

**SELF-SELECTION IN HOME CHOICE:
USE OF TREATMENT EFFECTS IN EVALUATING THE RELATIONSHIP BETWEEN
THE BUILT ENVIRONMENT AND TRAVEL BEHAVIOR**

By

Bin (Brenda) Zhou
Graduate Student Researcher
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
6.508. Cockrell Jr. Hall
Austin, TX 78712-1076
brendazhou@mail.utexas.edu

and

Kara M. Kockelman
(Corresponding author)
Associate Professor and William J. Murray Jr. Fellow
Department of Civil, Architectural and Environmental Engineering
The University of Texas at Austin
6.9 E. Cockrell Jr. Hall
Austin, TX 78712-1076
kkockelm@mail.utexas.edu
Phone: 512-471-0210
FAX: 512-475-8744

April 2008

Forthcoming in *Transportation Research Record*, 2008

ABSTRACT

The issue of self-selection's role in shaping travel patterns, by impacting one's home location choice, is a critical question. Developers, planners and policymakers regularly debate to what extent the built environment and land use patterns can alleviate roadway congestion, greenhouse gas emissions and myriad other urban problems. This study illustrates the use of Heckman's latent index model to ascertain travel impacts of neighborhood type in Austin, Texas. Under this approach, self-selection is formulated as sample selection bias in receiving a treatment. Here, treatment is defined to be one's residence in a suburban or rural zone, rather than Austin's central business district (CBD) and nearby urban zones. This treatment/no-treatment approach is a meaningful advance in models of self-selection effects, and requires estimation of three straightforward models. Depending on model specification used, results suggest that at least half (58% to 90%) of differences in vehicle-miles-traveled observed between similar households living in CBD/urban versus rural/suburban neighborhoods of Austin is due to the location or treatment itself, while self-selection of such treatment (by households that wish to meet special travel needs and/or preferences) accounts for the remainder.

INTRODUCTION

The interaction between land use and transportation has been recognized by researchers from different disciplines for decades. Facing the negative consequences of personal vehicle dominance, such as congestion, air pollution and global warming, New Urbanists propose land use patterns to moderate travel demand. They argue for changing the built environment to reduce the number of motorized trips, increase the share of non-motorized modes, reduce travel distances and increase vehicle occupancy of motorized trips (1). Since the early 1990s, a rich literature has investigated the relationship between the physical features of the urban landscape, transport policies, and travel behavior. Ewing and Cervero (2001) provide a comprehensive review of such studies (2).

Almost all studies use different data sets and geographic scales and focus on different aspects of travel behavior (e.g. household vehicle miles traveled, person miles traveled, number of trip chains, and mode split, for work or non-work trip purposes). They also draw different conclusions in terms of the statistical and practical significance of the built environment's impact on travel behavior. In general, early work has used more aggregate statistics and later work has used more disaggregate data. Several researchers have relied on quasi-experimental designs (e.g., pairing matched neighborhoods) in order to discern travel distinctions related to a few key design features (see, e.g., 3, 4, 5, 6). Others have used cross-sectional data and regression techniques to quantify the travel impacts of one's built environment (see, e.g., 7, 8). Krizek (2003) relied on longitudinal data involving household relocations to disentangle the relationship of travel behavior and the built environment (proxied by a single measure of neighborhood accessibility) (9). In general, much work supports, to some degree, the assertions of New Urbanists. However, use of attitudinal data in more recent studies, to correct for self-selection bias (due to residential sorting), can suggest little influence of the built environment, thereby highlighting the importance of self-selection issue.

Attitudes are typically difficult to measure, and experimental designs are sometimes infeasible. (For example, concerns regarding respondent burden often preclude the inclusion of attitudinal and stated preference questions at the end of a travel survey, and we typically cannot observe the same household living in different environments at nearly the same time.) Most researchers have had to apply appropriate econometric techniques to estimate the causal effects. Mokhtarian and Cao (2008) listed the following seven approaches to help address the self-selection issue, and possibly disentangle the relationship between the built environment and travel behavior: direct questioning, statistical control, instrumental variables models, sample selection models, joint discrete choice models, structural equations models and longitudinal designs (10).

This study falls under the sample selection approach, also called *latent index model*. This method not only provides consistent estimators of variables of interest, but also allows one to derive the treatment parameters that quantify the effects of self-selection in the context of specific environments. This method was applied to investigate daily vehicle miles traveled (VMT) by households surveyed in Austin, Texas. The following sections discuss related research, the data sets and methods used here, as well as model results, and conclusions.

LITERATURE REVIEW

Using different data and geographic scales, prior studies tend to draw different conclusions regarding the relationship between built environment and travel behavior. Most studies found that the built environment influenced travel behaviors. More recent studies that have controlled explicitly for attitudinal characteristics concluded that the built environment have insignificant impacts on travel behavior.

Cervero and Kockelman (1997) factorized built environment attributes into three principal dimensions (density, diversity and design, which they called the 3Ds) and examined how these variables influenced VMT per household, as well as work and non-work mode choice (*1*). They controlled for a variety of variables and found reduced trip rates and more non-auto travel when the household's built environment was characterized by higher densities, higher land use mixing and better pedestrian environments. This study analyzed the impacts of the built environment on travel behavior from a comprehensive perspective; but, as noted by the authors, the cross-sectional analysis limits the results, making them more associative, rather than causal.

