Submitted for *Transportation Research Part D: Transport and Environment*

Influences of built environment characteristics and individual factors on commuting distance: a multilevel mixture hazard modeling approach

Chuan Ding 1, Sabyasachee Mishra 2, Guangquan Lu 3, Jiawen Yang 4, Chao Liu 5

1 Assistant Professor, School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure System and Safety Control, Beihang University, Beijing 100191, China, Email: cding@buaa.edu.cn

2 Assistant Professor, Department of Civil Engineering, University of Memphis, Memphis 38152, United States, Email: smishra3@memphis.edu

3 Associate Professor, School of Transportation Science and Engineering, Beijing Key Laboratory for Cooperative Vehicle Infrastructure System and Safety Control, Beihang University, Beijing 100191, China, Email: lugq@buaa.edu.cn

4 Professor, School of Urban Planning and Design, Peking University, Shenzhen 518055, China, Email: yangjw@pkusz.edu.cn

5 Faculty Research Associate, National Center for Smart Growth Research, University of Maryland, MD 20742, United States, Email: cliu8@umd.edu

*Corresponding author: Chuan Ding, Address: New Main Building H-1101, Beihang University, 37 Xueyuan Road, Haidian District, Beijing 100191, China. Tel. /fax: +86-18518790913 Email address: cding@buaa.edu.cn*
Influences of built environment characteristics and individual factors on commuting distance: a multilevel mixture hazard modeling approach

Abstract

Concerns over transportation energy consumption and green-household gas (GHG) emissions have prompted a growing body of research into the influence of built environment on travel behavior. Studies on the relationship between land use and travel behavior are often at some aggregated spatial unit such as traffic analysis zone (TAZ), spatial issues occurs among individuals clustered within a zone because of locational effects. However, recognition of spatial issue in travel modeling was not sufficiently investigated. The object of this study is twofold. First, a multilevel hazard model was applied to accommodate the spatial context in which individuals generate commuting distance. Second, this research provides additional insights into examine the effects socio-demographics and built environment on commuting distance. Using Washington metropolitan area as the case, the built environment measures were calculated for each TAZ. To estimate the model parameters, the robust maximum likelihood estimation method for a partial function was used, and the model results confirmed the important roles played by the neighborhood and individual level factors in influencing commuting distance. Meanwhile, a comparison among the general multilevel model, single level and multilevel hazard models was conducted. The results suggest that application of the multilevel model obtains significant improvements over traditional model. The significant spatial heterogeneity parameter indicates that it is necessary to accommodate the spatial issues in the context of commuting distance. The results are expected to give urban planners and policy makers a better understanding on how the neighborhood and individual level factors influence the commuting distance, and consequently develop effective and targeted countermeasures.

Keywords

Built environment; Commuting distance; Land use; Spatial heterogeneity; Multilevel hazard model
Influences of neighborhood and individual level factors on commuting distance: a multilevel mixture hazard modeling approach

1 Introduction

Concerns over transportation energy consumption and green-household gas (GHG) emissions have prompted a growing body of research into the influences of built environments on travel behavior, especially vehicle miles traveled (VMT). In recent years, a growing body of literature has been conducted to examine the relationships between influential factors and travel behavior (Boarnet, 2011; Ewing and Cervero, 2001, 2010). Understanding the effects of built environment measures on travel distance can help planning agencies develop effective countermeasures to alter the car use to green travel, thereby reduce energy consumption and transport emission. However, the debate on the influences of various built environment measures on travel behavior is far away from reaching the consensus due to the confounding factors (e.g. socio-demographics), empirical contexts, geographical scale, and residential self-selection (Cao, 2015; Ding et al., 2014a; Antipova et al., 2011; Mokhtarian and Cao, 2008; Bhat and Guo, 2007; Limtanakool et al., 2006; Handy et al., 2005). As an important component of daily travel, commuting (i.e. home-to-work) trip is one of the major sources of traffic congestion and air pollution all over the world. Relative to travel outcomes of trip frequency, mode choice, and VMT, commuting trip distance has gained less attention (Manaugh et al., 2010).

In the field of transportation and urban planning, the study of the duration of travel behavior has recently obtained growing attention. Hazard-based modeling approach has been regarded as a powerful method for capturing the duration probability and dependence, therefore this method has been extensively used the fields of departure time choice (Hou et al., 2014; Gadda et al., 2009; Bhat and Steed, 2002) and traffic incident duration analysis (Hou et al., 2014; Hojati et al., 2013; Nam and Mannering, 2000). However, the application of hazard-based duration modeling approach in the fields of travel time (Guo et al., 2012; Yang et al., 2012) and travel distance (Anastasopoulos, 2012) is limited (van den Berg et al., 2012). Especially for the travel distance, it is typically considered as a travel outcome rather than a process, and hence the duration dependence is often ignored. Anastasopoulos et al. (2012) identified important factors that determined activity-based travel distance in urban area using the hazard-based modeling approach, and their model results empirically illustrated the applicability of the hazard-based modeling approach to spatial travel distance.

