The Journal of Transport and Land Use http://jtlu.org
Vol. 9, No. 1 [2016] pp. 9–28

JTLU

Quantifying residential self-selection effects: A review of methods and findings from applications of propensity score and sample selection approaches

Patricia L. Mokhtarian
Georgia Institute of Technology
patmokh@gatech.edu

David van Herick
University of California, Davis
dmvanherick@ucdavis.edu

Abstract: The phenomenon whereby individuals self-select into their residential environment based on previously determined preferences for how to travel is known as residential self-selection (RSS). Numerous studies have investigated the influence of RSS on the estimated effect of the built environment on travel behavior. However, surprisingly few have actually quantified its effect in terms of partitioning the total influence of the built environment (BE) on travel behavior into a component attributable to RSS and one attributable to the built environment itself. This paper reviews 10 analyses (found in seven studies) that have quantified the proportion of the total influence of the built environment that is due to the BE itself (which we call the BEP), using either propensity-score or sample-selection approaches to control for RSS. After first outlining the basics of each approach, we then explain the various methods used to compute the BEP, followed by a discussion of the empirical results. The estimated BEPs vary widely, ranging from 34 percent to 98 percent. A number of reasons for these disparities are suggested, but there is considerable divergence in estimates even when many of these factors are held constant. Additional research is called for to better understand the circumstances under which the BEP is higher or lower.

Keywords: Residential self-selection, propensity score, sample selection, travel behavior

Article history:
Received: October 28, 2014
Accepted: December 19, 2014
Available online: April 29, 2016

1 Introduction

The influence of the built environment (BE) on travel behavior (TB) is of considerable interest to transportation policy and planning, particularly since it is the most obvious limiting factor on whether individuals even have opportunities to make certain decisions with respect to their behavior. Without tracks, there is no rail transportation; without roads, there is no driving; and increased incidence of

http://dx.doi.org/10.5198/jtlu.2016.788
ISSN: 1938-7849 | Licensed under the Creative Commons Attribution – Noncommercial License 3.0

The *Journal of Transport and Land Use* is the official journal of the World Society for Transport and Land Use (WSTLUR) and is published and sponsored by the University of Minnesota Center for Transportation Studies. This paper and others from WSTLUR 2014 are also published with sponsorship from WSTLUR and the Institutes of Transportation Studies at University of California, Davis and University of California, Berkeley.

aesthetic walking paths leading to interesting destinations will surely encourage, if not lead to, more walking in an area. It remains an open question, however, how much the built environment itself brings about a certain kind of behavior versus how much one's proclivities for a certain kind of behavior influence one's choice of built environment. To the extent the latter is true, policies that attempt to induce more sustainable travel behavior by shaping the built environment could fall short of expectations if large numbers of people end up living in sustainable built environments (perhaps due to financial incentives or a sizable increase in the supply of such environments) without having the proclivity to travel more sustainably.

The phenomenon whereby individuals self-select into their residential environment based on previously determined preferences for how to travel is known as residential self-selection (RSS). In the past two decades, numerous studies have analyzed the influence of the built environment on travel behavior after controlling for self-selection. In the process, a variety of approaches have been employed: direct questioning, statistical controls, instrumental variables models, sample-selection models, propensity-score models, joint discrete-choice models, structural-equations models, mutually-dependent discrete-choice models, and longitudinal designs (Schwanen and Mokhtarian 2005; Chatman 2005; Cao, Mokhtarian, and Handy 2011; Stevens and Brown 2011).

The operationalization of these approaches has been diverse. In particular, the majority of studies simply provide a qualitative indication that both the built environment and self-selection matter, or that one of the two factors appears to be more important than the other. Several studies (e.g., Joh, Mai Thi, and Boarnet 2012; Larco et al. 2012) have found that only the built environment is important, but none has found that only self-selection matters. A much smaller number of researchers has actually quantified the shares of the total apparent influence of the built environment that are respectively attributable to the true influence versus to attitudinal predispositions, and they have done so in different ways.

Most of the handful of studies that quantify the role of RSS have been conducted since the review performed by Cao et al. (2011). Accordingly, it is worthwhile to assemble and examine those studies (both newer and older) to see what can be learned from them collectively. In particular, we provide an analysis both of the methods used to quantify the role of self-selection and the empirical results obtained. To keep the scope manageable, we limit ourselves to the two approaches that (so far) have, in our judgment, most commonly quantified the role of residential self-selection: propensity-score and sample-selection models. We also limit the discussion to situations where the travel behavior outcome of interest is treated as continuous-valued (and ratio-scaled).

The remainder of this paper is organized as follows. In the next section, we give a brief overview of the two approaches for dealing with RSS, and the corresponding methods used to quantify its role. In the third section we review seven recent studies (involving 10 models) that quantify the relative influences of the built environment and residential self-selection on travel behavior. The final section discusses the range of outcomes exhibited by the studies reviewed, including likely reasons for the differences, and calls for additional research to further investigate those reasons.

2 Brief overview of methodologies

In the following subsections, we first provide short descriptions of the propensity-score and sample-selection approaches for dealing with RSS. The final subsection presents the methods found in the literature for computing the proportion of the total effect of the BE on TB that is due to the BE itself (which we will refer to as the "built environment proportion" or BEP) as opposed to the proportion due to RSS (namely 1–BEP) for the propensity-score and sample-selection approaches.

2.1 Propensity scores

At the heart of this approach is the estimation of a *propensity score* for each case, which in our context is the probability of living in an urban neighborhood (the treatment condition), obtained from a binary discrete choice model of residential location. The propensity scores can be used in three (non-mutually-exclusive) ways.[1] The first is *Regression (PSR):* they can be entered into the TB model as a control variable, similar to entering attitudes in the statistical controls approach, but with the difference that the propensity score (a) can combine multiple attitudes (and other variables) into a single composite value, and (b) focuses on explaining the *propensity to live in a given environment,* not on explaining the *TB outcome itself.* The second is *Matching (PSM):* respective residential choice groups can be matched on their propensity score. For example, each urban resident can be matched to the "most similar" suburban resident (i.e., the one with the closest propensity score) until members of either group cannot be sufficiently matched, at which point the remaining cases are discarded. Then, the difference in TB for each matched pair is averaged over all pairs, and compared to the difference in average TB for all urban residents (matched or unmatched) versus all suburban residents. The third is *Stratification (PSS):* individuals can be stratified into a number of bins defined by specific ranges of the propensity scores (based either on the value of the score, e.g., subdividing the range from 0 to 1 into four or five intervals of equal width, or on the sample sizes within each stratum, e.g., subdividing the sample into four or five groups of roughly equal size based on the propensity score). Then, within each bin, differences in TB between the treatment and control group members can be compared, and a size-weighted average of those differences can be compared to the difference in average TB for all urban residents (matched or unmatched) versus all suburban residents.

Propensity-score matching and stratification employ related ideas; the purpose of each is to mimic a quasi-random experiment. By pairing or grouping cases with similar propensity scores, we can assume that the propensity to live in a given type of environment is similar for each person in such a pair or stratum, and that the "assignment" to a particular type of environment is random, given that propensity. Because the propensity to live in the treatment (or control) environment is similar for both cases, but their actual choices are different, any difference between the two groups with respect to the outcome variable of interest (i.e., travel behavior) is putatively capturing the "true" difference and not a difference that arises because of self-selection into the treatment or control environment. Compared to matching, stratification has the advantage that virtually the entire sample can be used, and the corresponding disadvantage that the propensities within a given stratum may be too disparate for the assumption of equivalence to hold.

In application, it is common to go beyond merely grouping cases on the basis of similar *propensity scores;* the ultimate goal is to correct for the *sources* of self-selection, namely the fact that the control and treatment groups are not initially equivalent, or "balanced," on the *key covariates* that influence the propensities to live in one type of location or another. Accordingly, some form of means comparison (such as paired samples t-tests or standard differences; D'Agostino 1998) is typically used on the covariates of the propensity-score equation for propensity-score matching or stratification to test (a) whether selection is taking place before grouping on propensity and (b) whether selection has been controlled for after grouping.

