NONPARAMETRIC REGRESSION ESTIMATION OF HOUSEHOLD VMT

Young-Jun Kweon
(Corresponding Author)
Graduate Student Researcher
Department of Civil Engineering
The University of Texas at Austin
ECJ 6.9, Austin, Texas 78712
Tel: (512) 471-8270
FAX: (512) 475-8744
Email: fire264@mail.utexas.edu

Kara M. Kockelman
Clare Boothe Luce Assistant Professor of Civil Engineering
The University of Texas at Austin
ECJ 6.9, Austin, Texas 78712
Email: kkockelm@mail.utexas.edu

Submitted for Presentation & Publication at the 2004 Annual Meeting of the Transportation Research Board

Word Count: 4230 words + 8 figures & tables = 6230 word-equivalents

ABSTRACT

The number of vehicle miles traveled (VMT) each year by households is a key variable of interest. It is used in most models of travel demand, as a control variable, a response variable, or both. This study employs nonparametric econometric techniques to examine the effects of household income, vehicle ownership and workers on annual household VMT using the 1995 Nationwide Personal Transportation Survey. The results are density functions and regression surfaces for VMT, in relation to these and other variables, including public transit availability, housing location (urban versus rural), and retirement.

As expected, households with more vehicles and workers exhibited much higher annual VMT. However, the effect of vehicle ownership tapers off, with VMT stabilizing at 5 or more vehicles. Interestingly, at low levels of household income, income increases are associated with lower VMT, essentially approximating a convex quadratic form, while controlling for vehicle ownership and workers.

These nonparametric regressions perform better than their ordinary-least squares counterparts in many ways: they permit full distributional flexibility of error terms, they result in higher R2 values, and they illuminate complex relationships that would not be contemplated by most analysts. However, they also require more computation, a large sample size in all data regions, and relative few control variables.

Key words: Nonparametric Regression; Vehicle Miles Traveled (VMT); Kernel Density Estimation; National Personal Travel Survey (NPTS)

1. INTRODUCTION

Household vehicle miles traveled (VMT) is perhaps the strongest single indicator of automobile dependence and a household's travel patterns. It also serves as a strong proxy for gasoline consumption, vehicle emissions, and crashes, which are of substantial interest to nations, communities, and their policymakers. It has been used in for urban planning (e.g. Miller and Ibrahim, 1998), travel demand analysis (e.g. Barr, 2000), traffic crash analysis (e.g. Kweon and Kockelman, 2003), and energy consumption and pollution analysis (e.g. Lyons et al. 2003). Almost always, rather than being measured directly, reports of VMT are based on other variables, such as a day's worth of travel destinations (e.g., Kockelman, 1997). short-period odometer readings on each personal vehicle, sample counts of key roadways around a region, and/or regional or state gas tax receipts.

Using these estimates, and knowledge of several explanatory variables of interest, VMT predictions have been made using parametric regression (e.g. Hansen and Huang, 1997; Kockelman, 1997) and classification or contingency tables (e.g. Kumapley and Fricker, 1996). The later is a non-parametric technique, but typically offer little to no information as to the distribution of VMT within each class. This paper takes a more flexible and informative approach, based on nonparametric estimation of VMT and its distribution, even while controlling for key predictor variables.

There have been studies to make use of VMT for various purposes. Kockelman (1997) and Miller and Ibrahim (1998) estimated VMT as functions of urban form and neighborhood characteristics. Applications of VMT estimates run the gamut, and are often quite interesting. For example, Lave (1996) investigated the improbable 41 percent increase in NPTS VMT estimates, between 1990 and 1983 using VMT data from three distinct data sets: the Residential Transportation Energy Consumption Survey (RTECS), California Smog-Check Data, and Federal Highway Administration data. Lave concluded that the 1990 NPTS VMT estimates were biased high, due to over-sampling of high-income households.

Using panel data for California urban counties between 1973 and 1990, Hansen and Huang (1997) applied a distributed-lag fixed-effects model and estimated elasticities of VMT with respect to the highway lane-miles to be quite high (at 0.6 to 0.9). Barr's (2000) study looked at VMT elasticities with respect to travel speeds (and other explanatory variables) and concluded that travelers spend 30 to 50 percent of the time savings by highway capacity improvements for additional travel. Using 1995 NPTS data and national crash data sets (FARS and GES), Kweon and Kockelman (2003) examined the bias in crash-risk conclusions based only on crash severity or vehicle ownership statistics.

Mingo and Wolff (1995) examined trucking VMT estimates using the U.S. Truck Inventory and Use Survey (now called Vehicle Inventory and Use Survey) and made recommendations regarding estimation methods. Similarly, Weinblatt (1996) proposed that the FHWA apply seasonal and day-of-week adjustment factors for its truck VMT estimates.

While nonparametric methods of VMT estimation do not appear in the published literature, such methods typically have been used for traffic flow investigations. Nearest-neighbor nonparametric regressions appear in studies forecasting traffic flow or travel time using large traffic flow datasets. For example, Clark (2003) recently examined relationships between flow, occupancy, and speed in order to generate short-term predictions of traffic flow. He employed a *k*-nearest-neighbor regression and relied on high-quality loop detector data from England. Smith et al. (2002) and Smith and Oswald (2003) used nearest-neighbor techniques to

forecast traffic flow based real-time traffic data. And You and Kim (2000) also used this technique to forecast travel time using traffic flow data on highways in Korea.

Although nonparametric regression techniques in transportation have focused on traffic flow, Kharoufeh and Goulias (2002) estimated kernel densities of various activity durations and conducted comparisons (on the basis of gender and presence of children in the household, for example). All studies using nonparametric regressions were heavily focused on traffic flow research largely because abundant data are available. No study using nonparametric regressions has been found in the transportation planning side of research; Kharoufef and Goulias (2002) is the one that comes closest.

This study explores variations in household VMT responses to the 1995 NPTS survey in relation to several key household characteristics: income, vehicle ownership, workers, public transit availability, residential area (urban/rural), housing type, and retirement. By using nonparametric techniques, VMT distributions are made explicit and can be visually compared. Moreover, the relationship of VMT and various household characteristics is estimated in flexible ways that cannot be expressed by relatively restrictive, parametric forms.

2. METHODOLOGY

Compared to parametric regression techniques, nonparametric density and regression estimation methods are largely unknown and complex. Density estimation using a kernel function is described here first, along with details of nonparametric regression. Other issues accompanying these nonparametric methods, such as bandwidth selection and dimensionality, are addressed.

2.1 Kernel Density Estimation

Rosenblatt (1956) introduced the idea of kernel estimation by imposing weights on each observation in a data set. These weights vary with distance from the central value, the kernel. A kernel function $K_h(x - X_i)$ provides these weights and is centered on each observation, X_i. A general kernel function *K* with a bandwidth *h* is defined as $K_h(x) = h^{-1}K(x/h)$. Kernel estimation permits full functional flexibility along with mathematical tractability. A smoothing parameter, often called bandwidth, *h* regulates the degree of smoothness for kernel smoothers and should be determined appropriately (as discussed in section 2.3 of this paper).