Most researches on the relationship between built environment and travel behavior are often at a certain aggregated spatial unit such as traffic analysis zone (TAZ), census tract, or the zip code level. Spatial issues (i.e. spatial dependency, spatial heterogeneity, and spatial heteroscedasticity) occur among individuals clustered within a zone because of locational effects. In general, ignoring accurate measure of spatial issues can result in inconsistent parameters estimation (Bhat and Zhao, 2002). To solve this problem, the multilevel modeling framework has been employed to the relationship analysis between built environment and travel behavior in recent years (Ding et al., 2014a, 2014b, 2016; Nasri and Zhang, 2014; Hong et al., 2014; Hong and Goodchild, 2014; Hong and Shen, 2013; Antipova et al., 2011). The empirical researches conducted by Ding et al. (2014a, 2014b, 2016) and Hong et al. (2014) showed that ignoring spatial relationship among TAZs could result in erroneous conclusions. Although previous studies highlight the need to accommodate the spatial context in which individuals make the duration travel decisions, the recognition of spatial issues in travel modeling is not still sufficiently investigated, which might be relevant to the complicated estimation for the multilevel model, relative to conventional methods.
Using different travel outcome dimensions, spatial scales, modeling approach, and estimation techniques, a substantial body of studies has examined the impacts of built environment on travel behavior. However, to our knowledge, limited efforts have been made to account for the spatial issues in the associated connection analysis between built environment and commuting distance using hazard-based modeling approach. The aim of this study is twofold. First, a multilevel mixture hazard model was applied to accommodate the spatial context in which individuals make the duration travel decisions (i.e. commuting distance). Second, additional insights for a better understanding the influences of socio-demographics and built environment (including both individual and neighborhood level attributes) on commuting distance were provided. Incorporating multilevel framework into hazard-based model not only reveals the influences of built environment on commuting distance, but also recognizes the spatial heterogeneity across residential location zones. Meanwhile, a comparison among the general multilevel model, single level hazard model and multilevel hazard model was conducted. This study is expected to provide additional insights into the linkage between built environment and commuting distance.

The remainder of this paper is organized as follows. The next section presents the modeling approach used in this study. Data sources and description is described in the third section. Then the proposed model is applied and empirical model results are discussed in the following section. The final section provides the conclusions and future studies.

2 Method

2.1 Hazard-based modeling approach

Taking the commuting distance $D$ as a random variable, let $f(d)$ be the probability density function of commuting distance. Then, according to the probability theory, the cumulative distribution function $F(d)$ represents the probability the random variable $D$ takes from distance zero to distance $d$ can be expressed as:

$$ F(d) = P(D \leq d) = \int_0^d f(u)du $$

(1)

The survival function and hazard function are the fundamental concepts of hazard-based modeling approach. Defined as the probability of the commuting process surviving until distance $d$, the survival function at commuting distance $d$ can be written as:

$$ S(d) = P(D > d) = 1 - P(D \leq d) = 1 - F(d) $$

(2)

According to the definition, given $d \to \infty$, $S(0)=1$ and $S(\infty)=0$. The hazard function at commuting distance $d$ is defined as the instantaneous rate of an end-of-duration occurrence (e.g. the commuting travel being finished), generally denoted by $h(t)$ and mathematically can be expressed as:

$$ h(d) = \lim_{\Delta d \to 0} \frac{P(d \leq D < d + \Delta d)}{P(D \geq d)\Delta d} = \frac{f(d)}{S(d)} $$

(3)

The hazard will fluctuate over the commuting distance if there is a relationship between the likelihood of the commuting duration ending soon and the distance that has traveled. Otherwise, the conditional probability of the commuting duration is independent with distance (i.e. the hazard will be constant). Therefore, the concept of hazard reflects the commuting duration dependence. In hazard-based modeling approach, regression models are used for analyzing the hazard function trend as well as the impacts of related covariates (i.e. influential factors) on the hazard. According to the formulation of baseline hazard function, full-parametric and semi-parametric
regression models are the two most widely used methods (see Appendix for details).

A problem with the full-parametric modeling method is that it will inconsistently estimate the hazard function when the assumed parametric distribution is incorrect. With regards to the semi-parametric modeling method, it is an optimal method that yields valid estimates of covariate effects on the hazard function while avoiding the specification of an underlying distributional function (Hou et al., 2014; Liu, 2012). The Cox proportional hazard model is the most widely used semi-parametric modeling method with the assumption that the covariates are multiplicatively related to the hazard function. Suppose the individual \( m \) and \( n \) with the covariates \( X_m \) and \( X_n \), respectively, the hazard ratio (HR\(_m\)) of covariate X can be expressed as:

\[
HR_m = \frac{h_m(d)\exp(\beta^T X_m)}{h_n(d)\exp(\beta^T X_n)} = \exp[(X_m - X_n)\beta^T] \tag{4}
\]

Independent of unspecified baseline hazard function, an uncomplicated indicator to measure the effect of a given covariate on the hazard rate can be provided by the hazard ratio (Liu, 2012). If a specific independent factor is a dichotomous variable, the hazard ratio displays an intuitive meaning of the relative risk. For example, if the estimated parameter for the factor of gender (1=male, 0=female) on commuting distance is 1.60, thereby the hazard ratio is \( \exp(1.60)=4.95 \), suggesting that male’s commuting distance is about 5 times as female’s. For a continuous independent variable, the hazard ratio displays the extent to which the risk increases (HR>1) or decrease (HR<1) with a unit increase in the value of that independent variable.