For both matching and stratification, the "before" difference on covariates is simply the difference (as indicated by t-tests or standard differences) between the full set of individuals in the treatment condition and the full set of individuals in the control condition with respect to the covariates of the propensity-score equation. For matching, the "after" difference is the difference between the set of matched treatment individuals and the set of matched control individuals with respect to the same covariates. A

[1] A fourth way, less often used and not further treated here, is to use the propensity score to weight cases according to the inverse probability of treatment. See Austin (2011) for a discussion of all four methods.

typical rule of thumb in epidemiology is that if the standard difference for a covariate is less than 10 percent, the treatment and control groups are considered balanced on that covariate (Oakes and Johnson 2006). With respect to stratification, "after" differences may be analyzed by using a two-way analysis of variance (Rosenbaum and Rubin 1984), with two treatments by s strata and an F-statistic for the main effect of neighborhood type after adjusting for propensity score quantiles and an F-statistic for the interaction effect between neighborhood type and propensity-score quantile, both with respect to each of the covariates. If after matching or stratification there is little or no significant difference between urban and suburban residents (in the same pair or stratum) on all covariates, then self-selection is considered to be controlled for. Of course, this will only be true to the extent that all relevant covariates have been measured and included, an important point to which we return later.

2.2 Sample selection

Sample selection in this context refers to a model variously known as an endogenous switching regression, a "mover/stayer" or Roy model, or a two-outcome version of the standard "Heckit" model (Heckman 1979; Heckman, Tobias, and Vytlacil 2001). The standard sample selection approach requires that the selection equation be a binary probit model (which could have the same explanatory variables as the propensity score equation), and the outcome equation for travel behavior is allowed to vary by residential choice. Specifically, we have:

$RC_i^* = \alpha' \mathbf{w}_i + u_i, \quad RC_i = 1 \text{ if } RC_i^* \geq 0, \quad RC_i = 0 \text{ if } RC_i^* < 0$ (the selection equation);
$Y_i = Y_{i1} = \boldsymbol{\beta}_1' \mathbf{x}_i + \varepsilon_{i1}$ when $RC_i = 1$ (the outcome equation under the treatment condition);
$Y_i = Y_{i0} = \boldsymbol{\beta}_0' \mathbf{x}_i + \varepsilon_{i0}$ when $RC_i = 0$ (the outcome equation under the control condition);

where

RC_i^* is the latent continuous utility or propensity that person i has for living in an urban neighborhood;

RC_i is the observed residential location choice, = 0 if person i lives in a suburban neighborhood (meaning that $RC_i^* < 0$) and = 1 if in an urban neighborhood ($RC_i^* \geq 0$);

α is the vector of selection equation coefficients, w_i is a vector of observed covariate values, and u_i is the net impact on RC_i^* of unobserved characteristics for individual i;

Y_i is the travel behavior outcome for person i; and

β_{RC_i} is the vector of outcome equation coefficients, x_i is a vector of covariate values, and ε_{iRC_i} is the net impact on Y_i of unobserved characteristics, for individual i living in neighborhood type RC_i. It is customary to improve identifiability by including at least one variable in the selection equation that does not appear in the outcome equations (Winship and Morgan 1999; Cameron and Trivedi 2005).

The error terms of the three equations are assumed to be trivariate normally-distributed, as follows (Greene 2007, Section E32.3.1):

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i0} \\ u_i \end{pmatrix} \sim N_3 \left[\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 & \rho_1\sigma_1 \\ 0 & \sigma_0^2 & \rho_0\sigma_0 \\ \rho_1\sigma_1 & \rho_0\sigma_0 & 1 \end{pmatrix} \right],$$

where σ_1^2 and σ_0^2 are the variances of ε_{i1} and ε_{i0}, respectively, Var(u_i) is normalized to 1 for identifiability, ρ_1 = Corr(ε_{i1}, u_i) and ρ_0 = Corr(ε_{i0}, u_i). The zero covariance between the error terms of the two outcome equations reflects the fact that individuals are only measured in one of the two (treatment or control) states, not both, so if errors are uncorrelated across individuals (per the usual assumption for random samples), they will be uncorrelated between mutually exclusive groups of individuals.

The mover-stayer model is different from an ordinary market segmentation model in that: (1) the selection model is often (though not necessarily) estimated simultaneously with the outcome equations using full information maximum likelihood, and more importantly, (2) in estimation, each outcome equation incorporates a "selection correction" factor that involves the probabilities from the selection equation. This factor should ideally correct the selectivity bias (i.e., any non-zero mean) of the error term of the outcome equation, and under certain relatively general assumptions, in the case of a binary probit selection equation, the inverse Mills ratio (IMR) is the appropriate factor to do so (Winship and Morgan 1999). The IMR is the ratio of the standard normal probability density function ϕ to the standard normal cumulative distribution function Φ, where the arguments of both functions are the observed utility of the selection equation (or its negative). The appropriate IMR differs for each outcome equation: $\frac{\phi(\alpha'\mathbf{w}_i)}{\Phi(\alpha'\mathbf{w}_i)}$ when $RC_i = 1$ and $\frac{-\phi(\alpha'\mathbf{w}_i)}{\Phi(-\alpha'\mathbf{w}_i)}$ when $RC_i = 0$. If these quantities are inserted into their respective outcome equations with coefficients $\rho_1\sigma_1 = \lambda_1$ and $\rho_0\sigma_0 = \lambda_0$, respectively, ordinary least squares or maximum likelihood estimation of the resulting equations will provide consistent estimates of β_1 and β_0, as well as λ_1 and λ_0.

In the two-step estimation method, the selection model is estimated in the first step, the resulting $\hat{\alpha}$ is used to estimate the IMR terms, and the two outcome equations (containing the $\hat{IMR}$s) are estimated in the second step. In this case (reflecting the increased uncertainty from using $\hat{IMR}$ rather than the true IMR), the asymptotic covariance matrix of the estimators of the coefficients of the outcome equations needs to be corrected to obtain consistency of the standard errors of the parameter estimates. The usual least squares estimate of the variance of the error term of each outcome equation needs to be corrected as well. In the case of (one-step) full-information maximum likelihood (FIML) estimation, consistent estimates are obtained without the need for correction.

A key difference between the PS and SS methods is that propensity scores do not correct for the bias resulting from correlation between unobservables of the propensity-score equation and unobservables of the travel-behavior equation, but only for correlation between *observables* of the propensity-score equation and *unobservables* of the travel-behavior equation (Winship and Morgan 1999, p. 679). By contrast, the sample selection approach not only allows the travel-behavior equations to differ by type of residential location, it also allows correlation between unobservables of the selection equation (analogous to the propensity-score equation) and unobservables of the travel behavior equations (Winship and Morgan1999; Heckman, Tobias, and Vytlacil 2001; Cameron and Trivedi 2005, Section 16.5.7 and Chapter 25).

2.3 Computing the Built Environment Proportion (BEP)

By construction the BEP is a fraction, which will ordinarily (but not necessarily) fall between zero and one, inclusive.[2] The denominator constitutes the *total effect* of the built environment. It represents the BE effect on TB observed when RSS is not controlled for. The numerator constitutes the *true effect* of the BE on TB, representing the BE effect observed when RSS is controlled for. It is of interest to note

[2] Let us write BEP = $\frac{BEeff}{BEeff + ATeff}$ for simplicity, where the two constituents respectively represent the true effect of the built environment and the effect of attitudinal predispositions on travel behavior. In general, we would expect *BEeff* and *ATeff* to have

that although the concept of the BEP seems to be a natural one (referred to, albeit not given a name, in Mokhtarian and Cao 2008), none of the travel behavior studies reviewed in this paper (all of which used it, though again without calling it as such) cited a precedent for it in the general treatment effects/sample selection literature, and the present authors could not find such a precedent.[3]

In the following discussion, several concepts are of importance:

- The ***average treatment effect (ATE)*** is the expected difference in outcomes (after versus before treatment) across the whole population, i.e., the average change in TB if a randomly selected person moved from a suburban neighborhood to an urban one.[4]
- The ***average treatment effect on the treated (TT)*** is the expected outcome (TB) for those who receive the treatment (in our context, for those who now live in an urban location), relative to what the outcome would have been if they did not receive the treatment (i.e., if they were to live in a suburban location).
- The ***average treatment effect on the untreated (TUT)*** is the expected difference in outcomes for those who now live in a suburban location, if they were to move to an urban location.

