Residential Geolocation of Households in a Large-Scale Activity-based Microsimulation Model and Development of a High Definition Spatial Distribution of Vehicle Miles Traveled

Yali Chen, Ph.D.
Data Analyst, Budget and Planning Office,
University of California Santa Barbara
yali@geog.ucsb.edu

Konstadinos G. Goulias, Ph.D.
Professor, GeoTrans and Department of Geography
University of California Santa Barbara
goulias@geog.ucsb.edu

Chandra R. Bhat
Professor, Department of Civil and Environmental Engineering
University of Texas Austin
bhat@mail.utexas.edu

Ram M. Pendyala
Professor, Civil, Environmental, and Sustainable Engineering Program
School of Sustainable Engineering and the Built Environment
Arizona State University, Tempe
Ram.Pendyala@asu.edu

Paper prepared for presentation at the 2014 TRB 93rd Annual Meeting and publication in Transportation Research Record.

GEOTRANS Working Paper 0613-001
Residential Geolocation of Households in a Large-Scale Activity-based Microsimulation Model and Development of a High Definition Spatial Distribution of Vehicle Miles Traveled

Abstract

This paper presents a methodology to distribute the Traffic Analysis Zone (TAZ) level synthesized households and their members to parcels according to the household and parcel attributes. Three Multinomial Logit (MNL) models are estimated to represent the residence location association of households and land parcels, one each for single person, two persons or more without children, and two persons or more with children household types. The estimated models are then used in an algorithm that assigns households to locations in the Los Angeles County. Daily Vehicle Miles Traveled (VMT) of each household is assigned in this way to the parcel the household is assigned to using the algorithm. The method illustrated here shows the feasibility of doing this assignment using millions of parcels and households. It also shows that the results are reasonable and that it is possible to estimate VMT at specific locations and for spatially disaggregate jurisdictions, enabling the assessment of policies at very fine levels of resolution. In addition, our findings and related maps challenge the claim that central city residents travel less miles and suburban residents travel more.
INTRODUCTION

Recent legislation in California aims at creating the framework for a new approach to the design of cities that provides incentives for projects able to decrease household Vehicle Miles of Travel (VMT). Many of these projects, by nature, work at a very fine level of spatial resolution because they need to be coordinated with housing policies (SCAG, 2009). For instance, one such project envisions fine resolution interventions such as infill development jointly with public transportation provision (http://opr.ca.gov/docs/Proposed_Appendix_M.pdf).

Assessing VMT reduction at fine spatial levels of resolution requires the development of procedures that are able to associate (allocate) household-level VMT with the parcel of land on which the household resides. This is feasible when a region has an activity-based model (or a high definition equivalent model) that is also synthetically generating all the households in a region and a detailed database of the residential parcels and their characteristics. Such activity-based models (Kitamura, 1988, Axhausen and Garling, 1992, Bhat and Koppelman, 1999, Vovsha et al., 2004, Henson et al., 2009) are becoming increasingly accepted today, and are being implemented by many small and large MPOs in the United States and elsewhere. As part of these models, which are applied at the disaggregate level of households and individuals, the entire resident population of a region is synthesized in terms of households and individuals (Henson et al., 2009, Goulias et al., 2013).

In this paper, we use the output from a recently developed activity-based micorosimulator for the Southern California Association of Governments labeled as SimAGENT (for Simulator of Activities, Greenhouse gas emissions, Energy, Networks, and Travel; see Goulias et al 2012a), and show how the VMT predicted by SimAGENT at the household level can be assigned to individual parcels in the region. SimAGENT is based on synthetically generating the activity schedules of people in a day, accommodating for intra-household interactions (Bhat et al., 2012). The models embedded within SimAGENT for predicting daily travel patterns and activity time allocations are
influenced by fine resolution accessibility indicators that recognize the important influence of land use on activity-travel behavior. In this way, the analyst is able to examine the shifts in activity-travel patterns not only due to transportation system changes, but also due to land use policies (see Goulias et al., 2012b and Pendyala et al., 2012).

One of the limitations of activity-based models to date, however, is the continued use of traffic analysis zones as the spatial unit of analysis. This is done for the residential location of households, employment and school locations, and activity locations. In essence, the model system, instead of representing each origin and destination of a trip (and the location of each activity) as a point that corresponds to a building, represents locations as a centroid of a zone. In an earlier research study Tang et al., 2013 presented a method that assigns activities to business establishments, offering a solution to the geolocation of jobs, schools, and activities. In the research presented here, we discuss the development of a method to assign simulated households to housing units (and therefore, parcels) for the entire County of Los Angeles. By doing so, we are then able to translate the household and individual activity-travel patterns predicted by SimAGENT to the fine spatial resolution of individual land parcels, which in turn enables the evaluation of VMT reductions at fine spatial levels of resolution.

