 70. This may actually be an advantage over using posted speed limits, since we're interested in how individuals actually drive, and speeding is very common.

In addition to the data described above, the Federal Highway Administration's Highway Performance Monitoring System (HPMS) provides national data on the geographic location, functional characteristics, and level of use of a subset of US roads, though not average speeds. Utilizing the function relating road use to average car speed derived using the NY and CA data, we could use this national road use data to predict counterfactual road use and corresponding traffic congestion at the national level. Unfortunately, this data is only available for 2011. Preliminary analysis has shown that the speed vs. car density regression remains largely the same with 2009 and 2011 data, so we may use this data in lieu of national 2009 data.

5. Estimation

5.1. **Household Travel Choices Model.** For this section of the model, we use the estimation method of Spiller (2012), a maximum score estimation technique based on Fox (2007). While there are two components to our household model, VMT and vehicle bundle choice based on indirect utility, the derivation of the indirect utility function from the VMT equation ensures that these two equations are consistent with one another, so only one needs to be estimated. Thus, we will only need to estimate the indirect utility functions. This simultaneous procedure has a major advantage over a two-step estimation technique where extensive and intensive margins are treated separately–it yields a single set of parameters, whereas two-step models yield two distinct values for each parameter that appears in both equations, making it difficult to interpret the estimation results.

The basic assumption for this estimation procedure is standard–households choose their optimal bundle. Define $V_{iJ} \equiv V_i(J)$ and $V_{J_i} \equiv V_i(J_i)$. Given this assumption, for any household i and observed bundle J_i, it must be that $V_{iJ_i} \geq V_{iJ}\ \forall J \neq J_i$. Then we can draw a sample of observed bundle/alternative bundle pairs for each agent and use maximum likelihood to recover the coefficients that best explain the set of observed choices. We use this method rather than a more standard discrete choice estimation technique like multinomial logit because the number of potential car bundles is extremely large. Discrete choice estimation procedures that attempt to account for the entire choice set will grow intractably complex as the size of the choice set increases. To avoid this problem, we limit choices to a subset of binary decisions. While this subset of choices becomes extremely small relative to the true choice set as the size of the choice set increases, Monte Carlo simulations (Fox (2007)) have shown that this technique can perform fairly well in recovering true parameters at a vastly smaller computational cost. Also following Spiller (2012), we limit these counterfactual J bundles to one vehicle swaps from the original bundle. That is, for any observed bundle J_i, we'll switch only one vehicle in the bundle for another from the set of possible vehicle choices. This will allow us to difference out the vehicle fixed effects in the indirect utility function, which, given the large set of available vehicles, will likely be in the hundreds or thousands of additional parameters. Specifically, we can choose pairs of households with different bundles J_1 and J_2. We then choose cars $j_1 \in J_1$ and $j_2 \in J_2$ such that $j_1 \neq j_2$ and define new bundles $J_1' = J_1/j_1 \cup j_2$ and $J_2' = J_2/j_2 \cup j_1$. Then, defining the non-fixed effects portion of an indirect utility function as $\widetilde{V}_{iJ} \equiv V_{iJ} - \sum_{k \in J} \theta_k$, for household 1 we have that

$$\widetilde{V}_{J_1} + \sum_{k \in J_2} \theta_k \geq \widetilde{V}_{J_1'} + \sum_{k \in J_2} \theta_k$$

Canceling out the unchanged vehicle fixed effects, we have

$$\widetilde{V}_{J_1} + \theta_{j_1} \geq \widetilde{V}_{J_1'} + \theta_{j_2}$$

For household two we similarly have

$$\widetilde{V}_{J_2} + \theta_{j_2} \geq \widetilde{V}_{J_2'} + \theta_{j_1}$$

Then, differencing the two inequalities, we have

$$\widetilde{V}_{J_1} + \widetilde{V}_{J_2} - \widetilde{V}_{J_2'} - \widetilde{V}_{J_1'} \geq 0$$

With this equation, we can construct a likelihood function giving the probability that a swap lowers the aggregate utility of the two households. Define $\overline{V}_{iJ} \equiv \widetilde{V}_{iJ} - \sum_{k \in J} \theta_k$. Since the errors are distributed normally, the sum of the four error terms, $\epsilon_{12} \equiv \epsilon_{2j_1} - \epsilon_{1j_1} + \epsilon_{1j_2} - \epsilon_{2j_2}$, will be distributed $N(0, 4\sigma^2)$ and the likelihood will be

$$Pr(\widetilde{V}_{J_1} + \widetilde{V}_{J_2} - \widetilde{V}_{J_2'} - \widetilde{V}_{J_1'} \geq 0) = Pr(\overline{V}_{J_1} + \overline{V}_{J_2} - \overline{V}_{J_2'} - \overline{V}_{J_1'} \geq \epsilon_{12})$$

$$= \Phi\left(\frac{\overline{V}_{J_1} + \overline{V}_{J_2} - \overline{V}_{J_2'} - \overline{V}_{J_1'}}{2\sigma}\right)$$