In addition to the complex nature of describing the built environment (hundreds of variables may be needed), self-selection is an issue in disentangling the travel-environment relationship. People may choose a residential location in order to realize desired travel patterns, so that observed differences in travel behavior of households living in different locations do not relate to differences in the built environment alone. In other words, the impact of environment on behavior would be over-estimated if individuals' behavioral preferences help determine the environment. As a result, the efficacy of land use policies that focus on altering the built environment to shape travel demand could be exaggerated. Several research efforts have addressed the self-selection issue by controlling for the attitudes of trip makers.

Using attitude surveys and travel diaries, Kitamura et al. (1997) examined the impacts of the built environment and attitudinal characteristics on measures of mobility (including the number and fractions of personal trips, transit trips, and non-motorized trips) (*11*). They calibrated two sets of regression models: one with demographics and neighborhood characteristics and the other with additional factors that reflect attitudes towards travel behaviors. The authors found that attitudinal variables explained the highest proportion of data variation and were more strongly related with travel behavior than with measures of the built environment.

Bagley and Mokhtarian (2002) applied a system equations model (SEM) using nine endogenous variables (*12*): two residential location types (traditional and suburban), three measures of travel demand (miles traveled by different modes), three measures of attitude (pro-high density, pro-driving, and pro-transit), and one measure of job location (commute distance). Attitudinal variables were found to have the greatest impact on travel behavior, while residential location classifications appeared to have little impact. The authors argued that the "observed" association between the built environment and travel behavior is due to correlations among built environment and attitudinal variables, and thus self-selection exists.

The above studies contribute to an understanding on environmental factors, but require additional information on travelers' attitudes, which are not available in most household travel surveys. Other researchers seek to disentangle such relationships using neighborhood matched pairs (conventional versus neotraditional designs, for example), but sample sizes tend to be small and the issue of self-selection remains. With only the observational data on hand, researchers have to apply appropriate econometric techniques to estimate the effects.

Aware of the potential correlation between built environment attributes and error terms, Boarnet and Sarmiento (1998) applied instrumental variable models to control for such correlation, and concluded that information on the relationship between (non-work) travel and local land use did not support or oppose New Urbanist assertions (13). In contrast, Greenwald (2003) found some variables that represent built environment are statistically significant *after* controlling for residential self-selection, using an “extended” latent index model (14). Essentially, he converted a binary selection into a multinomial model of residential choice (combination of locations and tenure), and then added the predicted probability of the residential choice into eight equations that represent ratios of travel times between modes by purposes. However, his model specification differs from a standard latent index model, and its ability to control for residential self-selection is unclear.

Finally, Bhat and Guo (2007) applied a joint model of residential location choice and car ownership decisions, considering observed and unobserved variations in sensitivity to the built environment (15). They concluded that built environment attributes affect residential choice decisions as well as car ownership decisions.

In contrast to prior work in this area, this study investigates the effect of environment on household VMT using latent index models, as introduced by Heckman (16, 17). This method not only provides consistent estimators of variables of interest, but also allows one to derive the treatment parameters that quantify the effects of individuals’ self-selection into a dichotomous measure of neighborhood type. This method was applied to investigate household daily VMT in the Austin, Texas region. The following sections discuss the data sets and methods used here, as well as model results, and conclusions.

DATA DESCRIPTION

The primary data sources used here are the 1998-1999 Austin (Household) Travel Survey results and ArcGIS-encoded zonal data for the Austin region (including Williamson, Travis and Hays counties), as obtained from the Capital Area Metropolitan Planning Organization (CAMPO).

The response of interest is household VMT on the survey day (either a weekday or a Sunday), as obtained via start and ending odometer readings on all the vehicles owned by the household over the survey period. Households reporting more than 1,000 VMT on the survey day or having a household head less than 18 years of age were eliminated, resulting in a final sample of 1,903 household observations.

Since anonymity protections permit household location reporting only at the zonal level, zone-level (rather than parcel-level) attributes were linked to each household, according to zone of residence. Figure 1 shows the locations of all traffic analysis zones (TAZs) having at least one observation in the final sample, with Williamson County (to the north) exhibiting relatively low coverage (resulting from its lower population). Figure 1 also identifies the location of zones coded as belonging to the region’s central business district (CBD) or urban areas, versus those containing land use patterns of a more suburban or rural type. These zones/area types were coded by CAMPO, based on the Texas DOT formula, using combination of employment and household density values – with defining thresholds of 8, 3 and 1 person-equivalents per acre (where equivalent population is simply zone population plus zone employment times the regional persons-per-job ratio). These thresholds are relatively arbitrary, and other values (which are suitable for the study area) may serve such purposes just as well or better. Since latent index models deal with a binary treatment index (either treated or un-treated), TAZs in the study area

that were classified as CBD and urban were grouped into one category, and rural and suburban TAZs were grouped into a second category.