There are various kernel functions one might use, and, interestingly, one's choice of such functions should not substantially affect the density/regression estimates and results in general (Hardle, 1990). For illustration, the Epanechnikov function is as follows:

$$K_h(x - X_i) = h^{-1}\frac{3}{4}\left(1 - \frac{(x - X_i)^2}{h}\right) I\left(\left|x - X_i\right| \le h\right) \quad (1)$$

where *h* denotes bandwidth, *x* is a specific value, X_i is observation *i*'s vaue, and $I(\cdot)$ is an indicator function (equaling one if the condition in the parentheses is true and one if false).

In order to meet density definitions, kernel functions are necessarily symmetric around 0 and integrate to 1, and their estimates represent densities – and do not depend on any choice of origin, *x* (Hardle, 1991).

Averaging over all observations' kernel functions leads to the sample data's kernel density estimator:

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right) \quad (2)$$

where *n* is the number of observations.

2.2 Nonparametric Regression

2.2.1 General Concept

Regression estimation aims to find a relationship between a dependent variable and a set of independent variables. Often, regression equations are specified as the following:

$$Y_i = m(X_i) + \varepsilon_i$$

where ε_i denotes a random term with mean zero and variance σ^2 and defines the variation of Y_i around its mean, $m(X_i)$. $m(x)$ can be expressed as:

$$m(x) = E[Y_i \mid X_i = x] = \frac{\int y \cdot f(x,y)dy}{\int f(x,y)dy} = \frac{\int y \cdot f(x,y)dy}{f(x)} \tag{3}$$

where $f(x,y)$ denotes the joint density of X_i and Y_i and $f(x)$ denotes the marginal density of X_i.

For nonparametric prediction of *Y* values, one may weigh observed *Y* values, Y_i, associated with X_i values that are in the neighborhood of *x*. The weight for each observation Y_i depends on the distance of X_i to *x*. The general form of a nonparametric regression estimator (also called smoother) can have the following form of:

$$\hat{m}_h(x) = n^{-1}\sum_{i=1}^{n} W_{hi}(x) \cdot Y_i \tag{4}$$

where $W_{hi}(\cdot)$ is a weight function depending on the bandwidth *h* and the sample's values of the explanatory variable *x*.

Most nonparametric regression techniques can be viewed as a weighted average of the response variable Y_i, where the weight function $W_{hi}(x)$ relies on the specific technique employed and on the bandwidth-scaled distance between *x* and X_i (Hardle, 1999).

2.2.2 Kernel Regression

Nadaraya (1964) and Watson (1964) developed the Nadaraya-Watson kernel regression estimator of $m(x)$, based on locally weighted averages of Eq. 3's numerator and denominator. The denominator can be replaced by the kernel density estimate, and the numerator's joint density $f(x,y)$ can be estimated using a multiplicative kernel:

$$\hat{f}_{h_1,h_2}(x,y) = n^{-1}\sum_{i=1}^{n} K_{h_1}(x - X_i)K_{h_2}(y - Y_i) \tag{5}$$

where $K_{h_1}(\cdot)$ is a kernel density with a bandwidth *h*.

Then Nadaraya-Watson kernel estimator can be expressed as

$$\hat{m}_h(x) = \frac{n^{-1} \cdot \sum_{i=1}^{n} Y_i \cdot K_h(X_i - x)}{n^{-1} \cdot \sum_{i=1}^{n} K_h(X_i - x)} = n^{-1} \cdot \sum_{i=1}^{n} W_{hi}(x) \cdot Y_i \tag{6}$$

where $W_{hi}(x) = \dfrac{n^{-1} \cdot K_h(X_i - x)}{\hat{f}_h(x)}$ are normalized weights for each Y_i value. (Hardle, 1991).

2.2.3 Local Polynomial Regression

Nonparametric local polynomial regression estimators can be viewed as the solution to the least squares problem, $\min_{\boldsymbol{\theta}} \sum_{i=1}^{n} [Y_i - m_i(x, \boldsymbol{\theta})]^2 K_h(X_i - x)$, for some parametric model $m_i(x, \boldsymbol{\theta})$. If $m_i(x, \boldsymbol{\theta}) = \alpha + \beta(X_i - x)$, the local linear estimator can be obtained as follows:

$$\hat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{[s_{h2}(x) - s_{h1}(x)(X_i - x)]}{s_{h2}(x) s_{h0}(x) - s_{h1}(x)^2} K_h(X_i - x) Y_i \qquad (7)$$

where $s_{hr}(x) = \left[\sum_{i=1}^{n} (X_i - x)^r K_h(X_i - x)\right] / n$ (Bowman and Azzalini, 1997).

The Nadaraya-Watson estimator can be thought of a local polynomial regression estimator with a local constant. In other words, the value of α that minimizes $\sum_{i=1}^{n} (Y_i - \alpha)^2 K_h(X_i - x)$ is $\alpha^* = m_i(x)$ (Pagan and Ullah, 1999). The Nadayara-Watson kernel estimator is unbiased only when the function being estimated is a simple constant, invariant with *x*. In such a case, the statistical properties of the local linear smoothing estimator suggest that a locally linear, nonparametric regression[1] would be superior to the Nadaraya-Watson regression (Blundell and Duncan, 2000).

2.3 Bandwidth Selection

If one knew the relation that existed between *y*'s and *x*'s, *m*(*x*), one could derive formulae for "optimal" bandwidths by minimizing the mean squared error (MSE) or the mean integrated squared error (MISE). However, the true relationship between *y* and *x* values is unknown, so several methods to determine an appropriate bandwidth value have developed. Four that were used for density estimation in this study are described here; they are the rule of thumb (ROT), unbiased cross-validation (UCV), biased cross-validation (BCV), and direct plug-in (DPI) methods. Manual bandwidth selection (where one tries several bandwidth values and chooses based on visual examination of the resulting observations and predictions) also may be used; however, it is time-consuming and rather subjective. Different methods for bandwidth selection can produce rather different values, so bandwidth selection remains a subjective process.