Compared with the full-parametric modeling method, the semi-parametric modeling method provides two main advantages: first, exploring the effect of covariates on hazard with an unspecified baseline hazard function, to avoid the estimated biases caused by the improper distribution assumption; second, separating the effect of covariates from the baseline hazard function reflecting the population heterogeneity, to make explicit understanding on the factor effects. Therefore, the semi-parametric modeling method are selected to develop for modeling the commuting distance and explore the effects of influential factors in this study.

### 2.2 Model development

Let \( D_j \) be the commuting distance for the individual \( i \) (\( i=1, 2, \ldots, I \)) living in residence zone \( j \) (\( j=1, 2, \ldots, J \)), \( X_q \) represents the individual-associated attributes (e.g. age, gender, occupation, car ownership), \( Z_j \) represents the neighborhood level factors (i.e. built environment attributes) measured at the TAZ scale. To investigate the influences of neighborhood and individual level factors on commuting distance and simultaneously identify the unobserved spatial heterogeneity of commuting distance across TAZs, the multilevel Cox proportional hazard mixture model is formulated in this study. For separating the effects of neighborhood and individual level factors on commuting distance, the multilevel framework is incorporated into the Cox proportional hazard model. The hazard function of the multilevel hazard model can be expressed as:

\[
h(d \mid X_q, Z_j) = h_0(d)g(\beta^T X_q)h(\gamma^T Z_j) \tag{5}
\]

where \( h_0(d) \) represents the baseline hazard function, \( g(\beta^T, X_q) \) and \( f(\gamma^T, Z_j) \) indicates the effect of individual level and neighborhood level factors on the hazard, respectively. \( \beta^T \) and \( \gamma^T \) are the corresponding coefficients on the different levels. In this study, the forms of \( g(\beta^T, X_q) \) and \( f(\gamma^T, Z_j) \) can be expressed as:

\[
g(\beta^T, X_q) = \exp(\beta^T X_q) \tag{6}
\]
\[ f(\gamma^T, Z_j) = \exp(\gamma^T Z_j + \xi_j) \]  

(7)

Conditional on \( \xi_j \) term, the multilevel Cox proportional hazard mixture model can be written as:

\[ h(d \mid X_{ij}, Z_j) = h_0(d) \exp(\beta^T X_{ij} + \gamma^T Z_j + \xi_j) \]  

(8)

where \( \xi_j \) is an unobserved random term assumed to be normally distributed with mean zero and variance \( \sigma^2 \). The variance of \( \xi_j \) can measure the spatial heterogeneity of the commuting distance across TAZs.

Integrating multilevel framework and traditional hazard-based model, the multilevel Cox proportional hazard mixture model can allow for spatial variation across the residential zones while accounting for the within zone homogeneous correlations among individual characteristics. To accommodate the nested structure of the data related to individuals nested within neighborhood zones, a TAZ clustering variable is included in the model. The multilevel hazard modeling framework is described in Figure 1. To estimate the model parameters, the robust maximum likelihood (MLR) estimation method for a partial likelihood function, extended from the Cox method, is used in this study. All estimations and computation are carried out by using the M-plus software package.

![Multilevel hazard modeling framework for commuting distance](image)

**3 Data sources and description**

In order to disentangle the influence of neighborhood and individual level factors on commuting distance, an integrated dataset that includes built environment variables as well as travel information is crucial for this study. The Washington D.C. metropolitan area is used as the case, as shown in Figure 2. Washington metropolitan area is one of the densest metropolises all over the United States, including relatively extensive public transit system, such as metro, commuter rail, and regular bus (Mishra et al., 2012, 2015). In this context,
a better understanding the influences of built environment on travel behavior is required. Land use data were collected from three major sources: the Metropolitan Washington Council of Governments (MWCOG), Maryland Department of Planning, and the US Census. The Washington metropolitan area includes the following one city and eighteen countries: District of Columbia, nine countries in Maryland (e.g. Montgomery Country, Prince Georges Country), and nine countries in Virginia (e.g. Arlington County, Fairfax Country). The whole region covers approximately 5500 mi² of land and accommodates over 5.8 million people.

During the year 2007-2008, the Transportation Planning Board (TPB) at the MWCOG, teamed with the Baltimore Metropolitan Council (BMC) on behalf of the Baltimore Regional Transportation Board, conducted a household travel survey in both the Washington and Baltimore Washington regions (NCRTPB-MWCOG, 2010). Data for the travel survey was collected from randomly selected households in the Washington metropolitan region. Each household completed a travel diary that documented the activities of all household members on an assigned day, including how, where, when, and why people travel in the region. Demographic information (e.g. individual and household attributes) was also collected. The surveys have been stored in a database, which contains records for approximately 11,000 households, 25,000 persons, 132,000 trips, and 16,000 vehicles.
Figure 2 Washington metropolitan regional household survey areas