One may expect that, on average, households living in low-density areas tend to drive more than those in downtown areas – everything else constant. In the following analyses, households living in rural or suburban areas are said to be “treated” while those living in CBD and urban areas are said to be “un-treated”. As suggested, only one treatment type is allowed in this framework, rendering the neighborhood conditions rather aggregate. This may not be much of a limitation in clinical trials of different drugs, but it can pose a serious significant limitation when wanting to appreciate the effects of *multiple* neighborhood types on travel behavior and the like. Nevertheless, the model framework is a useful one, and can illuminate certain relationships, as described here.

The following sections discuss how to consistently estimate the impacts of built environment on a key response variable (household daily VMT in this case), along with several measures of the “treatment” effects.

METHODOLOGY

Model Specification and Treatment Parameters

A common approach to dealing with selection bias is use of a latent index model, which relates the treatment to the likelihood of potential treatment outcomes. More specifically, individuals receive treatment if the net “utility” of doing so is positive and do not receive treatment if the net utility is negative. Potential-outcome equations (household daily VMT) are specified as follows:

$$Y_i^1 = \mathbf{X}_i \boldsymbol{\beta}^1 + \varepsilon_i^1 \quad (1)$$

$$Y_i^0 = \mathbf{X}_i \boldsymbol{\beta}^0 + \varepsilon_i^0 \quad (2)$$

where Y^1 and Y^0 are the potential outcomes of treated and untreated individuals, respectively. Of course, each observation has only one state, so either Y^1 or Y^0 is observed for each individual – not both. \mathbf{X}_i is a row vector of the observed explanatory variables for individual i , and ε^1 and ε^0 are unobserved random variables.

Letting D_i denote the observed treatment decision of individual i (with $D_i = 1$ meaning the receipt of treatment [rural or suburban location] and $D_i = 0$ implying no treatment [CBD or urban location]), a selection equation essentially generates D_i via latent variable D_i^* , as follows:

$$D_i^* = \mathbf{Z}_i \boldsymbol{\theta} + \varepsilon_i^D \quad (3)$$

where \mathbf{Z}_i is a row vector of observed explanatory variables for individual i , and ε_i^D is the unobserved random variable, affecting this outcome. Conventionally, $D_i = 1$ if $D_i^* \geq 0$, and zero otherwise. It is important to note that this utility function, which determines treatment assignments, can be a combination of preferences of individuals and others, as well as treatment availability (such as the number of homes for sale in rural or suburban TAZs). A key idea is that the treatment assignment is “random”, to some extent, depending on the size of the additive error term, ε_i^D , relative to the systematic component ($\mathbf{Z}_i \boldsymbol{\theta}$) of the “utility function” D_i^* .

Using the above specification, the measured outcome for individual i can be given as follows:

$$\begin{aligned}
Y_i &= D_i Y_i^1 + (1 - D_i) Y_i^0 = D_i (\mathbf{X}_i \boldsymbol{\beta}^1 + \varepsilon_i^1) + (1 - D_i) (\mathbf{X}_i \boldsymbol{\beta}^0 + \varepsilon_i^0) \\
&= D_i \mathbf{X}_i \boldsymbol{\beta}^1 + (1 - D_i) \mathbf{X}_i \boldsymbol{\beta}^0 + D_i (\varepsilon_i^1 - \varepsilon_i^0) + \varepsilon_i^0 = D_i \mathbf{X}_i \boldsymbol{\beta}^1 + (1 - D_i) \mathbf{X}_i \boldsymbol{\beta}^0 + u_i
\end{aligned} \tag{4}$$

In latent index models, ε 's are assumed to be independent of \mathbf{X} 's. However, u_i is correlated with D_i , which leads to endogeneity bias. In order to consistently estimate the model, the expected value of ε_i^D is needed to serve as a control variable in Equations (1) and (2). The expected value of ε_i^D can be estimated using Equation (3), given an (assumed) distribution for ε_i^D .

Following model estimation, interest lies in various measures of treatment effectiveness. The parameter estimates described in the above three models do not provide information on how individuals are self-selected. The post-processed treatment effect parameters serve this purpose. Heckman and Vytlačil (1999) and Heckman et al. (2001) focus on the following four: the average treatment effect (ATE), the effect of treatment on the treated (TT), the local average treatment effect (LATE), and the marginal treatment effect (MTE) (18, 19). Each of these describes a specific perspective in evaluating the effect of treatment.

Average Treatment Effect (ATE)

Among the four effects, the ATE is perhaps of greatest interest since it is the expected change in outcome (Y_i) from the treatment of a randomly selected individual. More specifically, it produces the expected VMT increase when moving a randomly selected household between an urban/CBD zone and a rural or suburban TAZ.

$$ATE(\mathbf{x}_i) = E(Y_i^1 - Y_i^0 \mid \mathbf{X} = \mathbf{x}_i) = \mathbf{x}_i (\boldsymbol{\beta}^1 - \boldsymbol{\beta}^0) \tag{5}$$

This parameter is conditional on the distribution of \mathbf{X} . The unconditional estimate is obtained by integrating the equation over \mathbf{X} 's multivariate distribution:

$$ATE = E(Y_i^1 - Y_i^0) = \int ATE(\mathbf{X}) dF(\mathbf{X}) \approx \frac{1}{n} \sum_{i=1}^n ATE(\mathbf{x}_i) \tag{6}$$