2.3.1 Rule of Thumb

The rule-of-thumb (ROT) method uses a reference distribution to choose the bandwidth, and the normal distribution (with parameters μ and σ) serves as a guide. The following estimate of bandwidth employs the Gaussian kernel:

$$\hat{h}_0 = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} \approx 1.06 \hat{\sigma} n^{-1/5} \qquad (8)$$

In order to make this estimate less sensitive to outliers, the interquartile range $\hat{R}$ may be used instead of $\hat{\sigma}$. The ROT becomes:

$$\hat{h}_0' = \left(\frac{4\hat{\sigma}^5}{3n}\right)^{1/5} \approx \frac{\hat{R}}{1.34} n^{-1/5} \qquad (9)$$

where $\hat{R} = X_{[0.75n]} - X_{[0.25n]}$ and $\hat{R} \approx 1.34\sigma^2$ for Gaussian data. Combining these two rules yield an even better ROT formula for bandwidth selection (Hardle, 1991):

$$\hat{h}_0 = 1.06 \times \min(\hat{\sigma}, \hat{R}/1.34) \times n^{-1/5} \quad (10)$$

2.3.2 Data-Driven Bandwidth Methods

Least-squares cross-validation and direct plug-in (DPI) are data-driven methods. The unbiased cross-validation (UCV), also known as least-squares cross-validation (LSCV), and the biased cross-validation (BCV) are described here, and these two along with the DPI approach are plotted, for purposes of comparison, in Figure 1.

Typically, the UCV bandwidth is less than that of BCV. The general CV formula can be found in Scott and Terrell (1987). Here, the formula using a Gaussian kernel is presented (Scott, 1992).

$$UCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{n^2 h\sqrt{\pi}} \sum_{i<j} \left(e^{-\Delta_{ij}^2/4} - \sqrt{8} e^{-\Delta_{ij}^2/2} \right) \text{ and}$$

$$BCV(h) = \frac{1}{2nh\sqrt{\pi}} + \frac{1}{64n^2 h\sqrt{\pi}} \cdot \sum_{i<j} \left(\Delta_{ij}^4 - 12\Delta_{ij}^2 + 12 \right) \cdot e^{-\Delta_{IJ}/4} \quad (11)$$

where n is the number of observations and $\Delta_{ij} = \frac{X_i - X_j}{h}$

2.4 Curse of Dimensionality

While a nonparametric approach seeks to "let the data speak for itself" and can be used with multiple independent/explanatory variables, it is subject to a "curse of dimensionality": as the number of variables increases, the number of observations must increase exponentially in order to perform a statistically meaningful analysis. In addition, in case of more than two dimensions (i.e., two explanatory variables), visualization becomes restricted (one must fix the other variable values to view marginal distributions). In spite of this "curse" nonparametric regressions with two or more independent variables are usually thought to be still useful.

3. DATA AND ESTIMATION

3.1 Data

There are two U.S. databases that offer VMT at the household level: the Nationwide Personal Transportation Survey (NPTS) and the Residential Transportation Energy Consumption Survey (RTECS). Both cover all 50 states and Washington D.C. The RTECS is a sub-sample of the Residential Energy Consumption Survey (RECS), maintained by U.S. Department of Energy's Energy Information Administration (EIA) (Harrison and Moorhead, 2000). The NPTS is the only comprehensive nationwide survey of personal travel in the U.S., and it has been administered by the Federal Highway Administration (FHWA) in collaboration with other agencies in the U.S. Department of Transportation (DOT)[2] (FHWA, 1997).

The NPTS data were chosen for this work because they have more commonly used for household VMT estimation, and they provide a larger sample size: 42,033 households, rather than 3,000. However, only 17,477 of the NPTS's surveyed household provided an estimate of their household's annualized VMT (based on odometer readings) for all their vehicles and complete income data. So only those observations are used here.

Table 1 describes the variables used for this study. The 17,477 sample households logged an average of 18,000 miles per year, using all the household's personal vehicles (excluding motorcycles, bicycles and mopeds). About 82 percent of sample households own (rather than rent) their home, and the average reported household income is around $46,000 (in 1995 dollars). Almost?? 60 percent of households reside in urban areas and 63 percent live in a community where public transit (including bus, subway, or street car) is available.

<TABLE 1>

The focal point of this study is the effect of income, workers, and vehicles on a household's annual VMT, assuming some other household characteristics (such as the retirement status of the household "head"[3]). In section 3.2, VMT density functions are estimated with and without pre-set levels of public transit availability, home location, dwelling type, retirement status of household head, number of workers, and vehicle ownership. And then, in section 2.3, VMT regressions are estimated with one, two regressors, and all seven regressors.

3.2 Density Estimation

VMT density estimation was conducted using a Gaussian kernel function and four different bandwidth selection rules (ROT, UCV, BCV, and DPI). Figure 1 shows the estimated density results, with a mode around 15,000 miles and rightward skew. Household annual VMTs in excess of 50,000 miles are highly unusual. The UCV bandwidth (1093.9 mi/year) results are very close to those for the DPI bandwidth (952.0) and reasonably close to the BCV-bandwidth (3042.1) results. All differ strikingly from the ROT results, however: the ROT bandwidth was based on Eq. 10 and was much higher, at 7145.9.

<FIGURE 1>

These results exhibit a boundary problem, by producing positive density estimates for negative VMT values. There are several methods to treat such problems, including logarithmic transformation of VMT, (though this did not produce satisfactory results here).

Figure 2 displays the effects of indicator variables on VMT density estimates. The dots encompassing the density estimates are standard errors of the corresponding estimates. Households living in communities with some form of public transit are found to generate less VMT (Figure 2a). Those living in urban areas also tend to drive less (Figure 2b); this is likely because distances between one's home and destinations tend to be shorter, due to the relatively high land use intensity of urban areas. Households with retired heads also drive less (Figure 2d), for obvious reasons. The mode values of VMT for households residing in different housing types (Figure 2c) are very close, with a slightly higher mode for those owning their dwellings.

< FIGURE 2>

Figure 3 reveals changes in VMT density functions with number of workers (Figure 3a) and vehicles (Figure 3b). Both variables have positive effects on VMT, consistent with expectations. As these factors increase, the range or variance in VMT also rises, rather dramatically.

< FIGURE 3>

Nonparametric methods were developed for, and thus are most appropriate, for continuous variables, with different densities applying for different levels of indicator variables. Special caution is necessary for estimation results based on multi-class discrete variables. While many regression results appear valuable in this work, density estimates for models of VMT as a function of income alone, and of workers and vehicles, performed poorly, perhaps because the data are discrete (even in the case of income, since the NPTS collects these in the form of 18 categories).

3.3 Nonparametric Regressions

3.3.1 Use of a Single Regressor

A nonparametric regression of VMT of household income using the Nadaraya-Watson (N-W), local linear (LP1) and quadratic (LP2) estimators, with a cross-validation bandwidth, resulting in Figure 4a. As expected, income has a positive effect on VMT, though this effect gradually decreases. The three estimators produced very similar results, and compared to an ordinary least squares regression line (also illustrated in Figure 4a), which does not appear to offer an appropriate approximation (due to the nonlinearity of income's effect).

Figure 4b presents the impacts of three different bandwidths based on the LP1 estimator. These vary from \$500 to \$15,000 yet *generally* produce very similar results. As is typical, larger bandwidths produce smoother density functions. Here, it appears that bandwidths of \$500 and \$1,000 are too tight to properly capture the density in the vicinity of high incomes, due to the large gaps produced by the 18 NPTS income categories in that range.