Where $\overline{V}_{iJ}$ is the deterministic portion of $\widetilde{V}_{iJ}$. Then, given a sample S of original and swapped bundle household pairs, the log likelihood function is

$$\mathcal{L} = \sum_S ln\left(\Phi\left(\frac{\overline{V}_{J_1} + \overline{V}_{J_2} - \overline{V}_{J_2'} - \overline{V}_{J_1'}}{2\sigma}\right)\right)$$

From here, one can simply maximize the likelihood function using a preferred numerical technique. Also following Spiller (2012), we put a particular limitation on how samples are drawn. Specifically, if two samples contain the same household i and that household swaps the same vehicle j in both cases, those two observations will have correlated error terms, because the error term ϵ_{ij} will appear in both inequalities. To avoid this, vehicle-household pairs must be chosen without replacement when constructing the sample.

Once the non-fixed effect portion of the indirect utility function is estimated, we can also estimate vehicle use, as the VMT equation is derived directly from the indirect utility function, and takes the same parameters. Also, we easily recover the vehicle fixed effects in a second estimation if desired, given the already estimated $\widetilde{V}'s$.

5.2. **Road Congestion Model.** Given our road congestion equation

$$S_{im} = \alpha_0 + \alpha_1 d_{im} + ... + \alpha_k d_{im}^k + \beta_1 S_{im}^{lim} + \beta_2 z_{im} + \epsilon_{im}$$

, we simply run an ordinary least squares regression for each specification of z_{im} and k.

5.3. **Finding Equilibrium Traffic and Vehicle Choices.** We now have models for road congestion and household choice, but our overall goal is to be able to answer questions about how traffic congestion influences household behavior and vice versa in counterfactual scenarios, such as when gas taxes are changed. To do this we can plug the speeds derived from the road congestion model under observed conditions into the household choice model to recover new household choices. These choices will include the intensive margin of vehicle ownership–VMT, which corresponds to traffic flow. Recall the example of a race track in Section 3.2. There, flow is the number times cars pass the finish line per hour. VMT is the number of miles driven per year, so VMT is in fact measured in units of flow. Further, we can convert VMT into hours of car travel per year by dividing VMT by the average speed. This then gives us units of vehicle density. Thus, we can sum the density's resulting from each household's decisions to predict a new level of overall vehicle density, which will in turn predict a new regime of household behavior, which will correspond to another level of density, and so on.

By iterating this procedure, we should approach a fixed point where the level of congestion is consistent with household behavior. We do not attempt to formally prove that this function is a contraction mapping or that such a fixed point exists, but we are fairly confident that we will see convergence. One potential obstacle to this is the discrete nature of some of the household's choices. Potentially, this procedure could get caught in a loop where agents jump from one discrete choice to another and back again forever. Should this become an issue, we may simplify the household's choice set to shrink the set of discontinuities where this problem could occur. In practice, we will likely not actually naively iterate through repeated inputs of the output of the previous input, but will instead use a procedure that uses information from previous iterations to predict the fixed point, and then guess that value.

Up to now, we've been vague about how speeds determined at the road segment level relate enter into the household's problem and how household level vehicle use estimates enter into a particular road's congestion estimation. We'll finally make that explicit. First, we can use the individual level NHTS data on the length of vehicle trips to estimate a distribution of trip lengths conditional on population (trips will typically be longer in more rural areas), VMT, and other factors. An unconditional distribution is shown in Figure 5.1. Given this distribution and the geographic location of the household (proxied by the centroid of the census block group) and the location of nearby traffic stations, we can assign weights based on the PDF of the distribution to each station conditional on its distance from the household. These weights represent the relative frequency of travel through that road segment by the given house, and we can easily normalize these weights to sum to one. For computational ease, we will drop all stations beyond a certain distance if there