Effect of Treatment on the Treated (TT)

The TT is the expected outcome gain from the treatment for individuals that select the treatment option. In this study, it represents the expected additional VMT of households located in a rural or suburban TAZ.

$$\begin{aligned}
TT(\mathbf{x}_i, \mathbf{z}_i, D_i = 1) &= E(Y_i^1 - Y_i^0 \mid \mathbf{X} = \mathbf{x}_i, \mathbf{Z} = \mathbf{z}_i, D_i = 1) \\
&= \mathbf{x}_i (\boldsymbol{\beta}^1 - \boldsymbol{\beta}^0) + E(U_i^1 - U_i^0 \mid -\mathbf{z}_i \boldsymbol{\theta} \leq \varepsilon_i^D)
\end{aligned} \tag{7}$$

This parameter is conditional on the joint distribution of \mathbf{X} and \mathbf{Z} , so integration over $F(\mathbf{X}, \mathbf{Z} \mid D_i = 1)$ leads to the unconditional estimate, as follows:

$$\begin{aligned}
TT &= E(Y_i^1 - Y_i^0 \mid D_i = 1) = \int TT(\mathbf{X}, \mathbf{Z}, D_i = 1) dF(\mathbf{X}, \mathbf{Z} \mid D_i = 1) \\
&= \frac{1}{n} \sum_{i=1}^n D_i \times TT(\mathbf{x}_i, \mathbf{z}_i, D_i = 1)
\end{aligned} \tag{8}$$

Local Average Treatment Effect (LATE)

The LATE is the expected outcome gain for individuals induced to experience the treatment by a

change in an instrument variable, from $\mathbf{Z}_i = \mathbf{z}_i$ to $\mathbf{Z}_i = \mathbf{z}_i$. The instrument variable is the “primary exclusion restriction” (19); it affects the treatment decision, but not the outcomes.

$$\begin{aligned} LATE(D[\mathbf{z}_i]=0, D[\mathbf{z}_i]=1, \mathbf{x}_i) &= E(Y_i^1 - Y_i^0 \mid D[\mathbf{z}_i]=0, D[\mathbf{z}_i]=1, \mathbf{X} = \mathbf{x}_i) \\ &= \mathbf{x}_i(\boldsymbol{\beta}^1 - \boldsymbol{\beta}^0) + E(U_i^1 - U_i^0 \mid -\mathbf{z}_i\boldsymbol{\theta} \leq \varepsilon^D \leq -\mathbf{z}_i\boldsymbol{\theta}) \end{aligned} \quad (9)$$

where $[w]$ denotes a function of variable w . The unconditional estimate of this effect is as follows:

$$\begin{aligned} LATE &= E(Y_i^1 - Y_i^0 \mid D[\mathbf{z}_i]=0, D[\mathbf{z}_i]=1) = \int LATE(D[\mathbf{z}_i]=0, D[\mathbf{z}_i]=1, \mathbf{X})dF(\mathbf{X}) \\ &\approx \frac{1}{n} \sum_{i=1}^n LATE(D[\mathbf{z}_i]=0, D[\mathbf{z}_i]=1, \mathbf{x}_i) \end{aligned} \quad (10)$$

As Heckman et al. (19, pp. 215) explain, this parameter “corresponds to the treatment effect for individuals who would not select into treatment if their vector \mathbf{Z} was set to \mathbf{z}_k (all other components of \mathbf{Z} unchanged), but would select into treatment if \mathbf{Z} was set to \mathbf{z}'_k (\mathbf{z}' in the original literature).”

Marginal Treatment Effect (MTE)

The MTE is the expected outcome gain for individuals with a given value of ε^D , which means this parameter measures the average outcome gain for the individuals “who are just indifferent to the receipt of treatment when the $\mathbf{z}\boldsymbol{\theta}$ index is fixed at the value $-\varepsilon^D$ ” (19, pp. 216).

$$\begin{aligned} MTE(\mathbf{x}_i, \varepsilon_i^D) &= E(Y_i^1 - Y_i^0 \mid \mathbf{X} = \mathbf{x}_i, E_i^D = \varepsilon_i^D) \\ &= \mathbf{x}_i(\boldsymbol{\beta}^1 - \boldsymbol{\beta}^0) + E(U_i^1 - U_i^0 \mid E_i^D = \varepsilon_i^D) \end{aligned}$$

(11)The unconditional estimate is as follows:

$$MTE(\varepsilon_i^D) = E(Y_i^1 - Y_i^0 \mid E_i^D = \varepsilon_i^D) = \int MTE(\mathbf{X}, \varepsilon_i^D)dF(\mathbf{X}) \approx \frac{1}{n} \sum_{i=1}^n MTE(\mathbf{x}_i, \varepsilon_i^D) \quad (12)$$

Trivariate Normality and Estimation

Under an assumption of trivariate normality across the three error terms, estimates of these four treatment effects enjoy closed-form expressions. In addition, the model can be consistently estimated using a straightforward two-step procedure (as described in 19). Beyond trivariate normality, Heckman and Vytlačil (2005) have described estimation procedures for more general error term specifications (20). In these cases, the four treatment effects can be expressed as weighted averages of MTE, involving integral calculations.

