RWI : Discussion Papers

Colin Vance and Ralf Hedel

# On the Link between Urban Form and Automobile Use

## Evidence from German Survey Data

No. 48

RWI ESSEN

# RWI : Discussion Papers
# No. 48

Published by Rheinisch-Westfälisches Institut für Wirtschaftsforschung,
Hohenzollernstrasse 1/3, D-45128 Essen, Phone +49 (0) 201/81 49-0

Editor: Prof. Dr. Christoph M. Schmidt, Ph.D.
ISSN 1612-3565 – ISBN 3-936454-75-2

RWI : Discussion Papers

No. 48

Colin Vance and Ralf Hedel

# On the Link between Urban form and Automobile Use

## Evidence from German Survey Data

RWI
ESSEN

**Bibliografische Information Der Deutschen Bibliothek**
Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.ddb.de abrufbar.

ISSN 1612-3565
ISBN 3-936454-75-2

**Colin Vance and Ralf Hedel ***

# On the Link between Urban Form and Automobile Use – Evidence from German Survey Data

*Abstract*

This study investigates the influence of urban form on automobile travel using travel-diary data from Germany. Two dimensions of car use are considered: the discrete decision to own a car and the continuous decision of distance traveled. Because these decisions are likely to be influenced by factors unobservable to the researcher, we apply censored regression models to evaluate the role of biases emerging from sample selectivity. Unlike much of the literature, we find that urban form variables are a significant determinant of both automobile ownership and use, a finding that holds even after using instrumental variables to control for endogeneity.

JEL classification: R14, R41

Keywords: urban form, non-work automobile travel, sample selectivity, instrumental variables

September 2006

---

*Colin Vance, RWI Essen, Germany; Ralf Hedel, Institute for Transport Research, German Aerospace Center (DLR), Berlin. – The authors wish to thank Elmar Brockfeld for assistance in creating the GIS measures of road density. We would also like to acknowledge the helpful comments of Manuel Frondel and Christoph M. Schmidt on an earlier draft. All correspondence to Colin Vance, Rheinisch-Westfälisches Institut für Wirtschaftsforschung (RWI Essen), Hohenzollernstr. 1–3, 45128 Essen, Germany, Fax: +49 201 / 81 49-200. Email: vance@rwi-essen.de.

## 1. Introduction

In Germany, as elsewhere in the industrialized world, the demand for motor vehicle travel has increased substantially in recent years, with the number of registered vehicles increasing by 12.1% and mileage increasing by 9.1% between 1993 and 2002 (Hautzinger, Stock 2005). Understanding the determinants of such trends has emerged as a major priority within the scientific and policy arenas given the range of negative externalities associated with private car use, including air and noise pollution as well as congestion and accidents on the public roadways. One important area of research has focused on the role of urban form as a cause of transport demand. The hypothesis of a link between urban form and automobile dependency has far reaching implications for transport and land use policy, as its verification would avail a broad palette of options to encourage alternative modes of travel. Establishing such a link empirically, however, has proved a vexing endeavor.

While several early studies uncovered evidence that land use planning – in particular high densities and mixed use – reduce automobile travel (e.g. Newman, Kenworthy 1989; Holtzclaw 1990; Friedman et al. 1994; cited in Handy 1996), more recent studies have suggested that the linkage is tenuous at best. Much of this research draws on household data collected from select cities and regions in the United States, where the widespread movement of populations away from urban areas has been argued to not only contribute to urban blight (Jargowsky 2001), but also to a loss of cultural heritage as open-space is replaced by helter-skelter development and disconnected residential communities (Kunstler 1994). Concern about these trends has spawned new paradigms in the planning community – alternatively referred to as the "new urbanism" or "neotraditional planning" – that emphasize compact design, mixed development and the provision of public transport as a means of integrating neighborhoods and reducing congestion. As cities and metropolitan areas develop strategies to implement these concepts, one of the most contentious conclusions emerging from the recent literature is that any such measures may have only muted effects. To the extent that households endogenously self-select themselves into communities that support their preferences for transportation- and other amenities, there will be less leverage to influence their behavior through changes in land use. The studies by Boarnet/Sarmiento (1998), Crane/Crepeau (1998), Krizek (2003), and Bento et al. (2005), all of which analyze household data collected from U.S. cities, are among those that find a modest role of individual measures of urban form in explaining automobile-based travel behavior.

To date, this issue has received very little empirical scrutiny in the European context, despite growing concerns about unmitigated land consumption and calls to avoid unbalanced urban planning (EC 2001). In 1993, the German government legally codified the concept of "decentralized concentration" into its regional planning guidelines (BBR 1993). Since that time, several German cities have adopted models predicated on "compact" development as a means of spatially integrating residential, recreational and commercial land uses to reduce automobile reliance (e.g. Dresden 2002). One of the few studies to examine the efficacy of such measures was recently commissioned by the Federal Ministry of Transport, Building and Housing, which drew on a cross-sectional survey of individual travel behavior (Siedentop et al. 2005). Although the authors are able to establish correlations between various measures of urban form and total travel, they find relatively weaker correlations between these measures and automobile travel. Moreover, they concede that their findings may be subject to the confounding influence of residential choice decisions.

The present study seeks to further contribute to this line of inquiry by estimating econometric models of car use on a panel of travel-diary data collected in Germany between 1996 and 2003. Two dimensions of car use are considered: the discrete decision to own a car and the continuous decision of distance traveled. Because these decisions are related and, moreover, are likely to be influenced by factors unobservable to the researcher, we explore alternative specifications using censored regression techniques to assess whether the results are subject to biases emerging from sample selectivity. Our focus is specifically on the determinants of non-work travel, as this variant generates a majority of the local area travel (Crane, Crepeau 1998) and tends to be more flexible and hence potentially more responsive to the built environment (Krizek 2003). In addition to modeling variables that capture the tools commonly advocated by land use planners to influence mobility behavior, including mixed use and public transit, an important contribution of this study is to control for endogeneity through the use of instrumental variables.