The method presented here uses household demographics data from a travel survey and recovers the residence characteristics of each household using spatial matching of addresses to land parcels. The resulting sample of households is used to develop models that correlate household characteristics with housing characteristics. Once estimated, the models are then used to predict the housing type for each synthetically created household of SimAGENT in each geographic subdivision in which the household resides. Finally, a matching routine of allocating households to specific housing locations (parcels) is applied to each geographic unit of the large scale microsimulation model. VMT simulated by SimAGENT for each household is then associated with each parcel to develop maps of VMT spatial distribution.
In the next section, we describe the data used in this method. This is followed by the residential assignment models and their estimation followed by the application algorithm and results. The paper ends with a summary and a list of next steps.

**DATA USED**

Two data sources are used to estimate the residential location assignment models. The first is the 2001 post-census SCAG region household travel survey and the second is the SCAG Parcel and Property Assessment Database. The first data set, the 2001 post-census SCAG region household travel survey (HTS), contains randomly selected households with their characteristics and their travel-activities within the SCAG region. The household characteristics include demographic information such as home location address, household size, income, residence type, and tenure of homeownership. The survey also provides demographic information for each household member, including age, gender, education, and ethnicity. The second data set, the SCAG Parcel and Property Assessment Database (PPAD), collected parcel information from each county office. This database consists of parcel shape files and assessment data, including address, land value, square footage, and number of bedrooms/bathrooms of the housing unit located in the parcel. The HTS data are processed to give them a housing unit through address matching with the PPAD using the following steps:

1. Process the addresses in the two databases to reconcile the different formats;
2. Join the two databases using processed addresses and identify the residence parcels for the household sample in the HTS if both of the addresses are correctly recorded and matched with each other;
3. If none of the parcel addresses matches with the location address of a household in the HTS, use other internet-based map and parcel shape files to locate the corresponding parcel for the household.

The above assignment process of households to parcels was undertaken for all households in the HTS database residing in Los Angeles County. This resulted in an
original sample of 6,714 households from the HTS. Of these, only 5,915 households were able to matched to residence parcels due to the address mismatching and other data related issues in the parcel data and the addresses in the HTS. These 5,915 households, along with their characteristics and the characteristics of the parcels and associated block-level demographics in which the parcel is located, are used to develop models that enable us to locate a given household in a certain parcel of land. A multinomial logit (MNL) model formulation is used for this purpose, though we segment the 5,915 households into three separate categories (and actually estimate three separate MNL models) to account for intrinsically different parcel preferences based on the following three household types: single persons, couples (two or more adults with no children), and couples with children (two or more adults with children).

Table 1 provides a summary of important sample characteristics of the 5915 households, segmented by the three household types just identified. The descriptive statistics are all reasonable. Of particular note is that single person households, relative to the other two household types, tend to reside in housing units that have fewer bedrooms, have a smaller square footage, are of a lower land value, and are in highly dense neighborhoods. Single person households also have lower car ownership levels and are much more likely to rent their dwelling unit.

Table 1 Sample Characteristics

<table>
<thead>
<tr>
<th>Variable</th>
<th>Single person</th>
<th>Two persons or more without children</th>
<th>Two persons or more with children</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Average</td>
<td>Standard Deviation</td>
<td>Average</td>
</tr>
<tr>
<td>Household size</td>
<td>1.00</td>
<td>0.000</td>
<td>2.00</td>
</tr>
<tr>
<td>Householder age</td>
<td>48.37</td>
<td>19.438</td>
<td>50.32</td>
</tr>
<tr>
<td>Number of workers</td>
<td>0.64</td>
<td>0.481</td>
<td>1.19</td>
</tr>
<tr>
<td>Number of students</td>
<td>0.16</td>
<td>0.368</td>
<td>0.25</td>
</tr>
<tr>
<td>Number of bedrooms</td>
<td>1.96</td>
<td>1.223</td>
<td>2.53</td>
</tr>
<tr>
<td>Square footage</td>
<td>1181.13</td>
<td>873.556</td>
<td>1470.4</td>
</tr>
<tr>
<td>Land value</td>
<td>115272.0</td>
<td>221881.5</td>
<td>122934.4</td>
</tr>
<tr>
<td>White % in the block</td>
<td>0.57</td>
<td>0.263</td>
<td>0.61</td>
</tr>
<tr>
<td>Hispanic % in the block</td>
<td>0.31</td>
<td>0.278</td>
<td>0.27</td>
</tr>
<tr>
<td>Asian % in the block</td>
<td>0.11</td>
<td>0.131</td>
<td>0.11</td>
</tr>
<tr>
<td>Population density</td>
<td>8.17e-3</td>
<td>7.496e-3</td>
<td>6.08e-3</td>
</tr>
<tr>
<td>Number of Cars in Household (%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
RESIDENTIAL LOCATION ASSOCIATION MODEL ESTIMATION