First, one assumes that the error terms (ε^1 , ε^0 , and ε^D) are jointly normally distributed:

$$\begin{bmatrix} \varepsilon^D \\ \varepsilon^1 \\ \varepsilon^0 \end{bmatrix} \sim N \left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \sigma^{1D} & \sigma^{0D} \\ \sigma^{1D} & \sigma^{11} & \sigma^{10} \\ \sigma^{0D} & \sigma^{10} & \sigma^{00} \end{bmatrix} \right)$$

Equation (5) (for conditional ATE) can not be further simplified because the distributional assumption does not change the functional form. However, Equation (7) (for conditional TT) can now be given as follows:

$$TT(\mathbf{x}_i, \mathbf{z}_i, D_i = 1) = \mathbf{x}_i(\boldsymbol{\beta}^1 - \boldsymbol{\beta}^0) + (\rho^1\sigma^1 - \rho^0\sigma^0) \frac{\phi(\mathbf{z}_i\boldsymbol{\theta})}{\Phi(\mathbf{z}_i\boldsymbol{\theta})} \quad (13)$$

where $\rho^1 = \text{Corr}(\varepsilon^1, \varepsilon^D)$, $\rho^0 = \text{Corr}(\varepsilon^0, \varepsilon^D)$, $\phi()$ is the standard normal density function, and $\Phi()$ is the standard normal cumulative distribution function. Due to normalization of the variance of the error term in the selection equation ($\sigma^{\text{DD}} = 1$), one finds that $\rho^1\sigma^1 = \sigma^{\text{1D}}$ and $\rho^0\sigma^0 = \sigma^{\text{0D}}$.

Equation (9) for the conditional LATE is then given as:

$$LATE(D[\mathbf{z}_i] = 0, D[\mathbf{z}_i] = 1, \mathbf{x}_i) = \mathbf{x}_i(\boldsymbol{\beta}^1 - \boldsymbol{\beta}^0) + (\rho^1\sigma^1 - \rho^0\sigma^0) \frac{\phi(\mathbf{z}_i\boldsymbol{\theta}) - \phi(\mathbf{z}_i\boldsymbol{\theta})}{\Phi(\mathbf{z}_i\boldsymbol{\theta}) - \Phi(\mathbf{z}_i\boldsymbol{\theta})} \quad (14)$$

And Equation (11) for the conditional MTE can be written as:

$$MTE(\mathbf{x}_i, \varepsilon_i^D) = \mathbf{x}_i(\boldsymbol{\beta}^1 - \boldsymbol{\beta}^0) + (\rho^1\sigma^1 - \rho^0\sigma^0)\varepsilon_i^D \quad (15)$$

It is apparent that when ε^D equals zero, *MTE* reduce to *ATE* because of the symmetry of the normal distribution.

Heckman (1976) proposed a two-step procedure to estimate such latent index models (16), and Heckman et al. (2001) provide a detailed procedure for doing so (19), as summarized here now:

Step 1. Run a binary probit model to obtain $\hat{\boldsymbol{\theta}}$ for the treatment decision and then use $\hat{\boldsymbol{\theta}}$ to compute the selection-correction terms (the expectation of the control variables). These expectations are defined as follows:

$$E(\varepsilon^D | \mathbf{z}_i\hat{\boldsymbol{\theta}}, D_i = 1) = \frac{\phi(\mathbf{z}_i\hat{\boldsymbol{\theta}})}{\Phi(\mathbf{z}_i\hat{\boldsymbol{\theta}})} \quad (16)$$

$$E(\varepsilon^D | \mathbf{z}_i\hat{\boldsymbol{\theta}}, D_i = 0) = -\frac{\phi(\mathbf{z}_i\hat{\boldsymbol{\theta}})}{1 - \Phi(\mathbf{z}_i\hat{\boldsymbol{\theta}})} \quad (17)$$

Step 2. Estimate two OLS regression models for *Y* values: one for groups that received the treatment (households living in rural or suburban Austin) and another for groups that did not, using the appropriate selection-correction terms. Use the estimation results (e.g. $\hat{\boldsymbol{\beta}}^1$, $\hat{\boldsymbol{\beta}}^0$, $\hat{\rho}^1\hat{\sigma}^1$ and $\hat{\rho}^0\hat{\sigma}^0$) to obtain estimates of the treatment parameters, given \mathbf{X} , \mathbf{Z} and \mathbf{Z} .

MODEL RESULTS

In this study, Y^1 denotes household daily VMT for those living in rural or suburban neighborhoods, and Y^0 denotes VMT for those in CBD or urban neighborhoods. Model specifications and analytical results are described below.

A Model of Treatment Selection

The explanatory variables in the binary probit model of treatment selection include an intercept, household size, number of workers, number of children under 5 years of age, household annual income, and age of household head. This last variable was chosen to serve as a “primary exclusion restriction” (19) that affects the treatment decision (e.g., residential location choice) but should not affect potential outcomes (e.g., household VMT). This variable is needed to

calculate the LATE. Attitudinal variables, such as strength of preference for living in less-populated area, are ideally suited to this purpose. However, age of household head appears to be the only variable in the Austin Travel Survey that may be able to fulfill this role. Household income, originally a categorical variable, was transformed into a single, quasi-continuous variable using mid-point values.