In general, these effects are not necessarily equal (Heckman, Tobias, and Vytlacil 2001), and it is not always clear which one is of the greatest interest. For example, as Ho et al. (2007, p. 204) point out, "Medical studies typically use the [TT] as the designated quantity of interest because they often only care about the causal effect of drugs for patients that receive or would receive the drugs." In our context, we could legitimately be interested in TT, as expressing the change in travel behavior that could be expected from specific new developments, *for the residents of those developments.* But if we wanted to project those results to a situation in which many such new developments were to be built, such that

---

the same sign, even if an individual is attitudinally mismatched with respect to his or her residential location. For example, all else equal, we expect those who *prefer* to live in an urban environment to drive less than those who do not, and we also expect those who actually *do* live in an urban environment to drive less than those who do not. It can be seen that BEP will lie outside the interval [0, 1] if and only if ATeff is opposite in sign to BEeff, in which case BEP > 1 if *ATeff* is smaller in magnitude than *BEeff*, and BEP < 0 if the converse is true. Having opposite signs would signify an extreme case of residential mismatch (e.g., in which people living in an urban environment actually drive *more* than those in suburbia, despite *wanting* to drive *less*), which is unlikely to occur as an average outcome across a large group. However, if either *BEeff* or *ATeff* is near 0 in reality, its *estimate* may lie just on the "other side" of 0 (compared to the sign of the larger effect) through random variation, which could lead to an estimated BEP slightly outside [0, 1].

[3] The evaluation literature includes a number of comparisons (e.g., Shadish, Clark, and Steiner 2008; Cook, Shadish, and Wong 2008) of treatment effects (TE) estimated from randomized experiments (RE, considered the gold standard), TE_{RE}, with those estimated from observational studies (OS), TE_{OS} – most rigorously using the same sample for both measurements. In such cases it is common to report *bias reductions* for various methods that control for selection bias. The absolute bias is computed as the absolute difference between the answer obtained from the randomized experiment and the one obtained from the observational study *without* controlling for selection bias ($TE_{OSwo\ control}$), namely $|TE_{OSwo\ control} - TE_{RE}|$. The answer obtained from the observational study *with* controls for selection bias ($TE_{OSw\ control}$) is presumably closer to the truth (TE_{RE}) than is $TE_{OSwo\ control}$, and so $\frac{|TE_{OSw\ control} - TE_{RE}|}{|TE_{OSwo\ control} - TE_{RE}|}$ is generally a quantity less than 1, denoting the absolute bias remaining after controls, expressed as a fraction of the amount in place before controls. The quantity $\left(1 - \frac{|TE_{OSw\ control} - TE_{RE}|}{|TE_{OSwo\ control} - TE_{RE}|}\right) \times 100\%$ is the *bias reduction*. The measure $\frac{|TE_{OSw\ control} - TE_{RE}|}{|TE_{OSwo\ control} - TE_{RE}|}$ is not unlike our concept of the BEP, for which the denominator represents the "uncorrected" effect and the numerator the "corrected" effect. In our case, however, a TE_{RE} is not available, and so our BEP corresponds to $\frac{TE_{OSw\ control}}{TE_{OSwo\ control}}$. Among other benefits, this literature is valuable in reminding us that $TE_{OSw\ control}$ is only *closer* to the truth (as estimated by TE_{RE}) than is $TE_{OSwo\ control}$, not *equal* to it. In general, the controls for RSS are going to be imperfect, for *any* method. Of course, it is also important to remember that even TE_{RE} can only be *estimated,* not known perfectly.

[4] Strictly speaking, there are two different averages involved (Cameron and Trivedi 2005). The average treatment effect (ATE) for a person with observed characteristics *x*, ATE(*x*), is an average over the distribution of unobserved characteristics. ATE, by contrast, is the average of ATE(*x*) over the distribution of x (observed characteristics). Technically, this might more clearly be labeled the average ATE(*x*), or AATE, but in keeping with convention, we will use "ATE" to refer to this double average.

many people would end up living there without deliberately self-selecting into them, then we would expect the TB impacts for those people to look like TUT, and ATE could be a more appropriate indicator of the overall effect.

To complicate matters further, the literature does not always clearly distinguish between ATE and TT (Oakes and Johnson 2006).[5] In our context, we consider it appropriate to focus on ATE, since the effectiveness of land-use policy as a tool for reducing vehicle travel depends on its large-scale application. However, PSM methods typically focus on TT,[6] whereas PSS and SS methods readily produce ATE as well as TT (and TUT) (Cameron and Trivedi 2005, pp. 872–875; Tucker 2010).

Accordingly, we can say that for both propensity-score and sample-selection models, the numerator of the BEP is a *treatment effect* for which self-selection has been controlled, but in application that effect has been computed as ATE for PSS and SS methods, and generally as TT for PSM methods. The denominator is again computed differently for each method; it consists of a quantity called the *observed influence* (obs. inf.) for propensity-score methods, and has two different formulations (presented in Section 2.3.2) in the sample-selection context.

Formally, ATE can be defined as $E[\ Y_{i1} - Y_{i0}\ ]$, where Y_{i1} is the (TB) outcome for person *i* when treated (living in an urban neighborhood), Y_{i0} is the outcome when *i* is untreated (living in a suburban neighborhood), and E is the expectation operator (after Winship and Morgan 1999). TT can be defined as $E[\ Y_{i1} - Y_{i0} \mid i \in T\ ]$, where T denotes the treatment group.

However, these conceptualizations inherently assume that we have information on all individuals for both the environment in which they actually do live *and* the environment in which they do not, but theoretically could, live.[7] Since it is rarely the case that data is available for both, the challenge is to find a way to estimate ATE and TT that is not subject to a self-selection bias.

In the subsections below, we present the computation of the numerator and denominator of the BEP for the propensity-score and sample-selection approaches, respectively, as it has been operationalized in the literature to date.

2.3.1 Propensity score methods

The simplest empirical measure of an effect is just the difference in means between the treatment and control groups: $\bar{Y}_{i\in T} - \bar{Y}_{i\in C}$, where T designates the treatment group and C the control group, Y_i is the TB outcome for person *i*, $\bar{Y}_{i\in T} = \frac{\sum_{i\in T} Y_i}{N_T}$, $\bar{Y}_{i\in C} = \frac{\sum_{i\in C} Y_i}{N_C}$, and N_T and N_C are the numbers of cases in T and C, respectively.

But if self-selection is not controlled for, this measure confounds the true effect of being treated

[5] For example, D'Agostino 1998, p. 2266, refers loosely to "treatment effects" and "average treatment effect at that propensity score." A key article by Becker and Ichino (2002) is titled "Estimation of average treatment effects…" but only discusses the average treatment effect on the treated. And Greene (2007 Section 32.2) occasionally uses "average treatment effect" as shorthand for what the context ostensibly makes clear is actually "average treatment effect on the treated." Other scholars, such as Cameron and Trivedi (2005), are careful to specify whether a given formula estimates TT or ATE.

[6] Conceptually, this is natural, since (as ordinarily applied), the entire purpose of the matching approach is to provide a proxy measure of the counterfactual *specifically for the treated cases.*

[7] In our opinion (experience), considerable confusion for the newcomer to this literature can arise from the failure to distinguish *group membership* from *treatment condition,* two dimensions that are conceptually (and therefore notationally) distinct but completely confounded in the typical observational study. It is important to understand that conceptually, in terms of group membership, the population is divided into two categories, here labeled "T" and "C," *separately from whether a given group member is actually treated or not.* Similarly, in terms of treatment condition a person can live in either of two types of neighborhood, RC = 1 (urban) and RC = 0 (suburban), *regardless of which group membership label he or she bears.* We only observe T members having RC = 1 and C members having RC = 0, but behind both the PS and the SS methods is the concept of the counterfactual, in which we imagine (and try to account for) what would happen if T members lived in suburbs (RC = 0) and/or C members lived in urban areas.

(moving from a suburban to an urban area) with effects that are due to self-selection. Thus, when self-selection is not controlled for, $\bar{Y}_{i\in T} - \bar{Y}_{i\in C}$ is the total *observed influence* of the BE, i.e., the denominator of the BEP. Various estimates of $\bar{Y}_{i\in T} - \bar{Y}_{i\in C}$ when self-selection is controlled for constitute alternative ways of computing the numerator of the BEP, the treatment effect.

Specifically, for PSM, the treatment effect (in this case, usually TT) is the difference in means of the outcome behavior of interest between the *matched* treatment and control residents (D'Agostino 1998; Cameron and Trivedi 2005); since they are matched to have similar *propensities* to live in a given area, any remaining differences in TB are presumed to be true influences of the BE in which they actually *live.* On the other hand, the denominator of the BEP, i.e., the observed influence, is the difference in means of the *unmatched* treatment and control residents; when pooled without regard to propensity, observed differences in TB are a conflation of both true BE influences and influences of predispositions to live in a certain type of BE.