Using the land use data collected, the built environment of residential neighborhood were measured at the geographic scale of TAZ, including residential density, employment density, land use mix, average block size, and distance from CBD (Ewing and Cervero, 2001, 2010). Density variable represents the level of sprawl and activity. Land use mix quantified the degree of balance across different land use types, representing the accessibility to various destinations. The variable of average block size represents the street network characteristic within the neighborhood. Distance from CBD measures the spatial centrality of residential location. These five measures were calculated for the total 2191 TAZs in ArcGIS 10.0 based on the method that was directly taken from author previous work (Ding et al., 2014a). Table 1 presents the descriptive statistics for built environment factors in the case area. The GIS shape files of TAZ were used for spatially processing the datasets and integrating the built environment measures into the travel survey records. Specially, the entropy index that indicates the extent of mixed land development was computed based on the four different land use
categories at the TAZ level, as shown below (Ewing and Cervero, 2010).

\[
\text{Land use mix} = - \left\{ \sum_{i=1}^{n} (P_i \ln P_i) \right\} / \ln(n) 
\]

(9)

where \( P_i \) is the proportion of land use type \( i \) found in the total TAZ land use, and \( n \) is the number of different land use types (residential, service, retail, and other). The value of land use mix index ranges from 0 to 1 and captures how evenly the residential, service, retail, and other areas are distributed within the TAZ. Higher values of land use mix indicate a more balanced land use pattern.

Table 1 Descriptive statistics of built environment factors and individual level factors

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Variable Description</th>
<th>Mean</th>
<th>St. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Socio-demographic factors at individual level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>Age of respondents in years</td>
<td>44.39</td>
<td>12.648</td>
</tr>
<tr>
<td>Gender</td>
<td>Male (1=yes)</td>
<td>0.52</td>
<td>0.500</td>
</tr>
<tr>
<td>Race</td>
<td>Caucasian (1=yes)</td>
<td>0.74</td>
<td>0.438</td>
</tr>
<tr>
<td>Occupation</td>
<td>Person works in a government agency (1=yes)</td>
<td>0.38</td>
<td>0.485</td>
</tr>
<tr>
<td>Household size</td>
<td>Number of persons in household</td>
<td>2.58</td>
<td>1.268</td>
</tr>
<tr>
<td>Household students</td>
<td>Number of students in household</td>
<td>0.70</td>
<td>0.985</td>
</tr>
<tr>
<td>Household workers</td>
<td>Number of workers in household</td>
<td>1.77</td>
<td>0.692</td>
</tr>
<tr>
<td>Number of vehicles</td>
<td>Number of household vehicles available</td>
<td>2.06</td>
<td>1.081</td>
</tr>
<tr>
<td>Household income</td>
<td>Income1: household income is less than $40,000 (1=yes)</td>
<td>0.08</td>
<td>0.264</td>
</tr>
<tr>
<td></td>
<td>Income2: household income is between $40,000 and $125,000 (1=yes)</td>
<td>0.59</td>
<td>0.492</td>
</tr>
<tr>
<td></td>
<td>Income3: household income is equal to or more than $125,000 (1=yes)</td>
<td>0.33</td>
<td>0.471</td>
</tr>
<tr>
<td>Travel mode</td>
<td>Commuting trip by car (1=yes)</td>
<td>0.78</td>
<td>0.413</td>
</tr>
<tr>
<td><strong>Built environment factors at TAZ level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residential density</td>
<td>Population/Area size (persons/ acre)</td>
<td>7.209</td>
<td>10.826</td>
</tr>
<tr>
<td>Employment density</td>
<td>Employment/ Area size (jobs/ acre)</td>
<td>12.421</td>
<td>51.552</td>
</tr>
<tr>
<td>Land use mix (entropy)</td>
<td>Mixture of residential, service, retail, and other employment land use types</td>
<td>0.418</td>
<td>0.242</td>
</tr>
<tr>
<td>Average block size</td>
<td>Average block size within TAZ (sq. mi.)</td>
<td>0.191</td>
<td>0.280</td>
</tr>
<tr>
<td>Distance from CBD</td>
<td>Straight line distance from CBD (mile)</td>
<td>15.716</td>
<td>11.898</td>
</tr>
</tbody>
</table>

*Note: 8327 persons, 1340 residential zones.*

In this study, the dependent variable is commuting distance (i.e. travel distance from residential location to workplace) that was obtained from the household travel survey. Travel distance for each trip was reported by the respondents on the survey day. After removing the cases with missing data and the respondents who are less than 16 years old, the final sample comprises 8327 respondents. In the sample, 78.2% of the respondents use automobile as their primary commuter mode and 21.8% use public transit. The distribution of commuting distance is shown in Figure 3. The mean of the commuting distance is 12.8 miles and the standard deviation is 11.863 miles. These commuting trips originate at 1340 residential zones. The average age of the sample is about 44 years old. 52% of the respondents are male and 74% are Caucasian people. The average household size is 2.58 and almost 19% of the respondents live alone. The average number of household workers is 1.77 and the average household car ownership is 2.06. Only 3.7% of samples have no vehicles and nearly 70% owned at least two vehicles. The detailed descriptive statistics of sample data is shown in Table 1.
4 Model results

Using the partial MLR estimation method, the results of the general multilevel model, single level and multilevel hazard models were estimated, including the estimated coefficient, hazard ratio, $t$-statistic, and $p$-value for each independent variable. For the hazard-based modeling approach, the estimated coefficients have the interpretation that the log hazard ratio is the change in the dependent variable from a one unit change on the independent variable all else held constant. For the hazard modeling approach, a lower hazard means a longer commuting distance. The estimated coefficients with a positive sign indicate that as the value of the independent variable increases so does the distance hazard, therefore a shorter commuting distance. Likewise, negative coefficient values result in a lower distance hazard, therefore a longer commuting distance. In this section, the model results are discussed in detail based on the individual level and TAZ level factors.