The model estimation shows that number of workers and number of children variables are not statistically significant, and were removed from the model specification. Table 1 provides summary statistics of all explanatory variables used in the final treatment-selection model, and binary probit model results are shown in Table 2.

Not surprisingly, household size is estimated to positively impact the probability of one's living in a rural or suburban area. Homes in these less intensely developed neighborhoods tend to be larger, single-family houses, and thus generally favored by larger households. The positive signs on household annual income and age of household head indicate that higher-earning households and those more advanced in age are more likely to live in rural or suburban areas, everything else constant.

Models of Treatment Outcome

Outcome models (for household daily VMT) are based on Equations (1) and (2), corresponding to the two treatment-specific groups (i.e., those living in rural or suburban areas, versus those living in CBD or urban areas). Explanatory variables include household size, number of workers, number of children, household annual income, number of vehicles, an indicator for whether proximity to work or school influenced their decision to locate in their current residential location (named Close to Work or School), an indicator for the presence of household member(s) driving a delivery vehicle (Delivery Driver[s]), and various neighborhood attributes (including median income of the home zone, population and household densities, and zonal employment, as determined by CAMPO for year 1997).

Estimation results suggest that household size, number of children, "Close to Work or School", median income of the home zone and household density are statistically insignificant in both equations, so these are not included in the final model. Since calculation of Heckman's four treatment parameters requires the same number of explanatory variables in each of the two equations, variables that are statistically insignificant in just one of the equations are retained in both models. Table 3 gives summary statistics of model variables for both treatment groups, and Table 4 provides estimates of all model parameters.

As expected, the numbers of workers and vehicles have positive impacts on household daily VMT, thanks to commute needs and easier access to vehicles. The presence of a delivery driver (as a household member) also increases the daily VMT. Interestingly, population and employment densities (of the home zone) are estimated to have opposing effects on household VMT, depending on location: significantly negative (as expected) in rural/suburban locations yet slightly positive in CBD/urban locations. It may be that households living in downtown Austin are making more shopping and recreational trips, often by car, when there are more opportunities nearby, and everything else constant.

Size of Treatment Effects

Heckman's four treatment parameters were calculated using Equations (5), (13), (14) and (15). The *ATE* is estimated to be 17.0 vehicle miles per day, which means a randomly selected

household is expected to increase its daily VMT by 17.0 miles when living in a rural or suburban neighborhood, as compared to living in a CBD or urban neighborhood (within the Austin region). Given the average Austin household's 61.3 daily VMT, 17.0 represents more than a 27% savings in daily VMT.

The *TT* was estimated to be 29.2 miles, suggesting that a household living in a rural or suburban neighborhood can be expected to exhibit 29.2 more daily VMT than one living in a CBD or urban neighborhood, all other attributes constant. Based on the size of these two effects (*ATE* and *TT*), the impacts of the “built environment” on household daily VMT (i.e., the VMT increase due to living in a rural or suburban area of Austin, rather than centrally or downtown) is estimated to be 58% of the as-observed differences in treated and non-treated households. This implies that self-selection accounts for 42% of observed VMT differences across Austin households in suburban/rural versus CBD/urban zones. Essentially then, moving all rural or suburban residents into CBD or urban zones may be expected to yield lower VMT savings than analysts may perceive at first glance.

It is worth mentioning that results on treatment effects are not highly robust to model specifications. When using the presence of four or more visitors on the travel survey day as the “primary exclusion restriction” variable (which affects the treatment decision but not the potential outcomes), the *ATE* is estimated to be 20.2 vehicle miles per day, the *TT* is estimated to be 22.5 vehicle mile per day, and the impacts of the “built environment” on household daily VMT are then estimated to be 90% of the observed differences in treated and non-treated households. However, the number of visitors varies from day to day, and so does not robustly represent a household, like age of household head. Certainly, different contexts will impact this ratio as well – for example, VMT by households residing in a transit-oriented station area versus those residing just outside. Ideally, Heckman's latent index model will be tested in a variety of contexts, with a variety of specifications, using more appropriate variables than one usually finds in standard travel surveys.

CONCLUSIONS

Land use-transportation interactions present a complex problem, and self-selection in location choice is a difficult issue to address properly. To date, different approaches, using different data sets and geographic scales, draw rather different conclusions. Seeking a statistically defensible approach, this study applied Heckman's latent index model to estimate the impact of neighborhood type on household daily VMT. Based on a sample of 1,903 Austin households residing in either rural or suburban zones (“treated”) versus CBD or urban zones (the “untreated” population), results suggest that this binary measure of neighborhood conditions is associated with significant changes in a household's daily VMT: 17.0 more VMT per day per household when moving a household from a CBD/urban zone to a rural/suburban location, everything else constant. Essentially, the data indicate that at least 58% of observed VMT differences are due to the household's location while the remaining 42% can be attributed to self-selection. This result provides support for New Urbanist claims, in that “built environment” attributes account for the majority of observed VMT differences. However, self-selection also is evident, as households, to some extent, choose a residential location in order to realize desired travel patterns. Therefore, modifications of land use patterns and the “built environment” can moderate automobile reliance, but to a lesser extent than is often argued, based on standard analyses of VMT data.