For PSS (Imbens 2004), the ATE and TT are each computed as (different) weighted averages of the differences in means of the outcome variable for the stratified groups. In both cases, within each stratum, separate averages are taken for the treatment and control cases, and the difference in the averages is calculated. Then those differences are averaged across strata: for the ATE, the differences are weighted by the share of individuals in the overall sample who fall within each stratum (without regard to treatment status), and for the TT they are weighted by the share of the sample of *treated* individuals that falls within the stratum. The observed influence for PSS is the same as for PSM: the actual difference between the treatment and control groups, without matching or stratification.

Not all individuals will be living in their preferred environment. Consider the extreme case where individuals are all "consonant" (Schwanen and Mokhtarian 2005) with their preferences. In this case, the average TB difference between urban and suburban residents will tend to be the most pronounced. At the opposite extreme, where all individuals are "dissonant," then the behavior of individuals will likely be much closer to one another, since each group may behave in part as it normally would based on preferences, but it would be affected by the environment in the opposite manner. For an experiment in which people could be randomly assigned to residential neighborhood type, the TB difference would lie somewhere between these two extremes (Cao 2010). However, the actual observed influence will likely not be the same as for a random experiment, due to some self-selection taking place. Through matching or stratification, the goal is to control for that difference between the observed influence and what it would be for a random experiment.

We were unable to find any studies in which a BEP was calculated for a propensity-score regression (PSR) model; for this paper we limit the discussion to methods that have been applied.

2.3.2 Sample selection model

For the sample selection model, authors often consider the ATE, estimated as (Heckman, Tobias, and Vytlacil 2001): $\hat{ATE} = \hat{\boldsymbol{\beta}}_1'\bar{\mathbf{x}} - \hat{\boldsymbol{\beta}}_0'\bar{\mathbf{x}} = \left(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0\right)'\bar{\mathbf{x}}$, where $\hat{\boldsymbol{\beta}}_1$ and $\hat{\boldsymbol{\beta}}_0$ are the vectors of estimated coefficients for the outcome models respectively associated with the treatment (urban) and control (suburban) conditions, and $\bar{\mathbf{x}}$ is the vector of sample means of the ordinary explanatory variables (not including the IMR term defined in Section 2.2, although the IMR term is included when *estimating the coefficients of the outcome equations*, thereby allowing the βs to be consistently estimated). Note that this formulation requires both sets of covariates for the two outcome equations to be the same, so the superset of all variables significant to either equation is included in both outcome specifications. The ATE is the expected change in travel behavior when moving a randomly selected individual from a suburban to an urban environment. Even though we still do not observe an individual in both environments, the genius of this method is that, by controlling for RSS through the inclusion of the IMR term in each

outcome equation, the consistently estimated β_1 and β_0 coefficients properly reflect the expected influence that the **x** variables *would* have if an individual with such characteristics were placed in each of the associated environments.

As mentioned, the ATE is the numerator for the BEP calculation: the true effect of the BE. For the denominator—the total effect of the BE—two quantities are of interest: the average treatment effect on the treated (TT), and the average treatment effect on the untreated (TUT). It can be shown that the treatment effect on the treated, for a *treated* person *i* with characteristics $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$, is $TT(x_i,w_i) = (\beta_1 - \beta_0)'x_i + [(\rho_1\sigma_1) - (\rho_0\sigma_0)]\left(\frac{\phi(\boldsymbol{\alpha}'\mathbf{w}_i)}{\Phi(\boldsymbol{\alpha}'\mathbf{w}_i)}\right)$, and the average treatment effect on the treated, TT, is estimated by replacing the unknown parameters in the above quantity with their estimated values, and averaging over the sample of treated persons (Heckman, Tobias, and Vytlacil 2001)[8]: $\hat{TT} = \frac{1}{N_1}\sum_{RC_i=1}\{(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)'\mathbf{x}_i + (\hat{\lambda}_1 - \hat{\lambda}_0)\left(\frac{\phi(\hat{\boldsymbol{\alpha}}'\mathbf{w}_i)}{\Phi(\hat{\boldsymbol{\alpha}}'\mathbf{w}_i)}\right)\}$, where N_1 is the number of cases for which $RC = 1$.

The TT is the expected outcome gain from the treatment for *individuals that select the treatment option.* In the case of residential choice, it represents the expected change in the travel behavior of individuals *who have moved from a suburban to an urban residential location* (i.e., who have been treated). Note that the TT for a given treated case *i* can be considered to represent the total effect of the BE for such cases, which can be decomposed into the component constituting the true effect of the BE ($(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)'\mathbf{x}_i$, i.e., the ATE for such cases), and the component reflecting the influence of the propensity of the treated cases (urban dwellers) to live in an urban environment ($[(\rho_1\sigma_1) - (\rho_0\sigma_0)]\left(\frac{\phi(\boldsymbol{\alpha}'\mathbf{w}_i)}{\Phi(\boldsymbol{\alpha}'\mathbf{w}_i)}\right)$, i.e., the portion of the total effect that is due to self-selection). $\rho_1\sigma_1$ and $\rho_0\sigma_0$ are the coefficients of the IMR terms in the RC = 1 and RC = 0 outcome equations, respectively, and $\frac{\phi(\boldsymbol{\alpha}'\mathbf{w}_i)}{\Phi(\boldsymbol{\alpha}'\mathbf{w}_i)}$ is the IMR for treatment for treated person *i*, call it IMR_{i1}, with estimated counterpart $\left(\frac{\phi(\hat{\boldsymbol{\alpha}}'\mathbf{w}_i)}{\Phi(\hat{\boldsymbol{\alpha}}'\mathbf{w}_i)}\right) = \hat{IMR}_{i1}$.

Similarly, the TUT for a given *untreated* case *i*, with characteristics $\boldsymbol{x}_i$ and $\boldsymbol{w}_i$, can be written as: $TUT(x_i,w_i) = (\beta_1 - \beta_0)'x_i + [(\rho_1\sigma_1) - (\rho_0\sigma_0)]\left(\frac{-\phi(\boldsymbol{\alpha}'\mathbf{w}_i)}{\Phi(-\boldsymbol{\alpha}'\mathbf{w}_i)}\right)$ and the average treatment effect on the untreated, TUT, is estimated by replacing the unknown parameters in the above quantity with their estimated values, and averaging over the sample of untreated persons: $\hat{TUT} = \frac{1}{N_0}\sum_{RC_i=0}\{(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)'\mathbf{x}_i + (\hat{\lambda}_1 - \hat{\lambda}_0)\left(\frac{-\phi(\hat{\boldsymbol{\alpha}}'\mathbf{w}_i)}{\Phi(-\hat{\boldsymbol{\alpha}}'\mathbf{w}_i)}\right)\}$, where N_0 is the number of cases for which $RC = 0$. The latter represents *the expected change in travel behavior of individuals who did* ***not*** *select the treatment option, if they* ***were*** *to be treated.* In our context, for a given untreated case this represents the total effect of the BE for such cases, which can be decomposed into the component constituting the true effect of the BE $(\boldsymbol{\beta}_1 - \boldsymbol{\beta}_0)'\mathbf{x}_i$, i.e., the ATE for such cases), and the component reflecting the influence of the propensity of the untreated cases (suburban dwellers) to live in an urban environment $([(\rho_1\sigma_1) - (\rho_0\sigma_0)]\left(\frac{-\phi(\boldsymbol{\alpha}'\mathbf{w}_i)}{\Phi(-\boldsymbol{\alpha}'\mathbf{w}_i)}\right)$, i.e., the portion of the total effect that is due to self-selection). $\frac{-\phi(\boldsymbol{\alpha}'\mathbf{w}_i)}{\Phi(-\boldsymbol{\alpha}'\mathbf{w}_i)}$ is the IMR for treatment for untreated person *i*, call it IMR_{i0}, with estimated counterpart $\left(\frac{-\phi(\hat{\boldsymbol{\alpha}}'\mathbf{w}_i)}{\Phi(-\hat{\boldsymbol{\alpha}}'\mathbf{w}_i)}\right) = \hat{IMR}_{i0}$.