The remainder of the paper is structured as follows. The next section describes the data sources and their assembly for the quantitative analysis. Section 3 describes the econometric models, the explanatory variables included in the specification, and some technical details on the interpretation of the marginal effects. Section 4 catalogues the results, and section 5 concludes the paper.

## 2. Data assembly

The primary data source used in this research is drawn from the German Mobility Panel (MOP), a representative multiyear travel survey financed by the German Federal Ministry of Transport, Building and Housing. The panel is organized in overlapping waves, each comprising a group of households surveyed for a period of one week over each of three years. The data used in this paper cover eight waves of the panel, spanning 1996 to 2003.

Households that participate in the survey are requested to fill out a questionnaire eliciting general household information and person-related characteristics, including gender, age and employment status. In addition, all household members over 10 years of age fill out a trip log capturing relevant aspects of everyday travel behavior, including distances traveled, modes used, activities undertaken, and activity durations. Using the data from these logs, we derived a measure of the total distance driven by the household over the course of the five-day week for non-work activities, which serves as the dependent variable. For cases in which these activities were undertaken as part of tours that involved work stops (e.g. a stop at the supermarket on the way home from work), we subtracted twice the direct home-work distance in order to remove work-related travel from the measure.

While the MOP includes a variable indicating the zip code in which the household resides, it lacks sufficiently detailed geospatial information to derive measures of community attributes. Moreover, the MOP includes no direct measures of household level income. To redress these features, we augmented the data with additional information obtained from Navtech and infas GEOdaten GmbH, two commercial data providers. The Navtech data includes shapefiles of the zip code boundaries and the roads network in Germany. The infas GEOdaten data is drawn from the year 2001 and includes information on per capita disposable income and demographic composition, as well as counts of the number of building structures and business outlets of various ages and types (e.g. residential, retail, service, etc). As with the Navtech data, this data is measured at the scale of the zip code, the median size of which is roughly 27 square kilometers. Both the GEOdaten and Navtech data sets were merged with the MOP data using the zip code identifier. A final data source was provided by the German Insurance Association (GDV), which compiles information on car insurance costs nationally at the provincial level. Because the provinces – of which there are roughly 445 in the dataset – are delineated at a slightly higher level of spatial aggregation than zip codes, a GIS was used to merge them with the household data.

In total, the data contains 3064 households distributed across 2122 zip codes within the 1996–2003 time interval. Of these, 875 participated in one year of the survey, 878 in two years and 1311 in all three years, yielding a total of 6564 observations on which the model is estimated[1]. To correct for the non-independence of repeat observations, the model is specified using robust regression techniques that account for clustering on the household.

---

[1] To assess the possibility of attrition bias, we explored specifications that included dummy variables indicating the one and two-year participating households. As the dummies were either insignificant or had a negligible impact on the remaining coefficients, they were left out of the final specifications.

## 3. Empirical issues

### 3.1 The estimators

Roughly 14% of the households in the data do not own an automobile and for which the observations on distance driven are consequently censored at zero. To accommodate this feature of the data, two alternative estimators are specified: the Heckit (Heckman 1979) and the two-part model. The Heckit model is appropriate when there is a concern for sample selection biases that could otherwise arise from the existence of unobservable variables (e.g. attitudes toward public transit) that determine both the discrete and continuous choices pertaining to car use. Such biases may emerge from the possibility that the determinants of car ownership are not random: those households that would drive short distances are the same households that are less likely to own a car.

The model considers that observations are ordered into two regimes. In the context of the present example, these regimes are defined by whether the household owns a car. The first stage, referred to as the selector equation, defines a dichotomous variable indicating the regime into which the observation falls:

(1) $$S_i^* = \tau' Z_i + u_i$$

(2) $$S_i = 1 \text{ if } S_i^* > 0 \quad \text{and} \quad S = 0 \text{ if } S_i^* \leq 0,$$

where $S_i^*$ is a latent variable indicating the utility from car ownership, $S_i$ is an indicator for car ownership status, the $Z_i$ denote the determinants of this status, $\tau'$ is a vector of associated parameter estimates, and $u_i$ is an error term having a standard normal distribution. After estimating $\tau$ using the probit maximum likelihood method, the second stage involves estimating an OLS regression of distance traveled conditional on $S = 1$. To control for sample selectivity, this second stage regression appends the inverse Mills ratio (IVM) calculated from the linear predictions of the probit model as an additional explanatory variable. This second stage regression is referred to as the outcome equation and is written as:

(3) $$E(Y_i \mid S_i = 1, X_i) = \beta' X_i + \beta_\lambda \lambda_i + \varepsilon_i,$$

where $Y$ is the dependent variable, measured here as the kilometers of weekly non-work vehicle travel, $X$ are the explanatory variables, $\beta$ are the associated parameters to be estimated, $\varepsilon_i$ is an error term assumed to have a bivariate normal distribution with $u_i$, and $\lambda$ is the IVM, defined by the ratio of the density function of the standard normal distribution, $\phi$, to its cumulative density function, $\Phi$.

One difficulty in estimating the Heckit emerges when there is a high degree of collinearity between the independent variables and the IVM, resulting in high standard errors on the coefficient estimates and parameter instability. This problem has generated a vigorous debate in the literature concerning the usefulness of the model for handling censored data. Skeptics of the Heckit's reliability have advocated the two-part model (2PM) as an alternative (Hay et al. 1987; Duan et. al. 1984; Manning et al. 1987; Dow, Norton 2003). As with the Heckit model, the 2PM involves the estimation of a selector and outcome equation, but is distinguished by the exclusion of the IVM in the latter equation[2]. While acknowledging the merits of the 2PM under certain circumstances, Leung/Yu (1996) show that when the Heckit is the true model, it performs considerably better than the 2PM so long as there are no collinearity problems. Nevertheless, they also demonstrate that a t-test of the coefficient on the IVM – the conventional indicator for the presence of selectivity and hence the choice between the 2PM and Heckit – is unreliable when the degree of collinearity is excessive. Dow/Norton (2003) elaborate on this point by noting that high collinearity may cause the IVM coefficient to be unusually large, producing a t-test that incorrectly rejects the 2PM in favor of the Heckit.