The results of the general multilevel model, single level hazard model and multilevel mixture hazard model are shown in Table 2, Table 3 and Table 4, respectively. Three goodness-of-fit indices were used to assess the model fit, including log-likelihood value, Akaike information criterion (AIC), and Bayesian information criterion (BIC). Higher value of log-likelihood and lower values of AIC and BIC identify a better fitting model. Compared the model fit information among the three models, the goodness-of-fit indices suggested that the multilevel mixture hazard model was the better fitting model. The multilevel mixture hazard model outperforms the single level hazard model, which in turn offers significant improvements over the general multilevel model. Specifically, the log-likelihood value of the multilevel mixture hazard model was 83.342 points higher than the...
single level hazard model. In turn, the log-likelihood value of the single level hazard model was 3,178.286 points higher than the general multilevel model. The values of AIC and BIC of the multilevel mixture hazard model were 164.681 and 157.654 points lower than the single level hazard model, respectively. In turn, the values of AIC and BIC of the single level hazard model were 6,362.573 and 6,323.275 points lower than the general multilevel model, respectively. We can see that it is very necessary to apply the hazard-based modeling approach to capture the commuting duration probability and dependence.

Table 2 General multilevel model results for the commuting distance

<table>
<thead>
<tr>
<th>Variables</th>
<th>General multilevel model</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficient</td>
<td>t-statistic</td>
<td>p-value</td>
</tr>
<tr>
<td><strong>Socio-demographic factors at individual level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>-0.002</td>
<td>-0.212</td>
<td>0.832</td>
</tr>
<tr>
<td>Gender</td>
<td>2.526</td>
<td>10.618</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Race</td>
<td>-1.607</td>
<td>-5.393</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Occupation</td>
<td>0.007</td>
<td>0.027</td>
<td>0.979</td>
</tr>
<tr>
<td>Household size</td>
<td>-0.317</td>
<td>-1.708</td>
<td>0.088</td>
</tr>
<tr>
<td>Household students</td>
<td>0.531</td>
<td>2.355</td>
<td>0.019</td>
</tr>
<tr>
<td>Household workers</td>
<td>-1.011</td>
<td>-3.925</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Number of vehicles</td>
<td>0.769</td>
<td>4.417</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Household income-1</td>
<td>-2.810</td>
<td>-6.081</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Household income-3</td>
<td>1.650</td>
<td>5.212</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Travel mode</td>
<td>-2.393</td>
<td>-6.295</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Built environment factors at TAZ level</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residential density</td>
<td>-0.050</td>
<td>-4.633</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Employment density</td>
<td>-0.004</td>
<td>-0.770</td>
<td>0.441</td>
</tr>
<tr>
<td>Land use mix</td>
<td>-2.889</td>
<td>-3.930</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Average block size</td>
<td>2.759</td>
<td>2.084</td>
<td>0.037</td>
</tr>
<tr>
<td>Distance from CBD</td>
<td>0.318</td>
<td>12.923</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Spatial heterogeneity parameter across zones</strong></td>
<td>10.825</td>
<td>11.782</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>( \sigma^2_n )</td>
<td>9.394</td>
<td>6.204</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

\[
\begin{align*}
\sigma^2_n & = 10.825 \quad 11.782 \quad <0.001 \\
\sigma^2_h & = 9.394 \quad 6.204 \quad <0.001 
\end{align*}
\]

**Model fit information**

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-likelihood</td>
<td>-31383.847</td>
<td></td>
</tr>
<tr>
<td>Number of free parameters</td>
<td>19</td>
<td></td>
</tr>
<tr>
<td>Akaike information criterion (AIC)</td>
<td>62805.694</td>
<td></td>
</tr>
<tr>
<td>Bayesian information criterion (BIC)</td>
<td>62939.211</td>
<td></td>
</tr>
</tbody>
</table>

*Note*: sample size is 8327.

Comparing the model results between the single level and multilevel hazard models, the differences between the two model results indicate how accommodating the spatial context in which individuals make travel behavior can influence the model estimates. For example, the factor of average block size was found to have negative effect on the distance hazard at the 95% level in the single level hazard model. However, this effect was found to be statistically significant at the 90% level in the multilevel hazard model. This finding is consistent with the results from the studies conducted by Ding et al. (2014a, 2016) and Hong et al. (2014):
single level model estimation tends to produce more statistically significant variables. Due to the fact that single level model ignores the within-zone observation dependence, and thus violates the basic assumption underlying traditional model (i.e. independent observations, homoscedasticity). As a result, standard errors of parameter estimates would be biased downwards, resulting in a large Type I error and falsely rejecting a true null hypothesis in statistical significance testing (Snijder and Bosker, 2012). Consequently, analyzing multilevel data with traditional model can produce misleading conclusions. As shown in Table 4, the spatial heterogeneity parameter across zones is statistically significant with positive sign at the 95% level, indicating that there is an important spatial variation of commuting behavior at the TAZ scale, and a multilevel modeling approach is necessary in this study. Therefore, the subsequent discussion of the influences of neighborhood and individual level factors on commuting distance was based on the multilevel mixture hazard model.