In addition to providing such estimates, this paper sought to illustrate the modeling

paradigm's potential. Heckman's approach is powerful, for a variety of contexts. Nevertheless, several enhancements can be made. For example, detailed address information on home, workplaces, and trip destinations is obscured (with such locations tied only to a TAZ). Without exact locations, it is difficult to construct more careful and meaningful measures of neighborhood design (e.g., distance to the nearest shopping center or bus stop). Furthermore, the CBD/urban versus rural/suburban distinction (based on a threshold density of 3 person-equivalents per acre) is quite arbitrary, and more interesting land use distinctions can be posited (e.g., transit-friendly zones versus all others).

Of course, one very desirable extension of Heckman's methodology is the option of multiple treatments, via a multinomial model for location choice and/or neighborhood attributes. This would allow for more continuity and variety in built environment conditions. Of course, computation of treatment effects under such a setting will be more difficult, but may remain feasible. Another potentially useful extension is to consider discrete models of latent segmentation for consistent treatment of discrete outcomes (e.g., vehicle ownership decisions, rather than, say, VMT).

Finally, it should be mentioned that this model's estimation does rely on assumptions of normality in the three key equations error components, and may not be robust to departures from this distributional assumption (19). More robust approaches tend to be non-parametric in nature, and therefore more complicated. In addition, a single percentage result (e.g., 58% or 90% of VMT differences being attributable to one's location) is not highly robust to model specification. More finely defined location types and more appropriate control variables than those offered in the Austin Travel Survey would be helpful in disentangling the role of home location. Regardless, Heckman's past methodological contributions take transportation planning and design in a new and rigorous direction. The present work offers transportation analysts a sense of the magnitude of the self-selection issue, at least for the important case of central versus non-central locations.

REFERENCES

1. Cervero, R., and K. Kockelman. Travel Demand and the Three Ds: Density, Diversity, and Design. *Transportation Research Part D: Transport and Environment*, Vol. 2, No. 2, 1997, pp. 199–219.
2. Ewing, R., and R. Cervero. Travel and the Built Environment: A Synthesis. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1780, 2001, pp.187–114.
3. Cervero, R., and R. Gorham. Commuting in Transit versus Automobile Neighborhoods. *Journal of the American Planning Association*, Vol. 61, No. 2, 1995, pp. 210–226.
4. Rutherford, G.S., E. McCormack, and M. Wilkinson. Travel Impacts of Urban Form: Implications from an Analysis of Two Seattle Area Travel Diaries. Presented at the Travel Model Improvement Program Conference on Urban Design, Telecommuting, and Travel Behavior. Washington, D.C., 1996.
5. Khattak, A.J., and D. Rodriguez. Travel Behavior in Neo-Traditional Neighborhood Developments: a Case Study in USA. *Transportation Research Part A*, Vol. 39, 2005, pp. 481–500.
6. Shay, E., and A.J. Khattak. Automobile Ownership and Use in Neotraditional and Conventional Neighborhoods. *Transportation Research Record: Journal of the Transportation Research Board*, No. 1902, 2005, pp. 18–25.
7. Crane, R., and R. Crepeau. Does Neighborhood Design Influence Travel? a Behavioral Analysis of Travel Diary and GIS Data. *Transportation Research D*, Vol. 3, No. 4, 1998, pp. 225–238.
8. Salon, D. Cars and the City: an Investigation of Transportation and Residential Location Choices in New York City. Dissertation for Doctorate in Agricultural and Resource Economics, The University of California at Davis, 2006.
9. Krizek, K.J. Residential Relocation and Changes in Urban Travel: Does Neighborhood-Scale Urban Form Matter? *Journal of the American Planning Association*, Vol. 69, No. 3, 2003, pp. 265–281.
10. Mokhtarian, P.L., and X. Cao. Examining the Impacts of Residential Self-Selection on Travel Behavior: A Focus on Methodologies. *Transportation Research Part B*, Vol. 42, 2008, pp. 204–228.
11. Kitamura, R., P. L. Mokhtarian, and L. Laidet. A microanalysis of Land Use and Travel in Five Neighborhoods in the San Francisco Bay Area. *Transportation*, Vol. 24, 1997, pp. 125–158.
12. Bagley, M.N., and P. L. Mokhtarian. The Impact of Residential Neighborhood Type on Travel Behavior: A Structural Equations Modeling Approach. *Annals of Regional Science*, Vol. 36, No. 2, 2002, pp. 279–297.
13. Boarnet, M.G., and S. Sarmiento. Can Land-Use Policy Really Affect Travel Behavior? a Study of the Link between Non-Work Travel and Land-Use Characteristics. *Urban Studies*, Vol. 35, No. 7, 1998, pp. 1155–1169.