Because the difficulties arising from multicollinearity complicate selection of the appropriate model, we present estimates from both variants for comparison. We also attempt to gauge the extent to which collinearity afflicts the results of the Heckit by reporting the condition number, a diagnostic tool suggested by Belsley et al. (1980). This measure, which indicates how close a data matrix $X$ is to being singular, is computed from the eigenvalues of the moment matrix. A higher condition number indicates a greater likelihood of collinearity problems, whereby Belsley et al. (1980) suggest a maximum threshold of 30 on the basis of Monte Carlo experiments.

### 3.2 The explanatory variables

To effectively identify the Heckit model and control for sample selectivity bias in the second stage regression requires the inclusion of at least one variable that uniquely determines the discrete choice of car ownership but not the continuous choice of distance traveled. In the present example, this selection can be informed by consideration of the fixed costs incurred with owning a car but not with driving. As an identifying variable, we include an insurance cost index compiled by the GDV, which varies between one and 12 and measures the average regional insurance costs of car ownership. We expect this variable to have a negative effect on the probability of car ownership, as increases in the index indicate higher insurance costs.

[2] Despite appearances, Duan et al. (1984) show that the 2PM is not nested in the Heckit because, in the case of the former model, the correlation of the error terms in equations (1) and (3) is irrelevant for the purposes of estimation.

To model the effects of urban form, we use four variables capturing three key elements of new urbanist design: public transit, density, and diversity. Public transit accessibility is captured by a variable from the survey that measures the walking distance, in minutes, from the house to the nearest public transit stop. Density is measured by two variables, the first of which is taken from the vector layer of the road network in Germany. Using this layer, an algorithm was written to calculate the total length and width of paved roads of various classes in each zip code, from which the density of road kilometers per square kilometer of land area could be derived. To maintain the focus on localized travel, we excluded the German Autobahn network from the calculation of the measure. The second density measure is derived by summing the total number of commercial outlets in the zip code and dividing by the zip code's area. Given our interest in non-work travel, the construction of this measure includes only retail, service and entertainment establishments.

Developing a measure of diversity is less straightforward, as the aim here is to simultaneously account for both the variety and prevalence of different attractions in the region that would influence mobility. We draw on an entropy-metric commonly employed in the biological sciences, referred to as Shannon's diversity index, which is based on information theory (Shannon, Weaver 1949). The index is defined as:

(4) $$H = -\sum_{j}^{Q} p_j \ln p_j ,$$

where $Q$ is the total number of "species" in a patch (in this case the zip code) and $p_j$ is the fraction of individuals belonging to the $j^{th}$ species. As with the density measure, we define retail, service and entertainment as the three classes of outlets over which the index is summed. Aside from being readily computable from the available data, a desirable feature of the index is that it takes into account both the number (i.e. richness) as well as the relative abundance of each type of outlet.

Aside from the distance to the nearest public transit stop, which is expected to be positively correlated with both the probability of car ownership and distance driven, the urban form variables could have either positive or negative effects, and it is not possible to state *a priori* which effects are expected to prevail. Diversity, for example, could be considered an amenity that encourages trips for shopping and other activities, thereby increasing the number of vehicle trips taken. At the same time, diversity might reduce the distance traveled for these trips if the mix of services sought is reachable within a smaller area. A similar logic applies to the density measures. A higher road density could be expected to reduce the costs and hence increase the amount of automobile travel while simultaneously increasing accessibility,

Table 1

**Descriptive statistics**

| | Mean | S.D. |
|---|---|---|
| Dependent variables | | |
| Own car (1,0) | 0.863 | 0.343 |
| Non-work km driven in 5-day week | 104.409 | 103.401 |
| Urban form variables | | |
| Outlet density | 0.609 | 1.282 |
| Road density | 13.086 | 11.555 |
| Outlet diversity | 0.887 | 0.075 |
| Walking minutes to public transit | 5.581 | 4.658 |
| Control variables | | |
| Number under 18 in household | 0.436 | 0.827 |
| Number over 64 in household | 0.354 | 0.638 |
| Number of working females | 0.408 | 0.514 |
| Number of working males | 0.469 | 0.532 |
| Number with college preparatory degree | 0.517 | 0.703 |
| Identifying variable | | |
| Insurance cost index | 6.208 | 2.907 |

which would reduce the traveled distances. Such countervailing effects are reflected in the available empirical evidence. Crane/Crepeau (1998), for example, find that dense street patterns are associated with fewer car trips relative to less dense networks, while Bento et al. (2005) find that increases in road density has a positive effect on annual miles driven.

The remaining variables included in the model serve as controls for the socio-demographic attributes of the household. Descriptive statistics for these, the urban form variables, the insurance cost index, and the dependent variables are presented in Table 1. Year dummies are also included in the model to control for autonomous shifts in macroeconomic conditions that could affect the sample as a whole.