At least two studies have used the TT as the denominator in a calculation of BEP (i.e., the BEP is calculated as ATE / TT) (Zhou and Kockelman 2008; Cao 2009). However, while also relaxing the assumption of jointly normally distributed error terms (and thus using counterparts to the IMRs that are different from the formulas presented here), Bhat and Eluru (2009) use a weighted average of the

[8] At the time of this writing, Limdep 10.0 estimates TT by evaluating the above expression at the *overall* sample means of x and w (personal communication of William Greene to the first author, Sept. 27, 2014), which differs from the approach described by Heckman, Tobias, and Vytlacil (2001) in (a) using a nonlinear function (the IMR) evaluated at the average rather than the average of the functional values (see, e.g., Train 2009, pp. 29-30, for the dangers of doing this in similar contexts), and (b) using the entire sample rather than only the treated cases to compute the quantity. It is unknown how different the two estimates are.

TT and TUT as their denominator—specifically, $\frac{1}{N}\left(N_1 \hat{TT} + N_0 \hat{TUT}\right)$, where N_0 and N_1 are as defined above, and $N = N_0 + N_1$. As will be seen in the following section, this alternative formulation substantially affects the empirical results.

The weighted average of the TT and TUT differs from, but is related to, ATE. In the standard mover-stayer model context (i.e., with jointly normal error terms), note that

$$\frac{1}{N}\left(N_1 \hat{TT} + N_0 \hat{TUT}\right) =$$

$$\frac{1}{N}\{\sum_{RC_i=1}[(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)'\mathbf{x}_i + (\hat{\lambda}_1 - \hat{\lambda}_0)\hat{IMR}_{i1}] + \sum_{RC_i=0}[(\hat{\boldsymbol{\beta}}_1 - \hat{\boldsymbol{\beta}}_0)'\mathbf{x}_i + (\hat{\lambda}_1 - \hat{\lambda}_0)\hat{IMR}_{i0}]\}$$

$$= \hat{ATE} + (\hat{\lambda}_1 - \hat{\lambda}_0)\frac{\sum_{RC_i=1}\hat{IMR}_{i1} + \sum_{RC_i=0}\hat{IMR}_{i0}}{N}.$$

That is, the weighted average can again be decomposed into a component constituting the true effect of the BE (the ATE), and a component constituting the average influence of the propensity to live in an urban environment (i.e., the portion of the total effect that is due to self-selection).

We speculate that the studies that use the TT as the denominator may have been considering Heckman's original sample selection model, in which there is only one outcome (i.e., cases are either treated or not observed), instead of the endogenous switching model (for which all cases are observed, either as treated or untreated). In the endogenous switching context, it seems natural that it should not make any difference whether one environment or the other is labeled as the treatment or control—but using ATE/TT yields different results when treatment and control labels are switched. The weighted average of TT and TUT has the further satisfying feature that its computation is based on the entire sample—making it a more appropriate counterpart to the ATE numerator, which by construction is computed over the entire sample— whereas in theory the TT measure (including its ATE component, as described above) is computed only over the treated observations.[9]

The reader may be confused by the fact that TT constitutes the numerator of the BEP for the PSM method (as applied so far) but the denominator of the BEP for SS (in two out of three applications reviewed here). To help clarify matters, we can point out that the quantity labeled "treatment effect on the treated" (and similarly for the untreated) has substantively different content in the two contexts. In the SS context, it is clear from the equations and accompanying discussion that TT includes both a "true" effect and a bias term representing the effect of self-selection, whereas in the PSM context, the matching process, occurring before TT is computed, is intended precisely to remove that bias, leaving the true effect (on the treated). Thus, in both instances the incorporation of TT is consistent with the concept of the BEP: in the denominator when (for SS) it reflects a "total" effect of the BE, and in the numerator when (for PSM) it reflects a "pure" effect.

[9] Although, as mentioned in the preceding footnote, Limdep 10.0 computes TT by using arguments that are averaged over the entire sample.

3 Review of studies that have quantified the true BE vs. RSS effect

As indicated in the introduction, only a relatively small number of studies have quantified the proportion of the effect of the BE on TB that is due to the BE itself, i.e., the BEP. Table 1 summarizes ten models in seven studies we have identified that explicitly quantified a value between 0 percent and 100 percent for the share of total BE effect due to each factor. In the remainder of this section, we briefly describe each study, together with the approach it used to quantify the BEP.

3.1 Propensity scores

Xinyu (Jason) Cao and his co-authors have pioneered the application of the propensity score approach to the RSS context, so far producing four different studies involving this approach.

3.1.1 Cao (2010)

Taking data from 1553 responses to a self-administered Northern California survey distributed in 2003, Cao (2010) used stratification on the propensity score as a means of controlling for self-selection. Neighborhood type ("traditional" and "suburban" residence) was used as the binary location variable for the propensity-score (logit) model with suburban residence as the refer¬ence category (i.e., for any analysis involving treatment effects, treatment = traditional, control = suburban). Neighborhoods were purposely chosen to vary systematically by neighborhood attributes, size of metropolitan area, and region in the state. Traditional neighborhoods were built mostly before World War II, and suburban ones were built more recently. A number of residential preferences and travel attitudes were available and incorporated into the propensity score equation. Residential preferences dealt with accessibility, physical activity options, safety, socializing, attractiveness, and outdoor spaciousness, while travel attitudes included being pro-walk/bike, pro-transit, pro-travel, travel-minimizing, car safety-conscious, and car dependent.

The study provides estimates of the BEP for two types of walking behaviors: strolling frequency (recreation) and walking-to-store frequency (transportation). The BEP was computed as the average treatment effect (ATE) divided by the observed influence (obs. inf.), where (see, e.g., Imbens 2004, p. 18) the ATE is calculated by first finding the difference in travel behavior between traditional and suburban residents within five strata of equal size (307–309 cases in each), and then taking the average of these differences, weighting each difference by the total number of individuals in the stratum for which the difference is calculated. To compare the differences before and after stratification, he used independent sample t-tests with Levine's test for inequality of variances, and F-statistics for main and interaction effects after stratification. The study found that after self-selection had been controlled, the influence of the BE that remained was 86 percent and 61 percent of the total for strolling frequency and walking-to-store frequency, respectively.

Table 1: Studies that quantify RSS

	Study	N	Location, Year Data Collected	BE Variable(s)	Control Neighborhood Type	Treatment Neighborhood Type	TB Variable	% of Influence of BE Due to Self-Selection ((1-BEP) × 100%)
Propensity Scores	Cao 2010 (stratification)	1,553	8 neighborhoods in No. California, 2003	Neighborhood type	Suburban	Traditional	Strolling frequency	14%
							Walking to store frequency	39%
	Cao, Xu, and Fan 2010 (matching)	3,376	Raleigh, North Carolina, 2006	Neighborhood type	*Respectively:* Urban, Urban, Inner-ring suburb, Urban, Inner-ring suburb, Suburb	*Respectively:* Inner-ring suburb, Suburb, Suburb, Exurb, Exurb, Exurb	Vehicle-miles driven	2–52%
	Cao and Fan 2012 (matching)	5,537	Raleigh, North Carolina, 2006	Density	Low density	High density	Person-miles traveled	28%
							Driving duration	66%
							Transit duration	49%
	Cao 2015 (matching)	1,682	8 neighborhoods in No. California, 2003	Neighborhood type	Suburban	Traditional	Vehicle-miles driven	25%
Sample Selection Modeling	Zhou and Kockelman 2008	1,903	Austin, Texas, region, 1998-99	Neighborhood type	CBD/urban	Rural/suburban	Vehicle-miles traveled	10-42%
	Cao 2009	1,479	8 neighborhoods in No. California, 2003	Neighborhood type	Traditional	Suburban	Vehicle-miles driven	24%
	Bhat and Eluru 2009	3,696	San Francisco Bay Area, 2000	Neighborhood type	Neo-urbanist	Conventional	Vehicle-miles traveled	17%

3.1.2 Cao, Xu, and Fan (2010)

Cao, Xu, and Fan (2010) used propensity score matching with multiple treatments to control for RSS with respect to vehicle-miles driven (VMD). The data comprised 3376 households completing a regional travel diary in the Research Triangle metropolitan area of North Carolina—the Greater Triangle Travel Study conducted in 2006. Variables available included socioeconomic traits, residential preferences with respect to length of commute, access to transit, access to a desirable school, neighborhood safety, and neighborhood amenities.