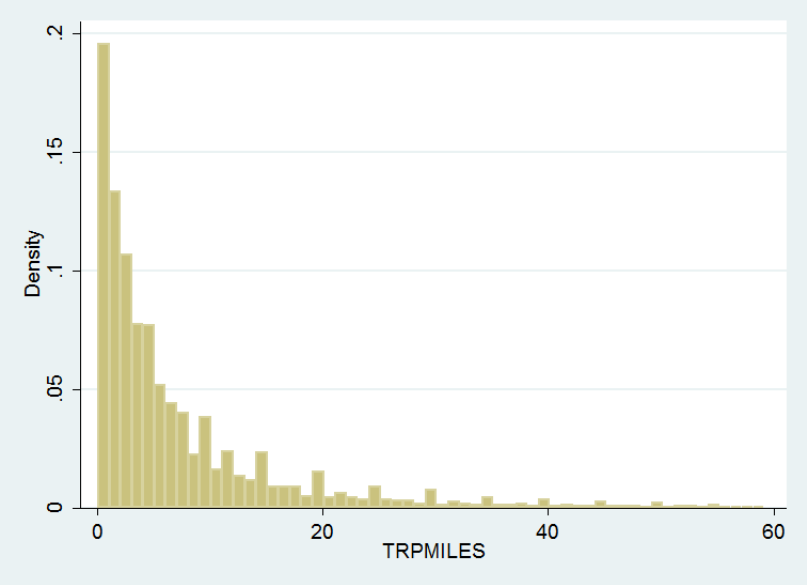

FIGURE 5.1. Histogram of the distances of trip legs in miles for trips using private motor vehicles.

are sufficiently many stations within that distance. Since these more distant stations will have very low weights to begin with, dropping them should make little difference. Formally, let T_k be the set of stations retained for household k. Let the weight for household k and station i be p_{ki} such that $\sum_{i \in T_k} p_{ki} = 1$. Then the measure of mean road speed used in the discrete-continuous choice problem for household k is $s_k = \frac{n}{\sum_{i \in T_k} \frac{p_{ki}}{s_i}}$, where s_i is the mean speed for station i, as estimated by the congestion regression.

Going from congestion to households, we need to aggregated speeds over roads, but going from households to congestion we need to do the opposite: to aggregate speeds over households. Analogous to our previous procedure, we'll drop households that are too distant to save computation time. Then let H_i be the set of households retained for station i. Define d_k as the total hours of car travel estimated for household k divided by number of hours in the year–this is the average contribution of the car to road density. A simple course of action would be to estimate density on a given road segment as the weighted average of the household densities: $d_i = \sum_{k \in H_i} p_{ki} d_k$. However, because we don't observe all households or all sections of roadway, this estimate will be a scalar multiple of the desired roadway density. Aside from that, due to the network structure of traffic flow which we do not fully observe, the actual densities we observe in the data will likely not match up exactly with the predicted densities, even up to scalar multiplication. Thus, for each station we will include a station specific scalar term that will account for both issues: $d_i = \alpha_i \sum_{k \in H_i} p_{ki} d_k$. Using the observed d_i, and computing the $d'_k s$ and $p'_{ki} s$ corresponding to the observed data, we can recover the $\alpha'_i s$. Then for counterfactual household behavior, we can compute the corresponding roadway densities by $\hat{d}_i = \alpha_i \sum_{k \in H_i} \hat{p}_{ki} \hat{d}_k$.

6. Results

6.1. **Household Travel Choices Model.** Not yet complete.

6.2. **Road Congestion Model.** The results of the road congestion model are generally intuitive, and because of the large sample size most coefficients are extremely statistically significant. We consider four specifications. The first and simplest specification estimates the harmonic mean of speed as a function of density, maximum speed[3] max speed. The fact that the coefficient is slightly less than one probably indicates that the 1st percentile of average speed is picking up some drivers prone to speeding. We see a small and statistically insignificant effect for the urban dummy. Finally, we see that people drive faster in summer and slower in winter, which makes sense given that there are generally more weather related driving hazards in winter.

In specification (2), we had higher order terms for density as well as interaction between density and max speed. Here, we see a positive coefficient for density of about 0.5, which is the opposite of what we'd expect. Note, however, that the second order term is negative, as is the interaction with max speed. Given that no road has a max speed below 30, and the overwhelming majority have maximum speeds closer to 60 or 70, the linear coefficient on density net of the max speed interaction will be much smaller for most observations. Thus, for almost all density observations, the higher order density terms will dominate and yield a negative effect for density. The small positive effect seen in the small set of observations with very low density and max speed are probably indicative of an area of the parameter space where the model doesn't fit well because it is so sparsely observed. The negative coefficient on the density and max speed interaction also seems reasonable, as the difference between the speed of gridlock traffic (a few miles an hour) and the road's maximum speed increases in rough proportion to maximum speed, so that the increase in density must, on average, have a larger negative effect on speed on faster roads. The urban dummy now has a larger and more statistically significant effect on road speeds. This is likely due to urbanicity proxying for the number of intersections a road passes through, which slow down traffic, especially when there are many cars on the road.

In specification (3), we add dummies for road type as well as interactions between density and road type, density and number of lanes, and season and latitude. The coefficient results are qualitatively unchanged for the variables included in specification (1). We can see that, relative to the baseline of major highways, minor arterial (medium traffic flow) and major collector (low traffic flow) roads have slower average road speeds relative to the maximum, with major collectors being the slowest. This is likely due to the fact that minor roads are more likely to have intersections than major highways.