### 3.3 Interpretation of the marginal effects

With respect to the interpretation of the marginal effects from the Heckit model, two clarifications are warranted. While it is well-known that the marginal effects from the probit selector equation are given by $\phi(\tau' Z)\tau_k$, it is often neglected that an adjustment is also required to interpret the coefficients from the outcome equation when the variable additionally appears in the selector equation. Following Sigelman/Zeng (1999), this adjustment, which yields the conditional marginal effect, is given by:

$$\frac{\partial E(y| S > 0, X)}{\partial X_k} = \beta_k - \tau_k \beta_\lambda \lambda(\lambda + \tau' Z). \quad (5)$$

The intuition underlying equation (5) is that the effect of $X_k$ can be decomposed into two parts. The first part, given by $\beta_k$, measures the effect of $X_k$ on the distance driven among those with a car, while the second part, represented by $\tau_k \beta_\lambda (\lambda + \tau' Z)$, represents the effect of a change in $X_k$ on the probability of owning a car (Saha et al. 1997). Clearly, when no sample selectivity issue is present, $\beta_\lambda = 0$ and the right hand side of (7) reduces to $\beta_k$. In this case, the conditional effect corresponds to that of the 2PM.

In addition to calculating the conditional marginal effects, we also calculate their statistical significance. This step is complicated by the fact that equation (5) comprises multiple parameters that makes analytical computation of the variance impossible, which is presumably one explanation for why it has been generally ignored in the literature. In the rare instances in which researchers present the marginal effects, the conventional practice seems to incorrectly rely on the standard errors of the unadjusted coefficient estimates for assessing significance. We circumvent this difficulty by applying the Delta method to calculate significance. This approach uses a first-order Taylor expansion to create a linear approximation of a non-linear function, after which the variance and measures of statistical significance can be computed[3].

## 4. Results

### 4.1 Models with urban form variables

Table 2 presents the coefficient estimates from the Heckit and 2PM models and the associated marginal effects. In discussing the results, we focus on these latter effects because of their behavioral relevance. Turning first to the selector estimates, all of the socio-demographic variables are significant at the 5% level and have positive coefficients. The largest effect is seen for the number of male workers in the household, a unit increase in which increases the probability of car ownership by 0.13. The identifying variable capturing insurance costs has, as expected, a negative coefficient and is also significant at the 5% level.

Moving beyond the control variables, the model suggests that urban form is a statistically significant determinant of car ownership: three of the four urban form variables have p-values less than 0.05 and a likelihood ratio test suggests that their inclusion significantly improves the fit of the model ($P < 0.0001$). The two density measures both have negative – albeit small – coefficients, which is consistent with the intuition that households located in densely settled areas have less need for automobile transit. Specifically, a unit increase

[3] The nonparametric bootstrap is an alternative method for calculating statistical significance. Vance (2006) illustrates this approach and the Delta method, and provides the code for implementing both techniques using the STATA software.

Table 2

**Vehicle ownership and distance traveled, Heckit and 2PM results**

| | Selector equation | | Outcome equation | | |
|---|---|---|---|---|---|
| | | | Heckit | | 2PM |
| | Coefficient | dy/dx | Coefficient | dy/dx | dy/dx |
| Urban form variables | | | | | |
| Outlet density | –0.090 (0.005) | –0.013 | –5.322 (0.021) | –3.028 (0.159) | –0.915 (0.686) |
| Road density | –0.019 (0.001) | –0.003 | –1.807 (0.000) | –1.312 (0.000) | –1.235 (0.000) |
| Outlet diversity | 0.749 (0.159) | 0.108 | 23.002 (0.353) | 3.922 (0.890) | –6.705 (0.785) |
| Walking minutes to public transit | 0.013 (0.046) | 0.002 | 0.877 (0.005) | 0.541 (0.129) | 0.623 (0.047) |
| Control variables | | | | | |
| Number under 18 in household | 0.248 (0.001) | 0.036 | 7.882 (0.000) | 1.553 (0.566) | 4.126 (0.038) |
| Number over 64 in household | 0.111 (0.025) | 0.016 | 13.121 (0.000) | 10.285 (0.001) | 11.326 (0.000) |
| Number of working females | 0.418 (0.000) | 0.060 | 17.457 (0.000) | 6.820 (0.075) | 7.992 (0.018) |
| Number of working males | 0.871 (0.000) | 0.125 | 39.661 (0.000) | 17.476 (0.000) | 19.713 (0.000) |
| Number with college preperatory degree | 0.353 (0.000) | 0.051 | 23.231 (0.000) | 14.228 (0.000) | 15.659 (0.000) |
| Average income of zip code | 0.010 (0.019) | 0.001 | 0.060 (0.741) | –0.192 (0.337) | –0.231 (0.175) |
| Insurance cost index | –0.026 (0.044) | –0.004 | | | |
| Inverse Mills ratio | | | 103.264 (0.000) | | |
| Constant | 0.090 (0.867) | | 41.305 (0.125) | | 106.059 (0.000) |
| Condition number | | | 23.108 | | |
| (Pseudo) $R^2$ | 0.204 | | 0.052 | | 0.046 |
| Number of observations | 6564 | | 5668 | | 5668 |

p-values in parentheses; year dummies not presented.

in road density reduces the probability of car ownership by 0.003, while a unit increase in outlet density reduces the probability by 0.013. The distance to the nearest public transit stop has the expected positive effect but is also small in magnitude, with each additional walking-minute increasing the probability of car ownership by 0.002. Likewise, the diversity variable is positive, which may reflect an increased demand for automobile travel as the variety of services in the surrounding area increases, though the estimate is statistically insignificant.

Before examining the corresponding estimates of the outcome equation, several points regarding the presence of sample selectivity bear noting. First, the condition number, calculated to be 8.45, is well below the threshold of 30 that Belsley et al. (1980) suggest is indicative of multicollinearity problems. Second, the coefficient on the inverse Mills ratio is positive and highly significant, suggesting that, on net, unobservable factors that increase the probability of car ownership also increase the distance driven. Third, the inclusion of the selectivity coefficient in the calculation of the marginal effects has a substantial bearing on their magnitude, indicating that the interpretation of Heckit estimates should be cast in more specific terms than is conventionally the case. Finally, the calculation of the statistical significance of the marginal effects is clearly warranted given that the p-values vary considerably compared with the unadjusted coefficients, in some cases by several orders of magnitude.