Table 3 Single level hazard model results for the commuting distance

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient</th>
<th>Hazard ratio</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Socio-demographic factors at individual level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.001</td>
<td>1.001</td>
<td>1.004</td>
<td>0.315</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.238</td>
<td>0.788</td>
<td>-9.872</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Race</td>
<td>0.175</td>
<td>1.191</td>
<td>6.581</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Occupation</td>
<td>0.002</td>
<td>1.002</td>
<td>0.085</td>
<td>0.932</td>
</tr>
<tr>
<td>Household size</td>
<td>0.044</td>
<td>1.045</td>
<td>2.816</td>
<td>0.005</td>
</tr>
<tr>
<td>Household students</td>
<td>-0.057</td>
<td>0.945</td>
<td>-3.137</td>
<td>0.002</td>
</tr>
<tr>
<td>Household workers</td>
<td>0.050</td>
<td>1.051</td>
<td>2.146</td>
<td>0.032</td>
</tr>
<tr>
<td>Number of vehicles</td>
<td>-0.080</td>
<td>0.923</td>
<td>-5.132</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Household income-1</td>
<td>0.238</td>
<td>1.269</td>
<td>4.445</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Household income-3</td>
<td>-0.106</td>
<td>0.899</td>
<td>-4.269</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Travel mode</td>
<td>-0.006</td>
<td>0.994</td>
<td>-0.228</td>
<td>0.819</td>
</tr>
<tr>
<td><strong>Built environment factors at TAZ level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residential density</td>
<td>0.009</td>
<td>1.009</td>
<td>5.810</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Employment density</td>
<td>0.001</td>
<td>1.001</td>
<td>0.242</td>
<td>0.809</td>
</tr>
<tr>
<td>Land use mix</td>
<td>0.184</td>
<td>1.202</td>
<td>3.230</td>
<td>0.001</td>
</tr>
<tr>
<td>Average block size</td>
<td>-0.220</td>
<td>0.803</td>
<td>-2.899</td>
<td>0.004</td>
</tr>
<tr>
<td>Distance from CBD</td>
<td>-0.028</td>
<td>0.972</td>
<td>-17.628</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>

**Model fit information**

| Log-likelihood                     | -28205.561  |
| Number of free parameters          | 16          |
| Akaike information criterion (AIC) | 56443.121   |
| Bayesian information criterion (BIC)| 56555.558   |

Note: sample size is 8327.

Results from the multilevel mixture hazard model suggest that with respect to the socio-demographic factors at individual level, the factors of gender, race, household size, number of household students, workers, vehicles, and household income were found to have significant effects on commuting distance at the 95% level. Specifically, the factor of male was found to significantly decrease the distance hazard, compared with the baseline of female. This indicates that males were more likely to travel longer distance than females. This
finding is consistent with most previous studies (Van Acker and Witlox, 2011; Liu and Shen, 2011; Manaugh et al., 2010). The gender difference is partly due to the fact that females remain primarily responsible for household maintenance tasks so that they more likely to choose workplace near home location. The factor of race was found to have a positive effect on distance hazard, indicating that the Caucasians tended to travel shorter distance to work, compared with other races. This may be due to the fact that Caucasians are more likely to live near to the workplace such as the CBD. The factors of household size and household workers were found to be associated with positive signs and increasing distance hazard. This finding coincides with the research conducted by Manaugh et al. (2010). The main reason may be that as the household size and number of workers increase, they are more likely to choose residential location near to the employment center. The factor of number of vehicles was found to have a negative effect on the distance hazard. This result is consistent with the research conducted by Anastasopoulos et al. (2012) and Cao (2009). The factors of low and high household income showed that as the household income increases, the commuting distance increases. This finding is consistent with most previous studies (Van Acker and Witlox, 2011; Manaugh, 2010). The main reason is that high income commuters are more likely to reside in suburban areas with longer distance to the employment center. It should be noted that the factor of travel mode on the commuting distance didn’t show a significant effect. This may be due to fact that most commuter use car as their primary mode in the case area. Meanwhile, metro and commuter rail can meet their needs of traveling long commuting distance to work.