The study considers six pairs of "treatment" and "control" neighborhood types—inner-ring suburb and urban, suburb and urban, suburb and inner-ring suburb, exurb and urban, exurb and inner-ring suburb, and exurb and suburb—with binary logit propensity-score models estimated for each pair. These were classified based on the network distance between households' residence and the city center point. In contrast to many other self-selection studies (but similar to Bhat and Eluru 2009), this assignment was determined after the survey through analytical methods, as opposed to neighborhoods purposefully being selected to fit those classifications before the survey was administered. It is interesting to note (as presented below) that the range of BEPs found in the Cao, Xu, and Fan (2010) study is quite wide relative to other studies. However, this may represent a more realistic approach to assessing BE effects across the full spectrum of geographic diversity, rather than confining the analysis to contrasts between "traditional" and "suburban" neighborhoods, which have been selected precisely to exemplify stereotypes. Specifically, entire regions are not "stereotypically suburban" and capable of being turned into "stereotypically traditional." In contrast, the approach of Cao, Xu, and Fan (2010) could be useful for showing what could be expected if an "inner-ring suburb" became "urban," or if a "suburb" became an "inner-ring suburb," which would be much more realistic (even if applied to only the relevant subset of a region) than projecting more dramatic built environment changes across an entire region.[10]

For the matching itself, a treatment individual was randomly selected and the control case with the closest propensity score was matched to that individual. They used one-to-one without replacement (meaning that once a control case was matched to a treatment case, it was removed and not available for further matching), and a caliper width of 0.01 (if a control case could not be found having a propensity score within 0.01 of the treatment case's score, then the treatment case was not matched).

To test whether self-selection was controlled for, they used the standard difference (D'Agostino 1998): $\delta = 100\% \times \frac{(\bar{x}_T - \bar{x}_C)}{\sqrt{\frac{s_T^2 + s_C^2}{2}}}$, where $\bar{x}_T$ is the sample mean of a continuous covariate of the propensity-score equation for urban residents, $\bar{x}_C$ is the sample mean for suburban residents, s_T^2 is the sample variance for urban residents, and s_C^2 is the sample variance for suburban residents (for discrete covariates, the expression is conceptually the same but operationalized slightly differently; see,

[10] We note in passing that this point illustrates the complexity of our application of classic program effectiveness evaluation approaches. In many such applications there are two crisp conditions, treatment and control, and the designation/identification of cases as one or the other is relatively unambiguous (albeit with some potential for contamination and crossover). In our context there are several complications, in that "treatment" (1) could be considered a spatial continuum (e.g., as approximated by population density); (2) is actually multivariate (based not just on density, but also on diverse land uses, availability of transit, walk/bicycle-friendliness, and so on), with the particular set of relevant variables being not at all well-defined and consistently agreed-upon (Cao 2009); (3) could reasonably apply not just to the residential location, but to work and potentially to other frequently visited locations as well; and (4) could be episodically applied and "unapplied" as individuals move to different types of locations over a lifetime. Furthermore, (5) as seen in Table 1 and throughout Section 3, there is not even unanimity in the literature on which condition constitutes the "treatment" and which the "control." For these reasons, the classification of cases as treatment or control is actually rather muddy, although most studies in this context simplify the analysis by (1) discretizing the continuum, (2) selecting locations so as to exemplify the two contrasting types of neighborhoods, and (3) classifying cases only on the basis of their residential location at a single point in time.

e.g., Austin 2011, p. 412). This equation is applied before and after propensity matching to see whether residential self-selection is taking place (before) and whether it has been controlled for (after). That is, it is applied to see whether the covariates are "balanced" between treatment and control groups. The BEP is calculated as the treatment effect (TE)[11] divided by the observed influence, where the TE is the difference in mean VMD between the *matched* treatment and control residents and the observed influence is the difference in mean VMD for *unmatched* treatment and control residents. For the six (treatment/control) combinations, they found BEPs as follows: inner-ring suburb/urban = 48 percent, suburb/urban = 67 percent, suburb/inner-ring suburb = 78 percent, exurb/urban = 84 percent, exurb/inner-ring suburb = 98 percent, and exurb/suburb = 95 percent).

3.1.3 Cao and Fan (2012)

Cao and Fan (2012) use propensity-score matching to control for self-selection with respect to person miles traveled, driving duration, and transit duration. This time taking 5537 households from the same 2006 (North Carolina) Great Triangle Travel Survey as Cao et al. (2010), they designate high- and low-density locations as treatment and control, respectively. In this study, they used a binary probit model, whereas Cao et al. (2010) used a binary logit model.

In addition to estimating the BEP, the authors discuss in greater depth whether and when matching is an appropriate means of controlling for selection. They caution that the method assumes that all variables affecting treatment assignments are observed, and that there is no "hidden" bias. This assumption would be violated if, for example, attitudes were not measured (which was not the case in their study, as detailed in Section 3.1.2). While sample selection models can, in principle, correct for selection bias even in the case of omitted influential variables, propensity-score methods cannot. This is because (as mentioned in Section 2.2) sample selection models control for correlations between *unobservables* of the selection equation and *unobservables* of the outcome equations, while propensity scores only control for correlations between *observables* of the propensity score equation and *unobservables* of the outcome equation (Winship and Morgan 1999; Cameron and Trivedi 2005). If attitudinal variables (or other omitted variables) that are in reality influential on travel behavior are not observed and included, then none of the propensity-score approaches can compensate for the bias that results from that omission.

They perform a sensitivity analysis using various caliper widths, since it is possible to obtain varying results with different caliper widths and the extent to which this occurs often tells us something important about the outcome variable. For example, they found that varying caliper width produced the most variant results for transit duration among the three travel behaviors, and noted that such a result was not surprising given the high proportion of people who do not use transit at all. In other words, poor response to the sensitivity analysis in this case coincided with the travel behavior (outcome) variable taking on the same value most of the time. They also considered a 95 percent confidence interval for their estimates of the treatment effect,[12] since point estimates may be misleading. Based on these confidence intervals, BEP ranged from 5 percent to 97 percent (across all three travel behaviors). While they caution that point estimates can be inaccurate or misleading estimators of the true BEP, the point estimates they did find for BEP were 72 percent, 34 percent, and 51 percent for person miles, driving duration, and transit duration, respectively.

[11] The authors refer to this as the average treatment effect (ATE), citing D'Agostino (1998), but it appears to be the treatment effect on the treated, as described by Becker and Ichino (2002), Greene (2007), and others (personal communication of the first author with J. Cao, March 10, 2015). The same comment applies to the other two studies involving propensity -score matching

[12] See footnote 11.

3.1.4 Cao (2015)

Cao (2015) focuses on propensity-score matching, involving the same data that was used for propensity-score stratification in Cao (2010) and for sample selection modeling in Cao (2009). He discusses some of the differences between the two approaches we consider in the present paper (plus a third, the statistical controls method). Using 1682 respondents from the 2003 Northern California survey, self-selection is controlled for with respect to VMD; similar to Cao (2009, 2010), traditional and suburban are the binary choices for the propensity-score equation, with treatment = traditional and control = suburban. Similar to Cao, Xu, and Fan (2010), which also employed matching, one-to-one matching without replacement and a caliper width of 0.01 was used, standard differences of the covariates of the propensity-score equation were used to determine whether self-selection had been controlled for after matching, and BEP was calculated in the same way.[13] He finds that about 75 percent of the built environment's influence was attributable to the built environment itself. He compares this to his 2009 study (Cao 2009, described in Section 3.2.2 of the present paper), which applies a sample selection model to the same data, where the result was 76 percent for the BEP.

3.2 Sample selection

Sample selection models differ from ordinary regression outcome models in two ways. First, in our context the coefficients of the outcome equation are allowed to differ by each state of the endogenous selection variable (the binary, treatment-versus-control or urban-versus-suburban variable). Second, a selection correction term is added to each outcome equation in addition to the other covariates (as discussed in Section 2.3). The formulas for the selection correction terms depend on the assumptions made about the distribution of the error term of the selection equation. A probit selection equation is convenient since the error covariance among the three equations is then trivariate normal distributed as presented above, but certainly other structures are possible as discussed below (Bhat and Eluru 2009).

3.2.1 Zhou and Kockelman (2008)

Zhou and Kockelman (2008) applied the mover-stayer sample selection model to a sample of 1903 household observations from the 1998-99 Austin Area Household Travel Survey, supplemented with ArcGIS-encoded zonal data, to model daily vehicle-miles. Neighborhood type is defined at the level of Traffic Analysis Zones (TAZs), which are classified as "rural or suburban" or "CBD or urban." Counter to the convention used in this paper and elsewhere (e.g., Cao 2010; Cao and Fan 2012; and Cao 2015)—but consistent with the other two papers using sample selection approaches (Cao 2009; Bhat and Eluru 2009) as well as Cao, Xu, and Fan (2010)—they define treatment as being located in a suburban or rural zone (perhaps so that the expected change in vehicle-based travel behavior will be positive). As with most public sector-sponsored household travel surveys, attitudes were not available, and the authors note that the sample selection model was used specifically for this reason.