Turning to the estimates of the outcome equation for the Heckit and 2PM models, the qualitative findings generally mirror those of the selector equation. The coefficient on the number of male workers again has the highest magnitude, whereby the marginal effect from the Heckit suggests that each additional male worker increases the distance traveled for non-work activities by 17.5 kilometers. The corresponding estimate from the 2PM is slightly higher at 19.7 kilometers per 5-day week. A more substantial discrepancy is seen for the variable measuring the number of children under 18, the marginal effect of which is insignificant in the Heckit model but which is estimated at 4.1 and is highly significant in the 2PM. Otherwise, the coefficient estimates of the control variables are similar across the two models.

With respect to the urban form variables, only road density is found to be significant based on the calculation of the conditional marginal effects from the Heckit. Specifically, each unit increase reduces the kilometers traveled by 1.3 kilometers. A slightly lower estimate, 1.2 kilometers, is generated by the 2PM. The coefficient on the walking time to the nearest public transit from the 2PM is also significant, suggesting that each additional minute increases vehicle travel for non-work activities by 0.62 kilometers. Nevertheless, given the highly significant estimate of the IVM and the evidence that multicollinearity is unlikely to be a problem, a conservative interpretation would dictate referencing the Heckit result that the variable is insignificant.

### 4.2 Models with urban form instruments

The foregoing analysis assumes that urban form is an exogenous determinant of travel behavior. If we instead consider the possibility of endogeneity, with households selecting neighborhoods based on their preferences for the attributes embodied in the urban form variables, then the estimated coefficients on these variables would be inconsistent as a result of their correlation with the error term. Stated alternatively, if households who dislike driving locate in regions that support alternative transport modes, then any apparent causal effects running from the urban form variables to driving behavior could be spurious.

To explore this possibility, we estimated a second series of models using instrumental variables for both the Heckit and 2PM specifications. Selecting appropriate instruments requires identifying factors that determine land use in the area of the households residence but not their travel behavior. Here we follow the lead of Boarnet/Sarmiento (1998), whose work uses several non-transport amenities that measure the socio-demographic composition and architectural character of the neighborhood. We draw on four such variables from the infas GEOdata dataset, all measured at the zip code level:

- the percentage of buildings built before 1945
- the percentage of buildings built between 1945 and 1985
- the percentage of residents over 65 years of age
- the percentage of foreign residents.

As noted by Boarnet/Sarmiento (1998), these variables are likely to be correlated with the urban form variables but have little bearing on travel decisions. We thus assume that the associated instruments used in the estimation of the land use effects are exogenous, an assumption which is tested below.

Table 3 displays the results of instrumental variables regressions for each of the four urban form variables using the Heckit model. Columns 1–4 contain the coefficients estimated by an IV-probit model (Newey 1987), and columns 5–8 contain the corresponding estimates from a two-stage least squares model. To avoid clutter, only the marginal effects are presented.

Multicollinearity does not appear to be a problem for three of the four models, as evidenced by the relatively low values of the condition number. This is not the case for the model H4, however, which includes the instrument for the walking distance to the nearest public transit. The fact that the condition number is above 30, along with the insignificance of the IVM, suggests that the 2PM may be more appropriate in this case.

With respect to the identification of the models, two tests are reported at the bottom of the table. The first of these, the Anderson canonical correlations likelihood ratio statistic, provides a test of whether the excluded instruments are relevant. The null hypothesis is that the equation is under-identified. The test statistic is distributed as $\chi^2$ with degrees of freedom equal to the number of included and excluded instruments minus the number of regressors plus one. In all four models the null hypothesis is rejected, providing support for the relevance of the instruments. The second test reports the Hansen- Sargen statistic. The null hypothesis here is that the instruments are uncorrelated with the error term and that the excluded instruments are correctly excluded from the estimated equation, conditions which are required for the instruments to yield consistent and unbiased estimates. This test is also distributed as $\chi^2$, with

Table 3

**Vehicle ownership and distance traveled, Heckit using urban form instruments**
marginal effects

| | Selector equation (IV-Probit) | | | | Outcome equation | | | |
|---|---|---|---|---|---|---|---|---|
| | P1 | P2 | P3 | P4 | H1 | H2 | H3 | H4 |
| Instruments | | | | | | | | |
| Outlet density | –0.044 (0.000) | | | | –16.336 (0.000) | | | |
| Road density | | –0.039 (0.000) | | | | –1.928 (0.000) | | |
| Outlet diversity | | | –0.586 (0.338) | | | | 45.237 (0.740) | |
| Walking minutes to public transit | | | | 0.044 (0.000) | | | | 10.677 (0.000) |
| Control variables | | | | | | | | |
| Number under 18 in household | 0.035 (0.001) | 0.241 (0.000) | 0.048 (0.000) | 0.045 (0.004) | 3.343 (0.103) | 3.156 (0.133) | 4.263 (0.059) | 4.861 (0.014) |
| Number over 64 in household | 0.016 (0.039) | 0.122 (0.001) | 0.028 (0.001) | 0.005 (0.705) | 10.179 (0.001) | 11.313 (0.000) | 11.252 (0.000) | 8.107 (0.009) |
| Number of working females | 0.060 (0.000) | 0.420 (0.000) | 0.067 (0.000) | 0.082 (0.000) | 7.077 (0.036) | 7.556 (0.025) | 6.972 (0.055) | 11.572 (0.001) |
| Number of working males | 0.130 (0.000) | 0.848 (0.000) | 0.143 (0.000) | 0.154 (0.000) | 19.427 (0.000) | 18.730 (0.000) | 19.739 (0.000) | 21.726 (0.000) |
| Number with college preparatory degree | 0.050 (0.000) | 0.340 (0.000) | 0.043 (0.000) | 0.066 (0.000) | 16.179 (0.000) | 15.769 (0.000) | 13.697 (0.000) | 18.073 (0.000) |
| Average income of zip code | 0.001 (0.047) | 0.009 (0.000) | –0.002 (0.378) | 0.000 (0.672) | –0.144 (0.390) | –0.685 (0.686) | –0.476 (0.342) | –0.450 (0.002) |
| Insurance cost index | –0.004 (0.061) | –0.027 (0.003) | –0.012 (0.000) | –0.005 (0.140) | | | | |
| Inverse Mills ratio | | | | | 232.059 (0.000) | 145.074 (0.000) | –29.722 (0.524) | 300.163 (0.199) |
| Condition number | | | | | 8.933 | 9.434 | 15.308 | 48.185 |
| Anderson LR statistic | | | | | 422.367 (0.000) | 935.996 (0.000) | 227.061 (0.000) | 8.958 (0.062) |
| Hansen J statistic | | | | | 4.172 (0.243) | 0.970 (0.808) | 19.939 (0.000) | 14.458 (0.002) |
| Number of obs. | 6564 | | | | 5668 | | | |