Table 4 Multilevel mixture hazard model for the commuting distance

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient</th>
<th>Hazard ratio</th>
<th>t-statistic</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Socio-demographic factors at individual level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Age</td>
<td>0.001</td>
<td>1.001</td>
<td>1.189</td>
<td>0.235</td>
</tr>
<tr>
<td>Gender</td>
<td>-0.258</td>
<td>0.773</td>
<td>-10.374</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Race</td>
<td>0.167</td>
<td>1.182</td>
<td>5.371</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Occupation</td>
<td>0.009</td>
<td>1.009</td>
<td>0.347</td>
<td>0.729</td>
</tr>
<tr>
<td>Household size</td>
<td>0.042</td>
<td>1.043</td>
<td>2.331</td>
<td>0.020</td>
</tr>
<tr>
<td>Household students</td>
<td>-0.057</td>
<td>0.945</td>
<td>-2.636</td>
<td>0.008</td>
</tr>
<tr>
<td>Household workers</td>
<td>0.065</td>
<td>1.067</td>
<td>2.555</td>
<td>0.011</td>
</tr>
<tr>
<td>Number of vehicles</td>
<td>-0.070</td>
<td>0.932</td>
<td>-4.137</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Household income-1</td>
<td>0.295</td>
<td>1.343</td>
<td>5.245</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Household income-3</td>
<td>-0.115</td>
<td>0.891</td>
<td>-4.011</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Travel mode</td>
<td>0.025</td>
<td>1.025</td>
<td>0.0671</td>
<td>0.502</td>
</tr>
<tr>
<td><strong>Built environment factors at TAZ level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residential density</td>
<td>0.009</td>
<td>1.009</td>
<td>4.957</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Employment density</td>
<td>0.001</td>
<td>1.001</td>
<td>0.714</td>
<td>0.475</td>
</tr>
<tr>
<td>Land use mix</td>
<td>0.269</td>
<td>1.309</td>
<td>3.529</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td>Average block size</td>
<td>-0.155</td>
<td>0.856</td>
<td>-1.657</td>
<td>0.087</td>
</tr>
<tr>
<td>Distance from CBD</td>
<td>-0.031</td>
<td>0.969</td>
<td>-12.168</td>
<td>&lt;0.001</td>
</tr>
<tr>
<td><strong>Spatial heterogeneity parameter across zones</strong></td>
<td>0.094</td>
<td>1.099</td>
<td>6.272</td>
<td>&lt;0.001</td>
</tr>
</tbody>
</table>
The estimated results show that higher degree level of land use mix corresponds to shorter commuting distance, which is consistent with the research conducted by Antipova et al. (2011). This finding is mainly due to the fact that the intermingling of residences, jobs, shops, and recreational facilities in a compact urban environment induce people to carry out their daily activities within a much smaller geographical area. As a measure of street network connectivity, the factor of average block size was found to have a negative relationship with distance hazard with a moderate significant level. This finding is consistent with the relationship research between land use and VMT conducted by Nasri and Zhang (2014). Generally speaking, a smaller block size indicates a better street network connectivity, therefore the distance to destination would be shorter. The factor of distance from CBD was also found to decrease the distance hazard, indicating that commuters living further away from CBD tended to travel longer distance to workplace. This may be because that most jobs opportunities were located close to CBD, thereby to certain extent further away from CBD indicates further away from workplace.

To obtain the relative importance of individual level and neighborhood level factors influencing commuting distance, the standardized effect for each factor are computed using the coefficient estimates of multilevel mixture hazard model, as shown in Figure 4. The standardized effect is the mean change in standard deviation units of commuting distance hazard for a one standard deviation change in the independent variable. When comparing the magnitudes of the standardized effects, the sensitivity of each factor influencing commuting distance hazard can be gained. Overall, the standardized effects of the socio-demographic factors are greater than those of the built environment attributes. This result is consistent with the research conducted by Pinjari et al. (2008). Among the socio-demographic factors at individual level, gender was found to have the highest

<table>
<thead>
<tr>
<th>Model fit information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log-likelihood</td>
</tr>
<tr>
<td>Number of free parameters</td>
</tr>
<tr>
<td>Akaike information criterion (AIC)</td>
</tr>
<tr>
<td>Bayesian information criterion (BIC)</td>
</tr>
</tbody>
</table>

*Note: sample size is 8327.*
standardized effect ($\beta = -0.677$), followed by the factors of low household income ($\beta = 0.410$) and number of vehicles ($\beta = -0.401$). These findings indicate that gender, household income, and number of vehicles are the dominant factors directly affecting commuting distance. When comparing the magnitudes of the standardized effects of built environment factors at neighborhood level, distance from CBD was found to have the highest standardized effect ($\beta = 0.671$), followed by the factors of residential density ($\beta = 0.189$) and land use mix ($\beta = 0.108$). These findings indicate commuting distance was more sensitive to the effects of distance from CBD, residential density, and land use mix, relative to other factors.

\[
\begin{align*}
\text{Age} & \quad \text{Gender} & \quad \text{Occupation} & \quad \text{Household size} & \quad \text{Household income} & \quad \text{Number of vehicles} & \quad \text{Households/Neighborhood} & \quad \text{Streetmade} & \quad \text{Residential density} & \quad \text{Employment density} & \quad \text{Land use mix size} & \quad \text{Distance from CBD} \\
\end{align*}
\]

![Figure 4 Relative effects of neighborhood and individual level factors in the hazard model](image)

The above findings confirmed the important roles that the residential built environment plays in affecting commuting distance, which are consistent with many studies supporting that land use policies should be part of strategy in reducing VMT and energy consumption (Cervero and Murakami, 2010). These results can help the policy makers and urban planners thoroughly understand the influential factors on commuting distance and consequently make effective countermeasures to reduce commuting distance. At the same time, the urban land use-transportation system is such a complex entity that all the components in the system work collaboratively rather than separately. Policy makers and urban planners should be aware that no single land use policy or transportation technology action can offer a complete solution to reduce auto dependency and then energy consumption and transport emission (Liu and Shen, 2011). Instead, integrating different land use policies and transport demand strategies is necessary (Rubin and Noland, 2010). The mixed policies likely will involve transit-oriented development policies that reduce the travel demand from automobiles and the strategies that provide alternative travel modes (e.g. commuter rail and bus) and improve efficiency of transportation systems.