The estimation of a multinomial logit formulation for parcel preference using the sample described in the previous section requires that, for each household, we generate alternatives in addition to the parcel of land on which the household actually sits. To do so, we randomly selected 50 parcels from the universal choice set of 2,359,345 parcels in Los Angeles County as alternatives (along with the chosen parcel) for each household. The "utility" $U_{in}$ for each parcel alternative $i=1, 2, ..., J$ for an individual household $n$ is given by the functional form shown in equation (1).

$$U_{in}^n = V_{in}^n + \varepsilon_{in}^n = \beta_i X_i^n A_i^n + \alpha_i A_i^n + \varepsilon_{in}^n$$

where

- $n=1, ..., N$ (number of households)
- $i=1, ..., I$ (number of alternative parcels)
- $X_i^n$ – Household attributes (e.g., income, household size)
- $A_i^n$ – Parcel attributes (e.g., square footage, land value)
- $\alpha_i$ – Coefficient
- $\beta_i$ – Coefficient
- $\varepsilon_{in}^n$ – Random error term

The utility formulation above is similar to the one used in earlier studies of location choice (see, for example, Guo and Bhat, 2004, Wadell and Ulfarsson, 2003). The household attributes used in our analysis include household size, total number of vehicles, residence dwelling type, whether or not to own the house, household income, number of workers, number of students, highest education level of household, householder age (householder in the HTS survey is the main household respondent), child presence (a child is defined as 17 years old and younger), and race. The parcel attributes
are square footage, number of units, number of bedrooms/bathrooms, land value, and block level characteristics including opportunity based accessibility indicators for 15 industries (Chen et al., 2011) and census block demographics. The interaction variables of household and parcel attributes are included in the MNL models to reflect heterogeneity across households for their home location choice preference.

The three MNL models for single person household, two-person without children household, and couples with children household are estimated and presented in Table 2. The variables explaining the propensity of each household type for a different housing unit contain a set of variables describing the housing units (single family house, number of bedrooms, square footage, and land value), another set of variables describing the block within which the house is located (percentage of different race/ethnicity groups and population density), accessibility of the block within which the housing units is located (derived from Chen et al., 2011; these are opportunity counts within a buffer of 10 minutes driving on the surrounding network), and a set of variables representing the interactions of household structure with housing attributes.

Table 2 MNL model estimation results

<table>
<thead>
<tr>
<th>Variable</th>
<th>Single person</th>
<th>Two persons or more without children</th>
<th>Two persons or more with children</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coeff.</td>
<td>t-stat.</td>
<td>Coeff.</td>
</tr>
<tr>
<td>single family house</td>
<td>-1.730</td>
<td>-12.046</td>
<td>-1.421</td>
</tr>
<tr>
<td>number of bedrooms</td>
<td>-0.262</td>
<td>-3.099</td>
<td>-0.110</td>
</tr>
<tr>
<td>square footage</td>
<td>-0.419</td>
<td>-3.132</td>
<td>-0.910</td>
</tr>
<tr>
<td>land value</td>
<td>-0.003</td>
<td>-7.6</td>
<td>-0.006</td>
</tr>
<tr>
<td>white % in the block</td>
<td>-1.013</td>
<td>-6.436</td>
<td>-0.861</td>
</tr>
<tr>
<td>Hispanic % in the block</td>
<td>-1.547</td>
<td>-13.097</td>
<td>-1.689</td>
</tr>
<tr>
<td>Asian % in the block</td>
<td>0.250</td>
<td>-1.252</td>
<td>-0.573</td>
</tr>
<tr>
<td>population density</td>
<td>0.033</td>
<td>9.163</td>
<td>0.015</td>
</tr>
<tr>
<td>accessibility by industry (AM peak)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Construction</td>
<td>-0.157</td>
<td>-3.068</td>
<td></td>
</tr>
<tr>
<td>Transportation</td>
<td>0.038</td>
<td>4.451</td>
<td>0.024</td>
</tr>
<tr>
<td>Information</td>
<td>-0.094</td>
<td>-5.100</td>
<td></td>
</tr>
<tr>
<td>Finance</td>
<td>-0.094</td>
<td>-5.506</td>
<td>-0.065</td>
</tr>
<tr>
<td>Professional</td>
<td>0.074</td>
<td>5.742</td>
<td>0.080</td>
</tr>
</tbody>
</table>
The parcel attributes are significant with negative coefficients in the three models, and suggest that single family houses, houses with many bedrooms, big square footage, and costly land value contain a lower number of households in the sample. However, the real insights arise when the effects of the parcel attributes are interpreted as interaction variables with household attributes (see the variables listed after the accessibility measures in Table 2). The positive signs of the coefficients for parcel attributes interacted with household size imply that households with more persons tend to choose single family homes and houses with more bedrooms. Household income has positive coefficients when interacted with parcel attributes, suggesting that households with high income are likely to live in single family homes, houses with more bedrooms and higher land value, and bigger houses with greater square footage.
The coefficients on the race/ethnicity percentages in the block, by themselves, do not provide much insights, because they need to be examined in combination with their interaction with the race/ethnicity of the household (see the last three rows of the table). The net implication of these variables is that there is clear and statistically significant ethnic spatial clustering. White households are more likely to live in the blocks with a higher percentage of white people, independent of household type. Hispanic households show a similar tendency for all household types. Although ethnic clustering effect is not significant for Asian households with a single person, the other two household types are likely to locate themselves in a block with higher Asian percentage.