< FIGURE 4>

Nonparametric regression can also be a very useful tool to test parametric model specifications (see, e.g., Bowman and Azzalini, 1997). For example, the linear or "LP1" specification of $m(x)$ (as shown in Figure 4a) was tested with various bandwidth values, ranging from \$1,000 to \$100,000. Under all bandwidth values specified, this linear model was rejected at the 0.01 significance level. This provides strong indication that a simple linear dependence of VMT on income, in whatever form (such as OLS) is inappropriate. It also supports use of a more flexible, perhaps nonparametric approach. R^2 values were computed for the linear model and the NW estimator, 0.043 and 0.054, respectively. The NW did not seem to provide large contribution over the linear model in terms of R^2.

3.3.2 Use of Two Regressors

As discussed in Section 2.4, high dimensionality in explanatory variable sets creates difficulties for nonparametric regression. However, regressions involving at least two regressors are often of great interest. Fortunately, in nonparametric regression, control for *indicator*-type variables does not exacerbate the dimensionality problem, and no additional observations are needed (for the same levels of statistical significance). For this reason, recognition of location, retirement status, and home ownership variables does not create problems here.

Two VMT regressions involving two regressors were estimated. The resulting surfaces are presented in Figures 5 and 6. Figure 5 shows the surface relating annual VMT to household

income and vehicle ownership. The relationship between VMT and income seems close to linear, whereas that between VMT and vehicle ownership appears logarithmic. Both have positive effects on annual VMT. However, regardless of income levels, VMT seems to reach a maximum at five or more vehicles.

< FIGURE 5>

According to Figure 6, both income and worker variables hold close-to-linear marginal effects on VMT, if the other variable is held constant. Interestingly, income's marginal effect diminishes as the number of household workers increases. At the extreme of five or six workers, income's effect becomes negligible.

< FIGURE 6>

3.3.3 Use of More Than Two Regressors

Although nonparametric regression with more than two explanatory variables is possible, in order to visualize the results (in three dimensions), values of all but two regressors should be fixed. Here, the reference values or household characteristics are home ownership in an urban area with public transit and a working (non-retired) head. Figures 7a through 7c plot estimates of mean VMT for each nonparametric VMT distribution versus household income and vehicle ownership. In each of these three plots, the number of workers is fixed (at one, two, or four).

< FIGURE 7>

As one would expect, Figure 7 reveals that household VMT rises with income and vehicle ownership. The relationships are noticeably affected by changes in the number of workers. As the series of surfaces suggests, the relationship of VMT to vehicle ownership become quite steep with an increase in the number of workers, resulting in average household VMTs on the order of 30,000 miles per year (for households exhibiting both high incomes and high vehicle ownership). This is likely because a household's workers make good use of any vehicles, and they are constrained if workers exceed vehicles.

The effects of income are rather small, particularly when compared to those of vehicle ownership. In fact, they are almost non-existent in households of four (or more) workers. Evidently, wealthy households can only purchase so much and make so many trips. In addition, high incomes should allow households to undertake purchases at more convenient, though more expensive, stores, as well as pay others to bring goods and services to them. In contrast, high vehicle ownership is a strong indicator that such vehicles are needed and will be used, adding substantially to VMT.

4. CONCLUSIONS

VMT has been widely used for various areas in transportation as a control variable or sometime as a main subject for investigation. Parametric regressions and classfication techniques have been the usual approaches to for VMT estimation, at the regional and/or household levels. This study employed a nonparametric approach to investigate changes in VMT – and in its distribution – based on a variety of control variables, using 1995 NPTS data.

Non-parametric density examinations and non-parametric regression permit modelers dramatic functional flexibility, thus avoiding common issues of overspecification, simplistic behavioral assumptions, and poor predictions. Relationships of response on control variables is permitted via parameterization of a mean function, and the results can either support or undermine the case for certain behavioral assumptions. For example, here a non-parametric regression model with a mean VMT as a linear function of income was strongly rejected. And locally-linear nonparametric regression with seven regressors indicated substantial changes in the relationship between annual VMT and income, vehicle ownership, and workers. For example average VMT in households with 4 workers does not appear to be affected by household income, but is considerably affected by the number of vehicles.

In summary, goodness of model prediction (as measured by chi-squared values for density functions and R2 values for the regression-based densities) was substantially enhanced by nonparametric techniques, relative to OLS results, underscoring the value of such techniques. However, nonparametric regression is more time-consuming and requires exponentially larger sample sizes for additional (non-indicator) control variables. And nonparametric regression results are difficult to appreciate in more than three dimensions (i.e., with more than two explanatory variables). Furthermore, concerns can arise with limited discrete (though not indicator) control variables, due to bandwidth selection issues. Depending on the application, however, nonparametric regression may be clearly superior and should be a handy device to have available in any modeler's toolbox.

Acknowledgements

This study was sponsored by the American Association of State Highway and Transportation Officials (AASHTO), in cooperation with the Federal Highway Administration (FHWA), and was conducted in the National Cooperative Highway Research Program (NCHRP) 17-23. The authors would like to thank Dr. Paul Wilson in the Department of Economics at the University of Texas at Austin for his helpful comments.

Endnotes

[1] The term locally linear means that, in a very narrow region, essentially a linear relationship is being estimated over each observation. The series of these relations is then weighted to produce that point's estimate.

[2] These other agencies are the Bureau of Transportation Statistics (BTS), the Federal Transit Administration (FTA), and the National Highway Traffic Safety Administration (NHTSA)

[3] A household head is self-determine by each responding household. This person may or may not be the eldest, and may or may not be male.

References

Barr, L. C. Testing for the significance of induced highway travel demand in metropolitan areas. *Transportation Research Record 1706*, 2000, pp. 1-8.

Bowman, A. W., and A. Azzalini. *Applied Smoothing Techniques for Data Analysis: The Kernel Approach with S-Plus Illustrations*. Oxford University Press, New York, 1997.

Blundell, R., and A. Duncan. Kernel regression in empirical microeconomics. *The Journal of Human Resources*, Vol. 33, 2000, pp. 62-87.

Clark, S. Traffic prediction using multivariate nonparametric regression. *Journal of Transportation Engineering*, Vol. 129, No. 2, 2003, pp. 161-167.

Federal Highway Administration (FHWA). *User's Guide for the Public Use Data Files: 1995 Nationwide Personal Transportation Survey*. Publication No. FHWA-PL-98-002, Washington D.C., 1997.

Fox, J. *An R and S-Plus Companion to Applied Regression*. Sage Publications, Thousand Oaks, CA, 2002.

Hansen, M., and Y. Huang. Road supply and traffic in California urban areas. *Transportation Research Part A*, Vol. 31, No. 3, 1997, pp. 205-218.