5 Conclusions

In recent years, a growing body of literature has been conducted to examine the relationships between influential factors and travel behavior. Understanding the effects of built environment measures on travel distance can help planning agencies develop effective countermeasures to reduce energy consumption and
transport emission. Although the hazard-based modeling approach has been regarded as a powerful method for analyzing the duration of different activities, the application of hazard-based duration models is limited in the field of urban planning and transportation. To accommodate the spatial context in which individuals make travel decisions, a multilevel mixture hazard model was proposed in this study. A comparison among the general multilevel model, single level hazard model and multilevel hazard model was conducted. Our study suggests that the application of the multilevel model obtains good model fit. The significant spatial heterogeneity parameter indicates that it is necessary to account for the spatial issues in modeling the commuting distance. This research provides additional insights into the influences of neighborhood and individual level factors on commuting distance.

Using the Washington metropolitan area as the study case, the built environment measures were calculated for each TAZ and the commuting trips of 8327 respondents were collected. The robust maximum likelihood (MLR) estimation method for a partial likelihood function was used to estimate the multilevel hazard model, and the model results confirmed the important roles played by the neighborhood and individual level factors in influencing commuting distance. Significant factors affecting the commuting distance were identified, including socio-demographic attributes and built environment measures. To be specific, at the individual level, eight factors (i.e., gender, race, household size, number of household students, workers, vehicles, and household income) were found to have significant effects on commuting distance. Meanwhile, four factors (residential density, employment density, land use mix, and distance from CBD) at neighborhood level were identified to be significantly associated with commuting distance. Overall, our findings can help policy makers and urban planners develop a better understanding on how the neighborhood and individual level factors can influence the commuting distance, and consequently develop more effective and targeted countermeasures.

One limitation of this research should be noted that the impact of residential self-selection was not captured. People who prefer to drive less commuting distance may select themselves into a compact and mixed land use residential neighborhoods. In this case, the residential built environment does not have an effect on commuting distance because the commuting distance is determined by the residential neighborhood choice. To solve this problem, more personal attitudes and preferences data or other modeling techniques are required to disentangle the influences of built environment and self-selection on commuting distance (Cao, 2015; Ding et al., 2015; Mokhtarian and Cao, 2008; Handly et al., 2005) in future studies. This model can only explain certain part of spatial heterogeneity (spatial autocorrelations among the individuals in the same zone). The correlated heterogeneity (spatial relationship among zones) was not considered in this study. For further study, we would like to extend this research to apply a discrete-continuous model to model car ownership, travel mode choice, and travel distance simultaneously accounting for the endogeneity problem.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (71503018) and the National Basic Research Program of China (2012CB725404).

Appendix

For the full-parametric modeling method, the dependent variable should be assumed to follow a known probability distribution. Accelerated failure-time (AFT) model and proportional hazards (PH) model are the two primary categories that are widely investigated. In the AFT model, the natural logarithm of the survival distance is assumed to be linearly associated with the covariates that can be expressed as:
\log d = \beta^T X + \varepsilon \quad (A1)

where $X$ are the covariates, $\beta$ are the coefficients, and $\varepsilon$ is the error term. In an AFT model, the effect of the covariate is to accelerate distance by a factor of $\exp(-\beta^T X)$. Different AFT models can be obtained by varying the distribution of the error term $\varepsilon$, including exponential, Weibull, extreme value, normal, logistic, lognormal, and gamma models (Liu, 2012). For example, let $\varepsilon$ have an extreme value distribution, the survival function $S(d)$ commuting distance will have a Weibull form, thereby the Weibull model is obtained. Similarly, if $\varepsilon$ is assumed to have a stand normal distribution, the duration of commuting distance $D$ before the trip ending will have a normal density, resulting in a lognormal model.

In terms of the PH model, the hazard function on the relationship between the associated covariates and commuting distance process can be written as

\[ h(d) = h_0(d) g(\beta^T, X) \quad (A2) \]

where $h_0(d)$ denotes a known baseline hazard function for commuting distance, $g(\beta^T, X)$ is a nonnegative function indicating the effect of covariates on the hazard, $X$ are the covariates, and $\beta$ are the coefficients. Generally, there is an inherent assumption that the concomitant covariates have a multiplicative effect on the hazard function. A popular choice is to let $g(\beta^T, X) = \exp(\beta^T X)$. For the PH model, different models can be obtained based on different parametric baseline hazard $h_0(d)$, including exponential, Weibull, and Gompertz models. For example, exponential model will be obtained if the baseline hazard is assumed to be constant. If the hazard function is assumed to vary monotonically over distance, it will lead to the Weibull model.

References


Liu, C., Shen, Q., 2011. An empirical analysis of the influence of urban form on household travel and energy
http://dx.doi.org/10.1016/j.compenvurbsys.2011.05.006


http://dx.doi.org/10.1007/s11431-012-4900-z