[3]We use the maximum speed here rather than posted speed limits because many stations do not have associated speed limit records, and would have to be dropped if speed limit were used.

Table 3: Estimated speed as a function of road density and other factors

	Specification			
	(1)	(2)	(3)	(4)
Density	-0.785	0.491	-0.888	0.439
	(0.005)**	(0.028)**	(0.012)**	(0.045)**
Density^2		-0.055		-0.044
		(0.002)**		(0.002)**
Density*max speed		-0.008		-0.011
		(0.000)**		(0.001)**
Max speed	0.844	0.932	0.825	0.912
	(0.004)**	(0.005)**	(0.006)**	(0.008)**
Urban	-0.068	-0.331	-0.921	-0.683
	(0.123)	(0.119)**	(0.127)**	(0.126)**
Spring	0.47	0.634	3.171	5.243
	(0.084)**	(0.079)**	(0.835)**	(0.818)**
Summer	1.213	1.337	3.235	5.942
	(0.082)**	(0.077)**	(0.783)**	(0.768)**
Winter	-0.905	-0.924	-4.817	-1.951
	(0.084)**	(0.079)**	(0.920)**	(0.902)*
Spring*latitude			-0.071	-0.127
			(0.023)**	(0.023)**
Summer*latitude			-0.051	-0.127
			(0.022)*	(0.021)**
Winter*latitude			0.112	0.031
			(0.026)**	(0.025)
Major collector road			-7.702	-1.988
			(0.275)**	(0.310)**
Minor arterial road			-3.846	-1.033
			(0.255)**	(0.266)**
Road type and density interactions			Yes	Yes
Higher powers of density (3^{rd} to 6^{th})		Yes		Yes
Number of lanes dummies	Yes	Yes	Yes	Yes
Lanes*density			Yes	Yes
Constant	5.803	-1.668	9.034	0.962
	(0.291)**	(0.308)**	(0.455)**	-0.555
R^2	0.47	0.54	0.52	0.54
N	79,871	79,871	76,641	76,641

* p <0.05; ** p <0.01

ave road type observations, and so were dropped for specifications (3) and (4).

Finally, in specification (4), we add the addition covariates from specifications (2) and (3). We see results very similar to those of specification (2) for the variables included in specification (2), and very similar to specification (3) for those variables included in (3). Generally, we see a great deal of consistency in the signs of most coefficients across the four specifications, and significant consistency in the magnitudes of quite a few coefficients, except of course for those that have interaction terms in one specification and not another.

7. Conclusions

Not yet complete.

References

[1] Hausman, J. A. (1981). Exact consumer's surplus and deadweight loss. The American Economic Review, 662-676.

[2] U.S. Department of Transportation, Federal Highway Administration, 2009 National Household Travel Survey. URL: http://nhts.ornl.gov.

[3] Geroliminis, Nikolas; & Daganzo, Carlos F. (2009). Existence of urban-scale macroscopic fundamental diagrams: Some experimental findings. University of California Transportation Center. UC Berkeley: University of California Transportation Center. Retrieved from: http://escholarship.org/uc/item/1b07t4q7

[4] Spiller, Elisheba, Household Vehicle Bundle Choice and Gasoline Demand: A Discrete-Continuous Approach (August 14, 2012). Available at SSRN: http://ssrn.com/abstract=2129396 or http://dx.doi.org/10.2139/ssrn.2129396

[5] Manski, Charles F. "Maximum score estimation of the stochastic utility model of choice." Journal of Econometrics 3, no. 3 (1975): 205-228.

[6] Feng, Ye, Don Fullerton, and Li Gan. Vehicle choices, miles driven, and pollution policies. No. w11553. National Bureau of Economic Research, 2005.

[7] Fox, Jeremy T. "Semiparametric estimation of multinomial discrete-choice models using a subset of choices." The RAND Journal of Economics 38, no. 4 (2007): 1002-1019.

[8] Dubin, Jeffrey A., and Daniel L. McFadden. "An econometric analysis of residential electric appliance holdings and consumption." Econometrica: Journal of the Econometric Society (1984): 345-362.

[9] Petrin, Amil. Quantifying the benefits of new products: The case of the minivan. No. w8227. National Bureau of Economic Research, 2001.

[10] West, Sarah E. "Distributional effects of alternative vehicle pollution control policies." Journal of public Economics 88, no. 3 (2004): 735-757.

[11] Parry, Ian WH, and Kenneth A. Small. "Does Britain or the United States have the right gasoline tax?." The American Economic Review 95, no. 4 (2005): 1276-1289.