Hardle, W. *Applied nonparametric regression*. Cambridge University Press, New York, 1990

Hardle, W. *Smoothing Techniques With Implementation in S*. Springer-Verlag, New York, 1991.

Harrison, I, and V. Moorhead. *Odometer Versus Self-Reported Estimates of Vehicle Miles Traveled*. July 24, 2000. http://www.eia.doe.gov/emeu/consumptionbriefs/transportation /vmt/vmt.html. Accessed July 17, 2003.

Hauer, E. *Observational Before-After Studies in Road Safety*. Pergamon, Tarrytown, NY, 1997.

Jovanis, P., and H.-L. Chang. Modeling the relationship of accidents to miles traveled. *Transportation Research Record 1068*, 1986, pp. 42-51.

Kumapley, R. K., and J. Fricker. Review of methods for estimating vehicle miles traveled. *Transportation Research Record 1551*, 1996, pp. 59-66.

Kockelman, K. Travel Behavior as a Function of Accessibility, Land Use Mixing, and Land Use Balance: Evidence from the San Francisco Bay Area. *Transportation Research Record 1607*, 1997, pp. 117-125.

Kweon, Y.-J., and K. Kockelman. Overall Injury Risk to Different Drivers: Combining of Exposure, Frequency and Severity Models. *Accident Analysis and Prevention*, Vol. 35, No. 4, 2003, pp. 441-450.

Lave, C. What really is the growth of vehicle usage? *Transportation Research Record 1520*, 1996, pp. 117-121.

Miller, E. J., and A. Ibrahim. Urban form and vehicular travel: Some empirical findings. *Transportation Research Record 1617*, 1998, pp. 18-27.

Mingo, R., and H. K. Wolff. Improving national travel estimates for combination vehicles. *Transportation Research Record 1511*, 1995, pp. 42-46.

Nadaraya, E. A. On estimating regression. *Theory of Probability and Its Applications*, Vol. 9, 1964, pp. 141-142.

Pagan, A., and A. Ullah. *Nonparametric Econometrics*. Cambridge University Press, New York, 1999.

Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons, New York, 1992.

Scott, D. W., and G. R. Terrell. Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, Vol. 82, No. 400, 1987, pp. 1131-1146.

Watson, G. S. Smooth regression analysis. *Sankhya Series A*, Vol. 26, 1964, pp. 359-372.

Weinblatt, H. Using seasonal and day-of-week factoring to improve estimates of truck vehicle miles traveled. *Transportation Research Record 1522*, 1996, pp. 1-8.

You, J., and T. J. Kim. Development and Evaluation of a Hybrid Travel Time Forecasting Model. *Transportation Research Part C*, Vol. 8, No. 1, 2000, pp. 251-256.

Smith, B. L., B. M. Williams, and R. K. Oswald. Comparison of parametric and nonparametric models for traffic flow forecasting. *Transportation Research Part C*, Vol. 10, No. 4, 2002, pp. 303-321.

Kharoufeh, J. P., and K. G. Goulias. Nonparametric identification of daily activity durations using kernel density estimators. *Transportation Research Part B*, Vol. 36, No. 1, 2002, pp. 59-82.

Smith, B. L., and R. K. Oswald. Meeting real-time traffic flow forecasting requirements with imprecise computations. *Computer-Aided Civil and Infrastructure Engineering*, Vol. 18, No. 3, 2003, pp. 201-213.

List of Tables

TABLE 1. Description of Variables

Variable	Description	Mean	Std. Dev.	Num. Obs.
AnnualVMT	Annual vehicle miles traveled (mi/year)	17,906	16,229	17,477
NumWorkers	Number of workers in a household	1.405	0.9318	
NumVehicles	Number of vehicles in a household	1.992	0.8563	
HHIncome	Household income ($/year)	$45,996	$27,900	
PublicTransit	Indicator of public transit availability	0.6286	0.4832	
Urban	Indicator that household resides in an urban area	0.5979	0.4903	
Owned	Indicator that household owns place of residence	0.8229	0.3818	
Retired	Indicator that household head is retired	0.2045	0.4034	

Notes: (1) VMT does not include motorcycles, non-motorized household vehicles or mopeds. (2) 18 NPTS income categories were converted to a continuous variable here, by assigning the middle value of each income category.

List of Figures

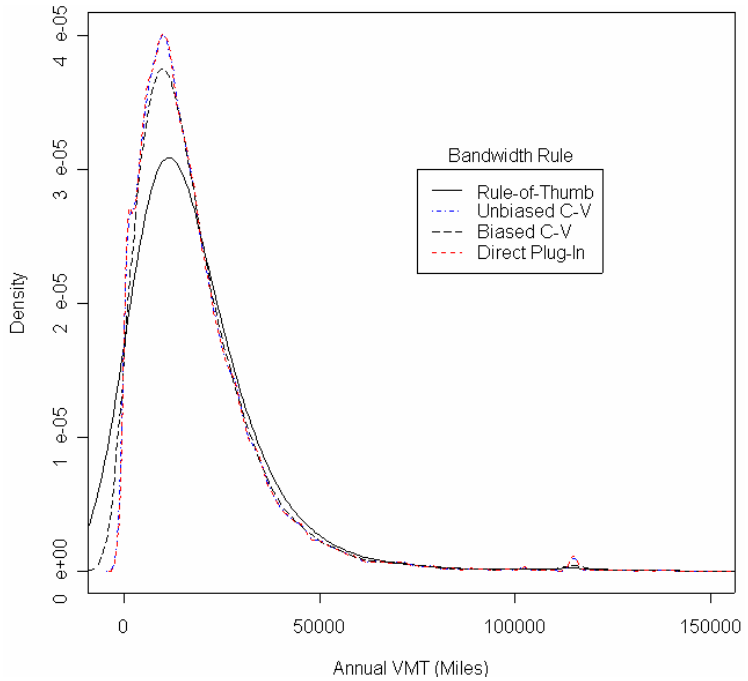


FIGURE 1. Kernel density estimates of annual VMT per household.
(Actual density extends beyond 150,000 VMT/year.)