They consider four treatment effects described in Heckman and Vytlacil (2005) and Heckman, Tobias, and Vytlacil (2001): the average treatment effect (ATE), the treatment effect on the treated (TT), the local average treatment effect (LATE), and the marginal treatment effect (MTE). For the calculation of the BEP, only the first two are of interest. The ATE is the increase in VMT that would be expected if an individual randomly chosen from the entire population were to move from the CBD/urban environ-

[13] See footnote 11.

ment (control) to the rural/suburban (treatment) environment. The TT is the increase in VMT (relative to the VMT if living in an urban location) that would be expected from an individual randomly chosen from among suburban residents (i.e., from among those who have selected the treatment of living in a suburb). They compute the relative effects (i.e., ATE/TT) as an indicator of the BEP (without using that terminology).

As far as we are aware, Zhou and Kockelman (2008) were the first to use the ratio of ATE/TT as a measure of the proportion of the total effect of the BE that can be considered "true," i.e., remaining after self-selection is accounted for. For example, although Rosenbaum and Rubin (1983, 1984) talk about *reduction* in selection bias (and its corresponding estimate), they do not explicitly speak of a ratio between ATE and observed influence. Zhou and Kockelman (2008) found an ATE of 17.0 vehicle-miles per day (indicating that a *randomly selected* household would increase VMT by 17.0 miles per day, on average, if moving from a CBD/urban environment to a rural/suburban environment), and a TT of 29.2 vehicle-miles per day (indicating that a household *living in a rural or suburban environment* can be expected to exhibit 29.2 more daily VMT than an observationally equivalent one living in a CBD or urban environment). This indicates a BEP of 58 percent. However, they also tested the sensitivity of the results to the selection model specification. When the presence of four or more visitors on the travel survey day was included in that model, they found that the ATE was 20.2 vehicle miles per day and TT was 22.5 vehicle miles per day, for a BEP of 90 percent. They conclude that further varying the specification will produce yet different results, but based on the two specifications tested, they suggest values of 10-42 percent for the role of self-selection (i.e., the complement of the BEP).

Note, however, that because Zhou and Kockelman (2008) reversed the labels of treatment and control, the TT of their study would correspond to the TUT of studies with the opposite labeling scheme (with the sign of the effect also reversed); i.e., in other studies it would be the TUT that would correspond to the average *decrease* in VMT if a randomly selected resident of a suburban neighborhood (i.e., an *untreated* person) could also be observed living in an urban neighborhood (i.e., receiving the treatment). Thus, if they had chosen the opposite labeling scheme, their computed BEP would differ by virtue of having a different denominator (while the numerator would remain the same except for a sign reversal). A similar problem would pertain to the propensity score matching methods as typically applied (except that it would be the observed-influence denominator that would remain the same except for the sign reversal, and the treatment-effects numerator that would differ). In contrast, as can be seen from the discussion in Section 2.3.1, propensity-score stratification (as applied in this context, to date) is indifferent to the choice of labels of "treated" versus "untreated."

3.2.2 Cao (2009)

Using 1479 respondents from a 2003 Northern California survey (same data as Cao 2010, 2015), Cao (2009) follows a procedure nearly identical to that of Zhou and Kockelman (2008). In addition, he considers measures for whether self-selection had been controlled for with respect to vehicle-miles driven. The selection variable is traditional (control) versus suburban (treatment) location, where the categories were pre-determined as part of the original sampling design. Residential preferences and travel attitudes were available and allowed to enter both the selection and outcome equations (see Section 3.1.1 for the available variables). Not surprisingly, residential attitudes were somewhat more prevalent in the selection equation than were travel attitudes (three versus two variables, respectively), but travel attitudes did enter and were also clearly significant in the travel-behavior equations, whereas residential preferences were not significant in the latter case. A log transformation of weekly VMT was adopted as the outcome variable, so the equations for ATE and TT used in the calculation of BEP were in exponential form. Cao found an ATE of 25.8 miles and a TT of 33.8 miles per week, suggesting a BEP of 76 percent. As

with Zhou and Kockelman, however, this figure would presumably differ (by virtue of the denominator being different) if the treatment and control labels were reversed.

It is notable that built-environment variables were not included in either the selection equation or either of the two outcome equations. Although it might be appropriate to allow built environment variables into the outcome equations, it is more appropriate *not* to include them in the selection equation. This is because there is no obvious way to measure the BE of the *non-chosen* neighborhood type, and including the BE only of the *chosen* neighborhood type would be a misspecification (since the choice is assumed to be based on a comparison of BE characteristics across *all* alternatives, not just a function of the characteristics of the *chosen* alternative). Put another way, associations between built environment variables and the residential choice are more likely to indicate *properties* of that *already-chosen* built environment rather than a causal influence on *choosing* it, and model fit statistics would artificially indicate a better causal model than is actually the case.

3.2.3 Bhat and Eluru (2009)

All of the models thus far have used conventional assumptions about the underlying error covariance structure of the equations involved. However, real data rarely follow these conventional assumptions to the extent we would like. For those cases where they do not, Bhat and Eluru (2009) explore alternatives to conventional assumptions. They use a sample selection model, but unlike other studies, use a "copula"-based approach to relax the assumption of bivariate normally distributed errors between the selection equation and each outcome equation, which they assert can be restrictive and inappropriate. They consider residential neighborhood choice (conventional vs. neo-urbanist, with conventional labeled the treatment condition, similar to Zhou and Kockelman 2008, but counter to most other studies) as a selection variable and daily household vehicle miles of travel (VMT) as the outcome variable, using 3696 observations from the 2000 San Francisco Bay Area Household Travel Survey (BATS). Exogenous variables include household characteristics, employment characteristics, and neighborhood characteristics; attitudinal data were not available.

Heckman's original formulation depends on using the IMR and assuming jointly normally distributed error terms for the selection and outcome equations. Lee (1983) generalized Heckman's approach by allowing univariate error terms to be non-normal, using a technique to transform non-normal variables into normal variables. This is one method commonly employed for dealing with sample selection models if the selection equation is not probit, or possibly not even binary. For example, Lee's method could be used in conjunction with multinomial logit (MNL) as the selection equation (although Dubin and McFadden, 1984, suggested an estimator more commonly accepted today to account for selection bias in the specific case of multinomial logit models, and Bourguignon et al. 2007 have suggested a very similar estimator to that of Dubin and McFadden's, involving a minor relaxation of a restriction placed on the correlations of the error terms).

Bhat and Eluru (2009, pp. 751–752) further generalize the possible relationships among error terms by using the copula—"a device or function that generates a stochastic dependence relationship (i.e., a multivariate distribution) among random variables with pre-specified marginal distributions." Since random variables may not have the same marginal distributions, this may be desirable since it allows for considerable flexibility in correlating random variables. They test six different possible copulas, respectively known in the literature as Gaussian, Farlie–Gumbel–Morgenstern (FGM), Clayton, Gumbel, Frank-Frank (F-F), and Frank-Joe (F-J). For each of those copulas, there is a different set of corrective terms entered into the regression equations (analogous to the IMR for Heckman's original model; see Bhat and Eluru, 2009, for more detailed formulas for the bias correction terms).

In the model they deem best (the F-F copula), they conclude that 83 percent[14] of the difference in VMT between households residing in conventional and neo-urbanist neighborhoods is due to "true" built environment effects, while 17 percent is due to self-selection. However, as indicated in Section 2.3.2, they use a different denominator than Zhou and Kockelman (2008) and Cao (2009). Specifically, their denominator is the weighted average of the treatment effect on the treated and the treatment effect on the non-treated. If they used only the treatment effect on the treated, consistent with the other two studies, then the value for BEP would be 51 percent—a substantial difference.