Accessibility measures for 15 industry types were included in the model estimation, but only a few indicators turn out to be significant. All of the three MNL models have negative signs for the finance industry, which suggest that households tend to stay away from blocks with high finance accessibility. Conversely, households are more likely to be in blocks with high professional and transportation accessibility. The results also indicate that high education accessibility has a negative impact on home location choice of households of two persons without children. None of the accessibility measures turned out to be statistically significant when interacted with household attributes (over and above the differences due to the segmentation by household type).

In addition to the Akaike Information Criterion (AIC) that is a function of the likelihood function but penalizes for the use of many variables, the performance of the estimated models is assessed by the percentage of correctly predicted parcels. The MNL model for single persons can correctly predict the living parcel for almost 20% out of all single person households and the other two models can predict more than 10%. Given the fact that it is not realistic to correctly predict the exact housing units as observed, the predicted housing type is introduced to serve as another measure. Table 3 shows that the three models can correctly predict housing type (single family housing or not single family housing) for more than 50% of households.

Table 3 Model validation results
### MNL Model Application

SCAG's SimAGENT model system generates households for each Traffic Analysis Zone (TAZ) in the SCAG area. A procedure is designed to assign the TAZ level households to individual parcels using the estimated models. Figure 1 describes the flow chart of the assignment procedure. The program written in C# performs the assignment for the 2,243 TAZs in Los Angeles County that we selected for this pilot exploration (because we had complete parcel information for this county). As shown in Figure 1, with the parcel and household data for a TAZ, the program calculates the "utility" values for each household and parcel pair within the TAZ using the estimation results of the three models developed in model estimation section. The assignment is performed in two steps for both parcels with single family housing units and multiple dwelling units. For single family housing units, the program identifies the household with the highest utility value for every parcel and assigns the household to the parcel. When a household is assigned to a parcel, the household will be deleted from the household list and parcel unit will be deleted from the list of housing units. Different from single family housing units, the program locates the households with the highest to (k-1) highest utility value instead of the highest utility for the parcels with k units. It is worth noting that there are more households than the total number of housing units in a few TAZs due to the difference in synthesized household data and real parcel data. The remaining households from the above mentioned two steps are then randomly assigned to the parcels with multiple dwelling units.

After this assignment, the vehicle miles traveled (VMT) for every parcel is calculated by summing up the VMT generated by households located in the parcel. Figure 2 presents the VMT distribution in a TAZ located close to the interchange of I-10 and I-405. Dark green represents the parcel with multiple dwelling units. Light green represents the single housing unit. Red parcels are commercial parcels without any