p-values in parentheses; year dummies not presented.

the degrees of freedom equal to the number of excluded instruments minus the number of endogenous variables. With respect to the instruments for outlet and road density, the test statistic is very small, suggesting that the instruments are orthogonal to the dependent variable. However, in the case of the instruments for diversity and the distance to public transit, the hypothesis of zero correlation is clearly rejected.

The most general observation to draw from the coefficient estimates is that the effects of virtually all the socio-demographic variables are somewhat smaller in magnitude as compared with the estimates in Table 2, while the estimates on the urban form instruments are all considerably higher. In the case of the density measures, for example, it is seen that a unit increase in outlet density decreases the probability of car ownership by 0.04 and decreases the distance traveled by some 16 kilometers per 5-day week. In both cases, the magnitudes are over a three fold increase compared to the corresponding es-

Table 4

**Distance traveled, 2PM results using urban form instruments**

Outcome equation

| | 2PM1 | 2PM2 | 2PM3 | 2PM4 |
|---|---|---|---|---|
| Instruments | | | | |
| Outlet density | –16.133<br>(0.000) | | | |
| Road density | | –1.866<br>(0.000) | | |
| Outlet diversity | | | 36.928<br>(0.775) | |
| Walking minutes to public transit | | | | 10.565<br>(0.000) |
| Control variables | | | | |
| Number under 18 in household | 3.428<br>(0.092) | 3.646<br>(0.070) | 5.309<br>(0.010) | 4.748<br>(0.031) |
| Number over 64 in household | 10.207<br>(0.001) | 11.484<br>(0.000) | 11.536<br>(0.000) | 7.991<br>(0.021) |
| Number of working females | 7.114<br>(0.038) | 7.749<br>(0.022) | 7.501<br>(0.033) | 11.549<br>(0.003) |
| Number of working males | 19.511<br>(0.000) | 19.133<br>(0.000) | 20.438<br>(0.000) | 21.577<br>(0.000) |
| Number with college preparatory degree | 16.229<br>(0.000) | 16.009<br>(0.000) | 14.253<br>(0.000) | 17.973<br>(0.000) |
| Average income of zip code | –0.145<br>(0.407) | –0.073<br>(0.670) | -0.428<br>(0.367) | –0.452<br>(0.019) |
| Constant | 96.060<br>(0.000) | 108.744<br>(0.000) | 57.846<br>(0.636) | 27.339<br>(0.099) |
| Anderson LR statistic | 1003.169<br>(0.000) | 2022.602<br>(0.000) | 253.686<br>(0.000) | 123.382<br>(0.000) |
| Hansen J statistic | 2.038<br>(0.056) | 0.856<br>(0.836) | 22.363<br>(0.000) | 0.360<br>(0.948) |
| Number of obs. | | 5668 | | |

p-values in parentheses; year dummies not presented.

timates in Table 2; moreover, the instrumented outlet density variable is estimated to be highly significant. Large differences are also seen for the estimates of the effects of road density, particularly in the selector equation. A unit increase in this variable reduces the probability of car ownership by 0.04, a 10 fold increase in magnitude over the effect of the non-instrumented variable, and reduces distance traveled over the week by 1.93 kilometers, a roughly 0.62 larger effect than in the non-instrumented case.

Turning to the 2PM models in Table 4, we see a notably tight correspondence between the estimates of the urban form instruments and the corresponding estimates from the Heckit with instruments (with the exception of the diversity measure, which is insignificant in both models). This finding holds despite the apparent presence of sample selectivity in models H1 and H2 of

Table 3, as evidenced by the highly significant IVM. The results of the identification tests are also uniform across the models with the exception of the Hansen J statistic in model 2PM4, which fails to reject the validity of the instrument measuring the walking time to public transit. Furthermore, in contrast to the non-instrumented variant, the coefficient on this variable is highly significant and suggests that each additional walking minute increases non-work automobile travel by 10.7 kilometers.

Taken together, the results from Tables 2, 3 and 4 lead us to conclude that while sample selectivity is evident in the Heckit models of automobile ownership and use, the practical implications of this issue for the interpretation of the results ultimately depend on the calculation of the conditional marginal effects and on the use of instruments for the endogenous regressors. In particular, we find that when these two factors are accounted for, the differences between the Heckit and 2PM are, in fact, negligible. By contrast, the pronounced discrepancies between the non-instrumented and instrumented coefficient estimates suggest that endogeneity is a problem warranting closer consideration. Indeed, the seemingly inflated estimates of the instrumented variables cast some doubt on their accuracy, despite the generally supportive evidence of the identification tests. Lassen (2005), who documents similar findings of studies using instrumental variables (Dee 2004; Milligan et al. 2003), speculates that one cause may be measurement errors in the independent variables, which can induce attenuation bias in the estimate. To the extent that instruments mitigate the effects of measurement error, the estimates would be expected to increase.