4 Discussion and conclusions

In the now-sizable body of research dealing with the influence of residential self-selection on the impact of the built environment on travel behavior, the number of studies that quantify the built-environment proportion, or BEP (defined as the proportion of the total influence of the built environment that is not due to self-selection), is still relatively small. However, it is of policy relevance as well as academic interest, both to analyze the collective information provided by those studies, and to promote additional efforts to quantify this key measure through promulgating a description of the methods used to do so. Accordingly, the purpose of this paper was to review and analyze that specific set of literature, with respect to (1) methods used, and (2) results found. To keep the focus and length of the paper manageable, we limited ourselves to studies that used either propensity-score matching/stratification or a sample selection approach.[15] We identified 10 analyses in seven studies that quantified the BEP using one of these methods.

With respect to computing the BEP, the specifics differed by the method used to control for self-selection. For propensity-score methods, the denominator was a simple difference of mean outcomes between the treated and untreated groups (the "observed influence"), while the numerator was (for stratification) an "average treatment effect" (ATE) or (for matching) "treatment effect on the treated" (TT) that had been purged (to the extent possible) of the self-selection bias. For sample selection, the numerator was the average treatment effect (computed differently), representing the average change in travel behavior after controlling for self-selection (i.e., the true impact of the built environment), and the denominator was either a "treatment effect on the treated" that retained a "true" as well as a "bias" component, or the sample-size-weighted average of that quantity with the counterpart "treatment effect on the untreated." The first of these two ways of computing the denominator is sensitive to the choice of which condition is labeled "treatment" versus "control." We recommend that future studies use the formulation that is robust with respect to the choice of labels (namely, the weighted average of the treatment effect on the treated and the treatment effect on the untreated).

With respect to the results found, they are far from unanimous, with estimates of the *true* influence of the built environment running the gamut from 34 percent to 98 percent of the *total* apparent influence. Clearly there are a number of factors that could account for this range: the studies involve several different samples, collected at various locations and times, using different definitions of residential location categories, different labeling conventions (of treatment versus control) in some cases, different explanatory variables, and different outcome variables of interest, as well as different methodologies. But we still see considerable diversity even when many of these factors are held constant. For example, a single propensity-score stratification study (Cao 2010) finds BEPs ranging from 61 percent to 86 percent, varying only the travel outcome variable within essentially the same dataset. Similarly, varying only the selection of which pairs of residential location categories to compare, Cao, Xu, and Fan (2010) finds BEPs ranging from 48 percent to 98 percent. The only evidence available that compares methodology

[14] Per their Section 6 and computation from Table 4 [(21.37/25.59) × 100 percent]. Their Section 5 refers to "87%", which we believe to be a typographical error.

[15] We are aware of only a handful of studies that have quantified the BEP using other approaches.

while holding other factors constant found the BEPs to be quite similar for propensity-score matching (75 percent, Cao 2015) and sample selection (76 percent, Cao 2009). However, more comparisons like these are needed to indicate whether this is a trend or a coincidence, especially given that (1) the latter study used the label-dependent denominator in computing the BEP (see the discussion in Section 3.2.1), whereas the BEP measure in the former study is label-independent, and (2) the latter study has ATE in the numerator of the BEP, whereas the former study apparently has TT.

Considerable work remains for future research. First and foremost, we simply need more studies that quantify the BEP. Once the evidence base is larger, it could be fruitful to analyze the pool of such studies to see if any systematic variations in the BEP estimates can be discerned with respect to the factors identified above. We are particularly interested in whether the methodology per se makes a sizable difference. Work is underway to investigate this question, and also to expand the family of ways to estimate the BEP by transferring, to the extent feasible, methods used with one approach (such as statistical controls) to other approaches (such as sample selection). We hope that this initial offering, together with the work in progress, will help stimulate additional contributions to improving our knowledge in this important area.

Acknowledgements

The authors wish to thank Daniel Rodriguez for suggesting the Cook, Shadish, and Wong (2008) and the Shadish, Clark, and Steiner (2008) references. We also benefited from comments by participants in the Conference on Low-Carbon Cities: Land Use and Transportation Intervention, Xi'an, China, June 12–13, 2014, and the World Symposium on Transport and Land Use Research, Delft, the Netherlands, June 24–27, 2014. We are grateful to Gouri Shankar Mishra for extremely valuable discussions, identification of pertinent literature, and thoughtful reviews of earlier drafts.

References

Austin, P. C. 2011. An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research* 46: 399–424.

Becker, S. O., and A. Ichino. 2002. Estimation of average treatment effects based on propensity scores. *The Stata Journal* 2(4): 358–377.

Bhat, C. R., and N. Eluru. 2009. A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological* 43(7): 749–765.

Cameron, A. C., and P. K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* New York: Cambridge University Press.

Cao, X. 2009. Disentangling the influence of neighborhood type and self-selection on driving behavior: An application of sample selection model. *Transportation* 36: 207–222.

Cao, X. 2010. Exploring causal effects of neighborhood type on walking behavior using stratification on the propensity score. *Environment and Planning A* 42(2): 487–504.

Cao, X. 2015. The effects of neighborhood type and self-selection on driving: A case study of Northern California. Chapter 10 in *International Handbook on Transport and Development*, edited by R. Hickman, M. Givoni, D. Bonilla, and D. Banister. Cheltenham, UK: Edward Elgar.

Cao, X., and Y. Fan. 2012. Exploring the influences of density on travel behavior using propensity score matching. *Environment and Planning B: Planning and Design* 39(3): 459–470.

Cao, X., Z. Xu, and Y. Fan. 2010. Exploring the connections among residential location, self-selection, and driving: Propensity score matching with multiple treatments. *Transportation Research Part A: Policy and Practice* 44(10): 797–805.

Cao, X., P. L. Mokhtarian, and S. L. Handy. 2011. Examining the impacts of residential self-selection on travel behavior: methodologies and empirical findings. Chapter 1 in *Urban Sustainable Mobility,* edited by E. Venezia. Milan: FrancoAngeli, pp. 15–100. Available as Research Report UCD-ITS-RR-08-25. http://www.its.ucdavis.edu/research/publications/publication-detail/?pub_id=1194.

Cook, T. D., W. R. Shadish, and V. C. Wong. 2008. Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons. *Journal of Policy Analysis and Management.* 27(4): 724–750.

D'Agostino, Jr., R. B. 1998. Propensity score methods for bias reduction in the comparison of a treatment to a non-randomized control group. *Statistics in Medicine.* 17(19): 2265–2281.

Greene, W. H. 2007. *LIMDEP Version 9.0 Econometric Modeling Guide Vol. 2.* Plainview, NY: Econometric Software Inc.

Heckman, J. J. 1979. Sample selection as a specification error. *Econometrica* 47(1): 153–161.

Heckman, J. J., J. L. Tobias, and E. J. Vytlacil. 2001. Four parameters of interest in the evaluation of social programs. *Southern Economic Journal* 68(2): 210–223.

Heckman, J. J., and E. J. Vytlacil. 2005. Structural equations, treatment effects, and economic policy evaluation. *Econometrica* 73(3): 669–738.

Ho, D. E., K. Imai, G. King, and E. A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15: 199–236.

Imbens, G. W. 2004. Nonparametric estimation of average treatment effects under exogeneity: A review. *The Review of Economics and Statistics* 86(1): 4–29.

Joh, K., N. Mai Thi, and M. G. Boarnet. 2012. Can built and social environmental factors encourage walking among individuals with negative walking attitudes? *Journal of Planning Education and Research* 32(2): 219–236.

Larco, N., B. Steiner, J. Stockard, and A. West. 2012. Pedestrian-friendly environments and active travel for residents of multifamily housing: The role of preferences and perceptions. *Environment and Behavior* 44(3): 303–333.

Oakes, M. J., and P. J. Johnson. 2006. Propensity score matching for social epidemiology. In *Methods in Epidemiology,* edited by M. J. Oakes and J. S. Kaufman. New York: John Wiley & Sons.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of the propensity score in observation studies for causal effects. *Biometrika* 70(1): 41–55.

Rosenbaum, P. R., and D. B. Rubin. 1984. Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* 79(387): 516–524.

Shadish, W. R., M. H. Clark, and P. M. Steiner. 2008. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random and nonrandom assignments. *Journal of the American Statistical Association* 103(484): 1334–1356.

Tucker, J. W. 2010. Selection bias and econometric remedies in accounting and finance research. *Journal of Accounting Literature* 29: 31–57.

Winship, C., and S. L. Morgan. 1999. The estimation of causal effects from observational data. *Annual Review of Sociology* 25: 659–707.

Zhou, B., and K. Kockelman. 2008. Self-selection in home choice: Use of treatment effects in evaluating the relationship between the built environment and travel behavior. *Transportation Research Record* 2077: 54–61.