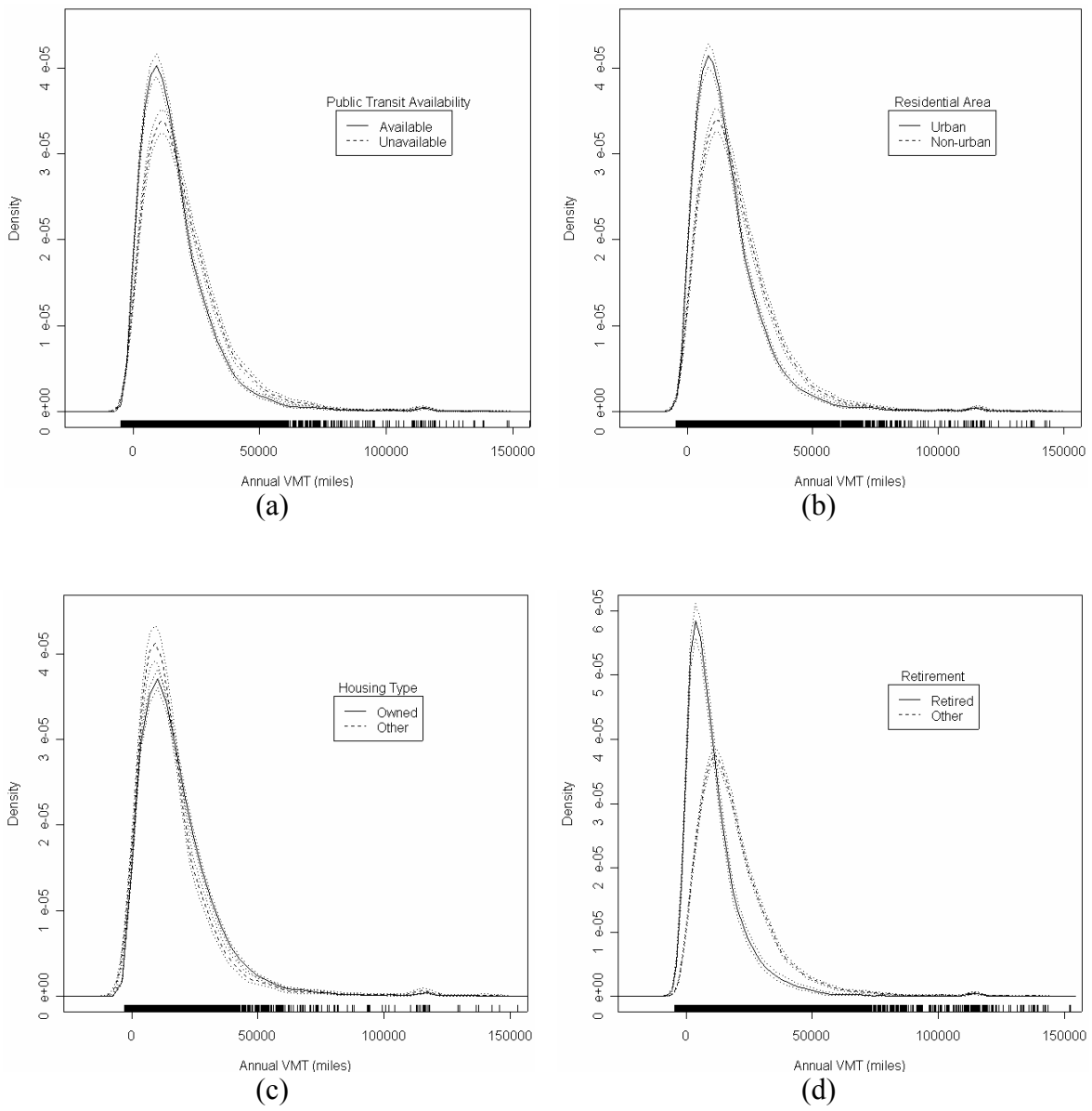
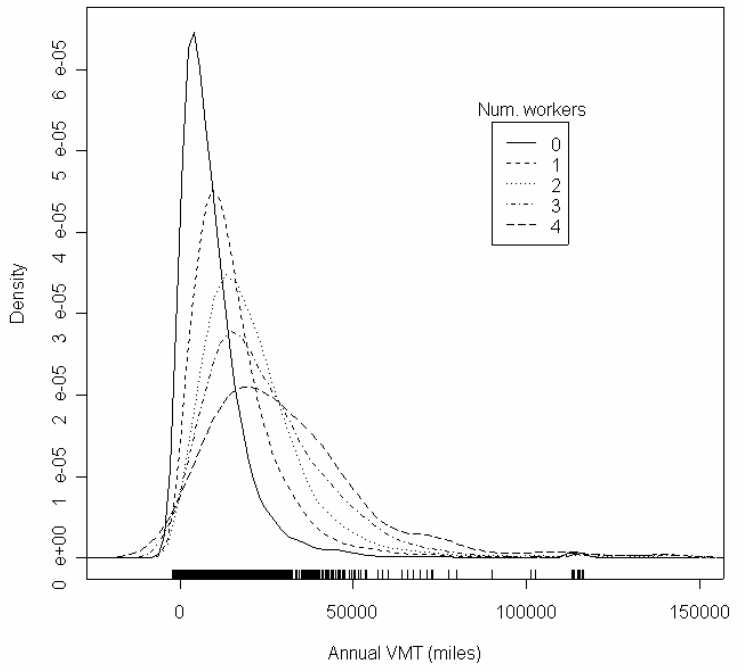
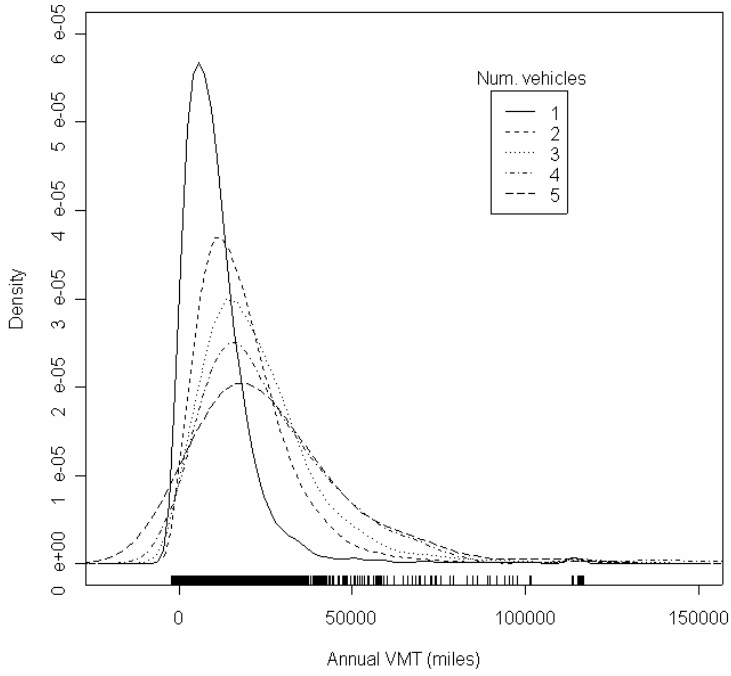


FIGURE 2. VMT density functions defined by public transit availability, urban/rural area, housing type, and retirement status.
(Dots represent standard errors of density value estimates.)