Because the issue of instrument validity is central to the question of whether the urban form variables have a causal effect on automobile use, we implement a final diagnostic check using a technique developed by Angrist/Krueger (1995), referred to as the split-sample IV estimator. This estimator involves randomly dividing the data into two sub-samples of equal size. The second sub-sample is used to estimate the first stage regression, the parameters from which are then used to construct fitted values and the second stage estimates using the first sub-sample. Unlike conventional IV estimates, which are biased toward OLS, these estimates are biased toward zero. They can be corrected, however, by multiplication with a parameter measuring the proportional attenuation bias. This parameter is derived by taking the inverse of the coefficient from a regression of the endogenous regressor on its predicted value (using the first sub-sample). Angrist/Krueger show that a major advantage of this approach is to reduce the risk of misleading inferences arising from the finite sample bias that can plague two stage least squares. As Hall et al. (1996) further note, another advantage is that the method is not reliant on canonical correlation tests, which they argue to be of dubious value in assessing the relevance of the instruments.

Table 5

**Distance traveled, 2PM results using split-sample instrumental variables**
Outcome equation

| | Outlet density | Road density | Outlet diversity | Minutes to public transit |
|---|---|---|---|---|
| Split sample instrumented variable (SSIV) | –11.377<br>(0.004) | –1.335<br>(0.006) | 80.204<br>(0.584) | 6.864<br>(0.007) |
| Attenuation bias | 0.806<br>(0.000) | 0.992<br>(0.000) | 0.999<br>(0.000) | 0.850<br>(0.000) |
| Unibased Split sample instrumented variable (USSIV) | –14.121<br>(0.006) | –1.346<br>(0.009) | 80.317<br>(0.656) | 8.075<br>(0.015) |
| Number of observations | 2836 | 2836 | 2836 | 2836 |

p-values in parentheses; control variables and year dummies not presented.

Table 5 presents the results from the application of this method for each of the urban form variables using the 2PM specification. The first row of the table shows the coefficient estimate of the split-sample instrumented variable (SSIV), the second shows the proportional attenuation bias, and the third shows the unbiased coefficient (USSIV) obtained by multiplying the inverse of the value in row two with the value in row one. With the exception of the diversity measure, all of the coefficients in row one are highly significant but smaller in magnitude than the corresponding estimates from Table 4, as is expected since this estimator is biased toward zero. The estimates of the proportional attenuation bias range from 80% for the case of the outlet density instrument to 99% for the road density instrument, suggesting that the instruments are strong and that the downward bias is relatively insubstantial. The unbiased coefficients in row three are – again with the exception of diversity – all highly significant but somewhat lower in magnitude than the conventional IV estimates of Table 4. These results provide further support for ascribing urban form, as measured by outlet density, road density, and the distance to public transit, a causative interpretation.

## 5. Conclusions

This paper has investigated the link between urban form, automobile ownership, and non-work automobile travel using variants of the censored regression model and instrumental variables, an approach which allowed us to explore the sensitivity of the estimates to sample selectivity and endogeneity. Unlike much of the work to date, we find that the urban form variables are a significant determinant of automobile ownership and distance driven, a finding that holds even after controlling for the possibility that these variables are jointly determined by other factors influencing residential choice decisions. In this regard, our specification of the urban form instruments – based on socio-demographic and architectural features of the surrounding area – follows closely the work Boarnet/Sarmiento (1998). While their analysis finds little evidence of a causative link between automobile travel and land use patterns, the evidence presented here is generally supportive that such a link exists. The results are robust to the presence of sample selectivity, and – with the exception of diversity – the validity of the urban form instruments is confirmed by identification tests. Moreover, the instruments retained their significance under split sample estimation, albeit with coefficient estimates that are of somewhat lower magnitude.

As this is one of the few studies to be conducted on this issue in a European context, it would be of interest to see whether the qualitative findings presented here are corroborated by studies using other data sets from within Germany and other European countries. A particularly useful line of inquiry would focus on the extent to which the spatial scale of the data affects the results. Data constraints precluded such an analysis in the present study, but it is one that is required before specific transport policies can be made based on the link between urban form and automobile travel. As has been demonstrated in the literature on spatial interactions (Geoghegan et al. 1997; Irwin, Bockstael 2001), individuals may respond differently to landscape patterns depending on whether these patterns characterize the individual's immediate activity space or are spatially removed. It is conceivable, for example, that higher road density would decrease automobile dependency immediately surrounding the household by making services and amenities more accessible by alternative modes, while increasing dependency in non-adjacent zones by making destinations there and beyond easier to reach by car.

These considerations lead to two concrete proposals for future research. The first would involve applying the modeling techniques used in this study to data that is more spatially disaggregate, as for example that used by Siedentop et al. (2005). As the work from Boarnet/Sarimiento's (1998) demonstrates, spatial scale matters, and higher resolution data would allow for comparative analyses by affording the possibility to aggregate the data to a coarser resolution. The second extension would involve further exploration of the role of spatial interactions through the inclusion of spatially lagged variables in the specification. In the context of the present data, for example, widow metrics could be calculated using a GIS that measure the urban form attributes over a ring surrounding – and spatially removed from – the zip code in which the household resides.

Given the paucity of available evidence to support the hypothesis of a causative effect of urban form on transport, these extensions, together with the techniques applied here, hold promise for further isolating the magnitude of the effect and thereby providing a platform for targeting policy responses to automobile dependency.

## References

Angrist, J.D. and A.B. Krueger (1995), Split-Sample Instrumental Variables Estimates of the Return to Schooling. *Journal of Business and Economic Statistics* 13 (2): 225–235.

Belsley, D., E. Kuh, and R. Welsch (1980), *Regression diagnostics: Identifying influential data and sources of collinearity*. New York: John Wiley & Sons.

Bento, A.M., M.L. Cropper, A.M. Mobarak, and K. Vinha ( 2005), The Impact of Urban Spatial Structure on Travel Demand in the United States. *Review of Economics and Statistics* 87 (3): 466–478.