<table>
<thead>
<tr>
<th>Model Description</th>
<th>Percentage of households which correctly predicted living parcel</th>
<th>Percentage of households with correctly predicted housing type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model for single person</td>
<td>19.5%</td>
<td>56.8%</td>
</tr>
<tr>
<td>Model with two or more persons without children</td>
<td>11.7%</td>
<td>60.1%</td>
</tr>
<tr>
<td>Model with two or more persons with children</td>
<td>10.2%</td>
<td>63.7%</td>
</tr>
</tbody>
</table>
assigned households. As one would expect, parcels with multiple housing units produce more VMT than the single housing parcels. Since no households are assigned to commercial parcels, these parcels produce zero passenger VMT. In addition, a few single housing parcels have zero VMT due to no trips generated in SimAGENT for the households living in the parcels on the simulation day. Figure 3 is the same depiction of VMT per parcel but this time contains a larger portion of the Santa Monica area in Los Angeles County.
Read parcel and household data for one TAZ

Divide the housing units in the TAZ into two groups
1. single family housing unit
2. multiple dwelling

Calculate the utility for each household and housing unit pair

single family housing unit

For \( j = 1 \) to total number of single family housing units

identify the household \( i \) with the highest utility \( U_i \) for housing unit \( j \)

If \( U_i = \) maximum \( U_i \) for household \( i \)

Yes

Assign the household \( i \) to housing unit \( j \) and delete them from the household and housing unit set

Matched housing unit and household pairs

No

Keep the housing unit in the housing unit set

Unmatched housing units and households

Repeat the process till all the housing units are assigned with a household
Multiple dwelling unit

For j = 1 to total number of single family housing units

identify the households with the 1-(k-1) highest utility $U_j$ for housing unit j (k units)

Yes

If $U_{ij}$ = maximum $U_{ij}$ for household i

Assign the household i to housing unit j and delete them from the household and housing unit set

Matched housing unit and household pairs

No

Keep the housing unit in the housing unit set

Unmatched housing units and households

Repeat the process till all the housing units are assigned with k households

Figure 1 Flow chart of the residence location assignment
Figure 2 Example of VMT distribution in a TAZ
One of the modeling and simulation objectives is to produce estimates of VMT per person under different land use policy scenarios. In addition, California developed targets that Metropolitan Planning Organizations should meet to satisfy recent legislative mandates (http://www.scag.ca.gov/factsheets/pdf/2009/SCAG_SB375_Factsheet.pdf). It is important then to also develop maps that show VMT per person and verify if residents of places with higher density produce more or less daily VMT per capita. This computation needs to be undertaken at the smallest possible spatial unit, and then may be aggregated to produces zonal averages. In this way, we have the data needed to test for the existence of the modifiable areal unit problem (or MAUP) that can distort spatial relationships and findings (Openshaw, 1984, Guo and Bhat, 2004). To examine this issue, we developed the map of Figure 4, which depicts the daily VMT generated by the persons who live in each parcel and each TAZ. The parcels show very different VMT per person even when they share almost the exact same spatial characteristics and the TAZs give the impression of homogeneity in behavior within the zone. This is the type
of behavioral heterogeneity that one should expect from a microsimulation model that attempts to mimic the real world and attempt to avoid artifacts of presentation such as the MAUP. However, we need to be cautious about these findings and develop a method to verify these findings further.

Figure 4 Daily VMT per Person Geolocated
SUMMARY AND CONCLUSION

This paper presents a methodology to distribute the TAZ level synthesized households to parcels according to the household, parcel attributes, and the US Census block in which the parcel is found. Three MNL models are estimated to represent the residence location association of households and land parcels. The estimated models are then used in an algorithm that assigns different types of households to locations in the Los Angeles County. Daily VMT of each household is assigned in this way to the parcel each household is allocated (geolocated) using the algorithm. The method illustrated here shows the feasibility of performing this task using millions of parcels and households. It also shows the results are reasonable and we are able to estimate VMT at specific locations and for spatially disaggregate jurisdictions enabling the assessment of policies at very fine levels of resolution.
There are, however, a few limitations and next steps. The MNL models can be refined further for a variety of different households using a richer set of attributes. Also, the data used here are more than a decade old. Using the new California Household Travel Survey database and the rich array of housing characteristics, one may estimate improved models that are able to assign households to parcels with higher fidelity. Spot checks of assigned households to parcels should also be done by developing a sampling strategy that enables validation of model outcomes.

Acknowledgments

Funding for this project was provided by the Southern California Association of Governments, the University of California (UC) Lab Fees program through a grant to UCSB on Next Generation Agent-based Simulation, and the UC Multicampus Research Program Initiative on Sustainable Transportation. This paper does not constitute a policy or regulation of any public agency. Srinath Ravulaparthy and Daimin Tang provided advise at different stages of this application. Hsi-Hwa Hu for SCAG provided the data used here and was the project manager for this research.
REFERENCES


Southern California Association of Governments (2009)