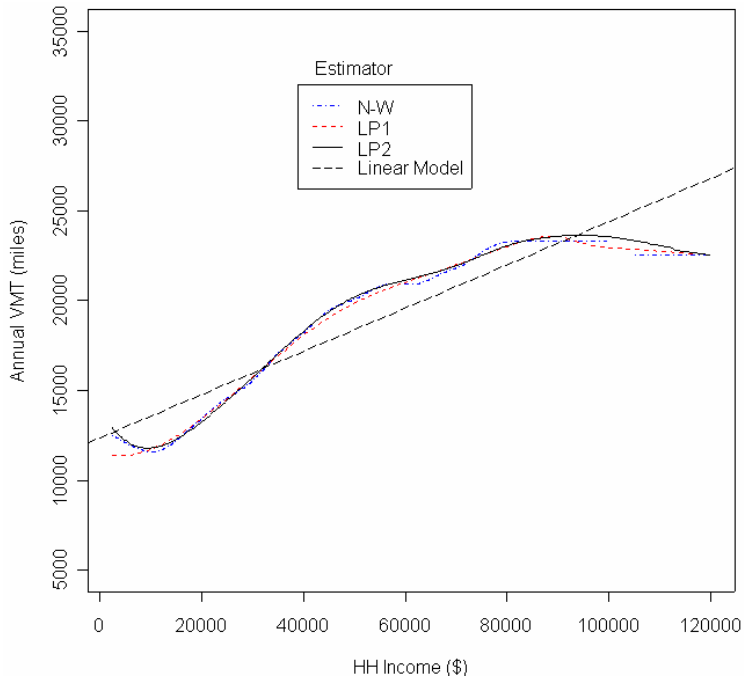


(a)

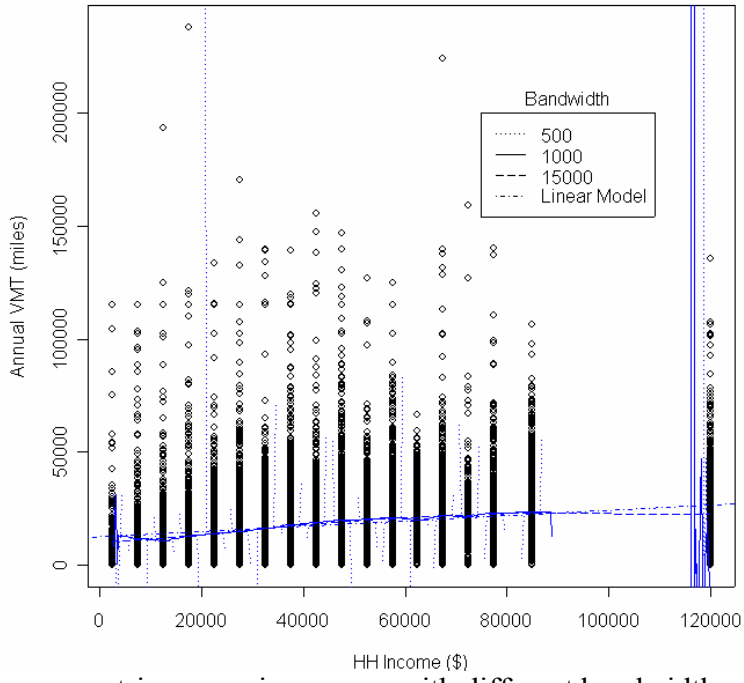


(b)

FIGURE 3. VMT density functions defined by the numbers of household workers and vehicles.



(a) Nonparametric regression curves with different estimators



(b) LP1-type nonparametric regression curves with different bandwidths, plus data points

FIGURE 4. Mean VMT estimates as a function of household income.

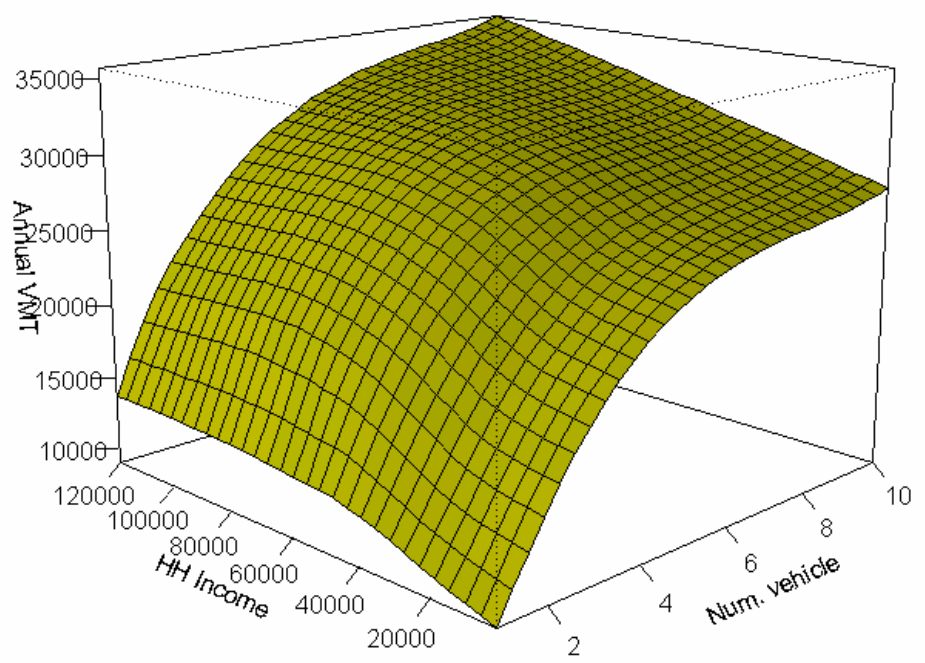


FIGURE 5. Mean VMT estimates as a function of household income and vehicle ownership.