14. Greenwald, M.J. The Road Less Traveled: New Urbanist Inducements to Travel Mode Substitution for Nonwork Trips. *Journal of Planning Education and Research*, Vol. 23, 2003, pp. 39–57.
15. Bhat, C.R., and J.Y. Guo. A Comprehensive Analysis of Built Environment Characteristics on Household Residential Choice and Auto Ownership Levels. *Transportation Research Part B*, Vol. 41, No. 5, 2007, pp. 506–526.
16. Heckman, J.J. The Common Structure of Statistical Models of Truncation, Sample Selection, and Limited Dependent Variables and A Simple Estimator for Much Models. *Annals of Economic and Social Measurement*, Vol. 5, 1976, pp. 475–492.
17. Heckman, J.J. Sample Selection Bias as a Specification Error. *Econometrica*, Vol. 47, No. 1, 1979, pp. 153–162.
18. Heckman, J.J., and E.J. Vytlaki. Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects. *Proceedings of the National Academy of Sciences*, Vol. 96, 1999, pp. 4730–4734.
19. Heckman, J.J., J.L. Tobias, and E.J. Vytlaki. Four Parameters of Interest in the Evaluation of Social Programs. *Southern Economic Journal*, Vol. 68, No. 2, 2001, pp. 210–223.
20. Heckman, J.J., and E.J. Vytlaki. Structural Equations, Treatment Effects, and Econometric Policy Evaluation. *Econometrica*, Vol. 73, No. 3, 2005, pp. 669–738.

LIST OF TABLES AND FIGURES

TABLE 1 Summary Statistics for Variables Used in the Location Choice Model

TABLE 2 Probit Model Results for Residential Location Choice

TABLE 3 Summary Statistics for Variables Used in the Models of Household VMT

TABLE 4 Model Results for Household Daily VMT

FIGURE 1 Classification of Austin's Traffic Analysis Zones

TABLE 1 Summary Statistics for Variables Used in the Location Choice Model

| Explanatory Variables | Minimum | Maximum | Mean | Std. Deviation |
|---|----------------|----------------|-------------|-----------------------|
| Household size (all persons) | 1 | 10 | 2.61 | 1.33 |
| Annual income (in \$1,000, for year 1996) | 5 | 150 | 44.93 | 30.97 |
| Age of household head | 18 | 99 | 42.34 | 15.76 |
| N _{obs} | 1,903 | | | |

TABLE 2 Probit Model Results for Residential Location Choice

| Explanatory Variables | Coefficient | t-statistics |
|------------------------------|--------------------|---------------------|
| Constant | -1.08 | -9.27 |
| Household size | 0.286 | 11.49 |
| Annual income (in \$1,000) | 0.00242 | 2.42 |
| Age of household head | 0.0113 | 5.88 |
| N _{obs} | 1,903 | |
| Log Likelihood | | |
| Market Share | -1286.5 | |
| Convergence | -1196.2 | |
| LRI | 0.0702 | |

Note: Residential locations in CBD or urban zones serve as the base.

TABLE 3 Summary Statistics for Variables Used in the Models of Household VMT

| Explanatory Variables | Minimum | Maximum | Mean | Std. Deviation |
|---|----------------|----------------|-------------|-----------------------|
| Living in Rural or Suburban Areas (Treated Sample) | | | | |
| Number of workers | 0 | 5 | 1.51 | 0.86 |
| Number of vehicles | 1 | 8 | 2.05 | 0.87 |
| Delivery driver(s) | 0 | 1 | 0.0550 | 0.228 |
| Population density (1,000/square mile) | 0 | 5.63 | 1.91 | 1.71 |
| Job density (1,000/square mile) | 0 | 11.19 | 0.403 | 0.776 |
| Household daily VMT | 0 | 962 | 71.03 | 88.14 |
| N _{obs} | 1,127 | | | |
| Living in CBD or Urban Areas (Untreated Sample) | | | | |
| Number of workers | 0 | 4 | 1.31 | 0.838 |
| Number of vehicles | 1 | 5 | 1.73 | 0.798 |
| Delivery driver(s) | 0 | 1 | 0.0374 | 0.190 |
| Population density (1,000/square mile) | 0 | 25.09 | 6.32 | 3.48 |
| Job density (1,000/square mile) | 0.136 | 164.8 | 3.85 | 8.26 |
| Household daily VMT | 0 | 981 | 47.04 | 81.55 |
| N _{obs} | 776 | | | |

TABLE 4 Model Results for Household Daily VMT (OLS)

| Explanatory Variables | Living in Rural or Suburban (Treated) | | Living in CBD or Urban (Untreated) | |
|---|--|--------------|---------------------------------------|--------------|
| | Coefficient | t-statistics | Coefficient | t-statistics |
| Constant | 30.43 | 2.41 | -16.30 | -1.23 |
| Number of workers | 13.55 | 4.08 | 8.97 | 2.30 |
| Number of vehicles | 16.75 | 5.05 | 16.03 | 3.81 |
| Delivery driver(s) | 18.89 | 1.70 | 26.87 | 1.77 |
| Population density (1,000/square mile) | -6.57 | -4.39 | 0.50 | 0.60 |
| Job density (1,000/square mile) | -5.23 | -1.60 | 0.90 | 2.60 |
| Selection-correction | -1.02 | -0.08 | -18.40 | -1.50 |
| Number of Observations | 1127 | | 776 | |
| R ² | 0.097 | | 0.072 | |



Note: All zones contain at least one sampled household, except for those labeled as “Study Area” zones, which complete the three-county region.

FIGURE 1 Classification of Austin’s traffic analysis zones.