Boarnet, M.G. and S. Sarmiento (1998), Can Land-use Policy Really Affect Travel Behavior? A Study of the Link between Non-work Travel and Land-use Characteristics. *Urban Studies* 35 (7): 1155–1169.

BBR – Bundesministerium für Raumordnung, Bauwesen und Städtebau (ed.) (1993), *Raumordnungspolitischer Orientierungsrahmen. Leitbild für die räumliche Entwicklung der Bunderepublik Deutschland*. Bonn.

Crane, R. and R. Crepeau (1998), Does Neighborhood Design Influence Travel? A Behavioral Analysis of Travel Diary and GIS Data. *Transportation Research D* 3 (4): 225–238.

Dee, T. (2004), Are There Civic Returns to Education? *Journal of Public Economics* 88 (9): 1697–1720.

Dow, W.H. and E.C. Norton (2003), Choosing Between and Interpreting the Heckman and Two-Part Models for Corner Solutions. *Health Services and Outcomes Research Methodology* 4 (1): 5–18.

Dresden – Landeshauptstadt Dresden (ed.) (2002), *Integriertes Stadtentwicklungskonzept*. Dresden: Stadtplanungsamt.

Duan, N., W. Manning, C. Morris, and J. Newhouse (1984), Choosing Between the Sample-Selection Model and the Multi-part model. *Journal of Business and Economic Statistics*. 2: 283–289.

EC – Commission of the European Communities (ed.) (2001), *White Paper – European Transport Policy for 2010*: Time to Decide. http://europa.eu.int/comm/transport/white_paper/

Friedman, B., S. Gordon, and J. Peers (1994), Effect of Neotraditional Design on Travel Characteristics. *Transportation Research Record*, 1466: 63–70.

Geoghegan, J., L.A. Wainger, and N.E. Bockstael (1997), Spatial Landscape Indices in a Hedonic Framework: an Ecological Economics Analysis Using GIS. *Ecological Economics* 23 (3): 251–264.

Hall, A.R., G.D. Rudebusch, and D.W. Wilcox (1996), Judging Instrument Relevance in Instrumental Variables Estimation. *International Economic Review* 37 (2): 283–298.

Handy, S. (1996), Understanding the Link Between Urban Form and Nonwork Travel Behavior. *Journal of Planning Education and Research* 15 (3): 183–198.

Hautzinger, H. and W. Stock (2005), Wer Fährt Eigentlich Wie Viel? Volumen und Struktur der Fahrleistung Deutscher Kraftfahrzeuge 1993/2002. *Internationales Verkehrswesen* 57: 368–373.

Hay, J., Leu, R. and P. Rohrer (1987), Ordinary least squares and sample-selection models of health-care demand. *Journal of Business and Economic Statistics* 5: 499–506.

Heckman, J. (1979), Sample Selection Bias as a Specification Error. *Econometrica* 47: 153–161.

Holtzclaw, J. (1990), Explaining Urban Density and Transit Impacts on Auto Use. Report to Natural Resources Defense Council. San Francisco: Sierra Club.

Irwin, E.G. and N.E. Bockstael (2002), Interacting Agents, Spatial Externalities, and the Endogenous Evolution of Residential Land Use Pattern. *Journal of Economic Geography* 2 (1): 31–54.

Jargowsky, P.A. (2001), Sprawl, Concentration of Poverty, and Urban Inequality. In G. Squires. (ed.), *Urban Sprawl: Causes, Consequences & Policy Responses*. Washington, DC: Urban Institute Press.

Krizek, K.J. (2003), Neighborhood Services, Trip Purpose, and Tour-Based Travel. *Transportation* 30 (4): 387–410.

Kunstler, J.H. (1994), *The Geography of Nowhere: The Rise and Decline of America's Man-Made Landscape*. New York: Touchstone.

Lassen, D. (2005), The Effect of Information on Voter Turnout: Evidence from a Natural Experiment. *American Journal of Political Science* 49 (1): 103–118.

Leung, S.F. and S. Yu (1996), On the choice between sample selection and two-part models. *Journal of Econometrics* 72: 197–229.

Manning, W., N. Duan, and W. Rogers (1987), Monte Carlo Evidence On the Choice Between Sample Selection and Two-part models. *Journal of Econometrics* 35: 59–82.

Milligan, K., E.N. Moretti, and P. Oreopoulos (2004), Does Education Improve Citizenship? Evidence from the U.S. and the U.K. *Journal of Public Economics*. 88 (9–10): 1667–1695.

Newey, W. (1987), Efficient Estimation of Limited Dependent Variable Models with Endogenous Explanatory Variables. *Journal of Econometrics* 36 (3): 231–250.

Newman, P. and J. Kenworthy (1989), *Cities and Automobile Dependence: An International Sourcebook*. Aldershot: Gower.

Saha, A., O. Capps, and P. Byrne (1997), Calculating Marginal Effects in Models for Zero Expenditures in Household Budgets Using a Heckman-type correction. *Applied Economics* 29: 1311–1316.

Shannon, C.E. and Weaver, W. (1949), *The Mathematical Theory of Communication*. Urbana: University of Illinois Press.

Siedentop, S., Stein, S., Wolf, U., Lanzendorf, M. and Hesse, M. (2005), *Mobilität im suburbanen Raum. Neue verkehrliche und raumordnerische Implikationen des räumlichen Strukturwandels*. Dresden et al.

Sigelman, L., and , L. Zeng. (1999), Analyzing Censored and Sample-Selected Data with Tobit and Heckman Models. *Political Analysis* 8: 167–182.

Vance, C. (2006), Marginal Effects and Significance Testing with Heckman's Sample Selection Model: A Methodological Note. RWI: Discussion Paper 39. RWI Essen: http://www.rwi-essen.de/dp.