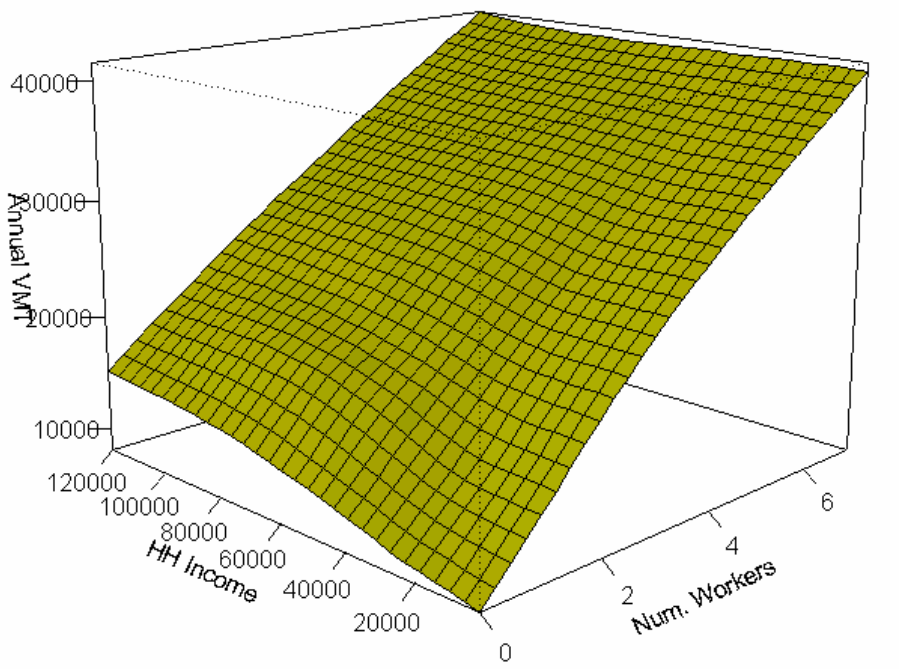
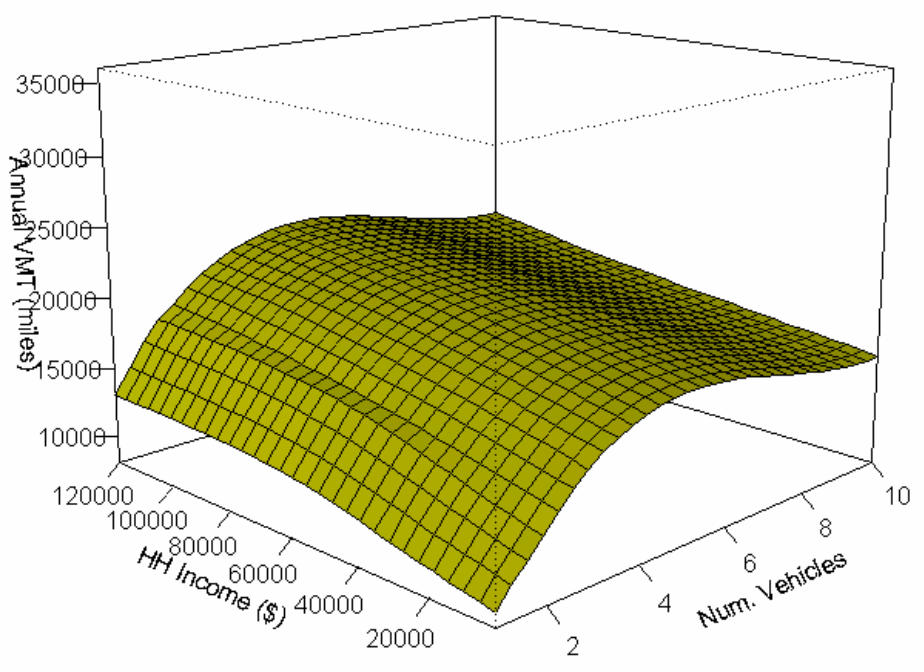
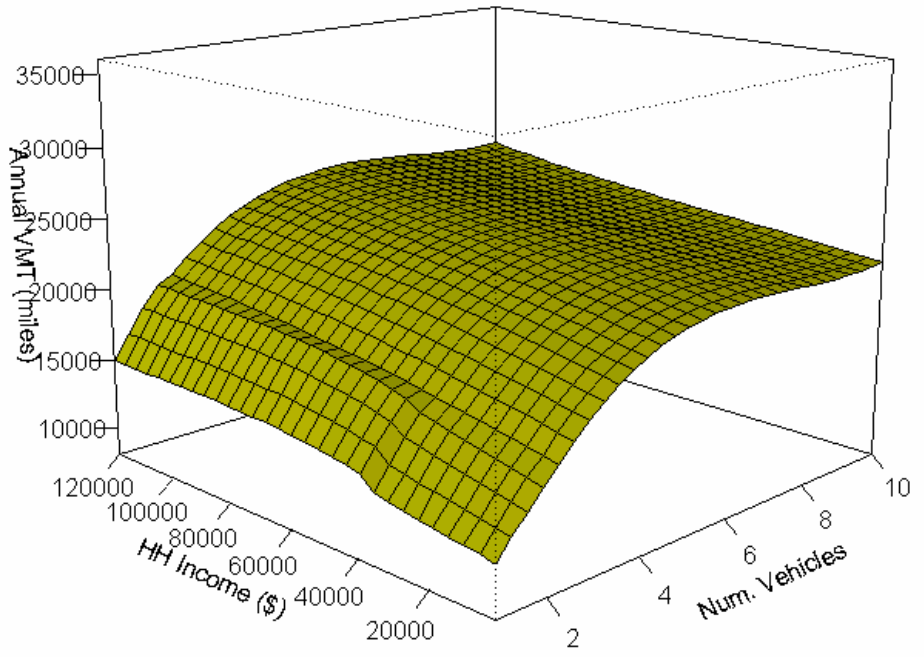


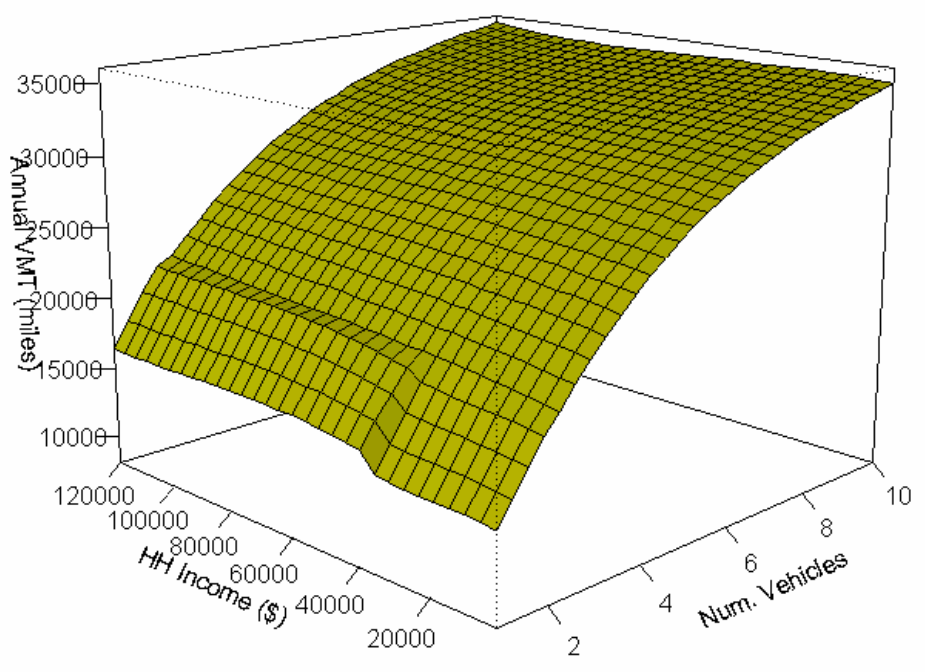
FIGURE 6. Mean VMT estimates as a function of household income and workers.



(a) Number of workers: 1



(b) Number of workers: 2



(c) Number of workers: 4

FIGURE 7. Mean VMT estimaties as a function of incomeand vehicle ownership, assuming a fixed number of workers (one, two and four).

(The presented data assume a household that owns its residence, in an urban area with public transit and a household head who is not retired.)