SimAGENT Population Synthesis

SimAGENT Population Synthesis

Ram M. Pendyala, Chandra R. Bhat, Konstadinos G. Goulias, Rajesh Paleti, Karthik Konduri, Raghu Sidharthan , and Keith P. Christian

January 31, 2013

Report Submitted to

SOUTHERN CALIFORNIA ASSOCIATION of GOVERNMENTS

GeoTrans Laboratory, 1832 Ellison Hall, University of California Santa Barbara, Santa Barbara, 93106-4060

SimAGENT Population Synthesis

Table of Contents

1. INTRODUCTION

Planning agencies are increasingly moving towards the development and deployment of tour-based and activity-based microsimulation models of travel demand as the complexity of transportation planning questions they must address becomes greater (Vovsha and Bradley, 2006). Activity-based microsimulation model systems are capable of simulating the activity-travel patterns of each individual in a region's population, essentially replicating a day in the life of a human. The model systems include a series of submodels or components that are sensitive to a host of socio-economic, land use, accessibility, and cost variables, thus providing the ability to assess the impacts of a wide range of travel demand management strategies and land use policies (Shiftan and Suhrbier, 2002). The Southern California Association of Governments (SCAG) embarked on a multi-year effort to develop a comprehensive continuous-time activity-based microsimulation model system so that impacts of alternative policy and land use scenarios could be accurately assessed in response to the mandates of California Senate Bill 375 (SCAG, 2010).

The Comprehensive Econometric Microsimulator of Daily Activity Patterns (CEMDAP) serves as the core engine of the activity-based model system being implemented in SCAG (Bhat et al, 2004). The overall model system, dubbed SimAGENT (Simulator of Activities, Greenhouse Emissions, Networks, and Travel), includes CEMDAP tied together with a series of additional model components needed to generate inputs for CEMDAP as well as process outputs from CEMDAP (Goulias et al, 2011). The key model components that provide inputs to CEMDAP constitute the focus of this chapter.

Virtually all activity-based travel microsimulation model systems require a complete synthetic population for the model region so that the activity-travel patterns of individual travelers can be simulated through the day (Bowman, 2009). The output of an activity-based model system is a series of travel records for each and every individual in the population. As micro data on the actual population is not available, it is necessary to generate a synthetic population of individuals and households such that the distributions of socio-economic and demographic attributes in the synthesized population match known true population distributions (usually available from a census database). There is an increasingly rich body of literature devoted to synthetic population generation, and although refinements continue to be made and variations in underlying algorithms do exist, the overall process for generating a synthetic population is quite well-established (Beckman et al, 1996; Guo and Bhat, 2004; Arentze et al, 2007; Pritchard and Miller, 2009; Auld and Mohammadian, 2010; Mueller and Axhausen, 2011).

A synthetic population is generated based on a set of control variables whose known (census) distributions drive the population synthesis process. When the synthetic population is drawn from a sample file, all of these control variables as well as a series of other attributes of the sampled records are written to the synthetic population file. This synthetic population then serves as the input to the activity-based model components which simulate daily activity-travel patterns for each individual in the population. While this process may be satisfactory, it does raise a key issue worth addressing. As the population of a region is likely to be much larger than the sample file from which synthetic households are drawn, the synthetic population will inevitably have many records that simply repeat themselves. This problem is particularly exacerbated in large scale activity-based microsimulation model deployments such as that for the Southern California Association of Governments. The base year (2003) population for the model region is more than 17 million people, while that for the future year (2035) is forecast to be more than 25 million people. When synthesizing such huge populations (from a sample file of one million records, for example), one is inevitably faced with rather large scale duplication of records. This results in a synthetic population that lacks the rich variance in population characteristics that would be desirable in the context of an activity-based microsimulation model implementation. Not only is there a lack of rich variance in population characteristics when the socio-economic modeling process is confined to the use of a synthetic population generator, but there is an absence of recognition that many socio-economic attributes are choices that people and households make in response to changing demographics. As a result, the socio-economic modeling process does not model choices related to education, employment, occupation, income, and housing type in response to changing population demographics. This lack of sensitivity or responsiveness in the socio-economic modeling process limits the potential application of the overall activity-based model system to analyze alternative demographic scenarios (e.g., implications of an aging population). In addition, while a few attempts have been made to model socio-economic choices of households and individuals (Goulias and Kitamura, 1992; Purvis, 1994; Sundararajan and Goulias, 2003; Morand et al, 2010), there is limited evidence on how well such model systems work in transportation modeling practice.

This chapter describes a comprehensive socio-economic model system that has been implemented in the context of the activity-based model development effort for the Southern California Association of Governments. The chapter presents evidence on the performance of the model system by comparing outputs of the model against known census distributions. The model system includes two major components. First, there is a synthetic population generator capable of synthesizing a population while simultaneously controlling for known distributions of both household and person level attributes (Ye et al, 2009). Second, there is a Comprehensive Econometric Microsimulator for Socioeconomics, Land-use, and Transportation System (CEMSELTS) module (Eluru et al, 2008) capable of modeling medium- and long-term socio-economic choices of individuals and households.

The remainder of this chapter is organized as follows. The next section provides an overview of the synthetic population generator while the third section provides an overview of the socio-economic microsimulator. The fourth section presents results of the application of the synthetic population generator while the fifth section presents results of the application of CEMSELTS for the Southern California region. Finally, concluding thoughts are offered in the sixth section.

2. THE SYNTHETIC POPULATION GENERATOR

The synthetic population generator that has been implemented within SimAGENT for the Southern California Association of Governments is PopGen (Pendyala et al, 2011). PopGen is capable of synthesizing a population while simultaneously controlling for both household and person level attributes of interest. The process implemented in PopGen is rather similar to earlier approaches, except that there is an additional algorithm that reallocates weights across sample households such that person-level control attributes are more accurately replicated in the synthetic population.

The synthetic population generation process in PopGen begins with the identification of a set of control variables for which marginal distributions are available. The control variables are those that are considered important in the transportation modeling context and for which true marginal distributions can be easily obtained, both in the base year and in the forecast year. In the case of

PopGen, control variables are identified both at the household level and the person level. In addition to synthesizing population in households, PopGen is also capable of synthesizing population in group quarters (both institutional and non-institutional) if group quarter control totals are available.

Once the household and person control variables, and their associated marginal distributions, are identified, an appropriate sample file that includes micro data records needs to be obtained. This micro data file serves two important purposes. First, it provides the seed joint distributions across the control variables of interest at the household and person level. Thus, if one has two household control variables, each with five categories, then the sample file provides a 5x5 joint distribution for the variables of interest. As the number of variables and categories per variable increases, the dimensionality of the joint distribution may become very large, leading to the presence of many sparse (or zero) cells in the sample joint seed distribution. Although PopGen incorporates procedures to account for the zero cell problem, due caution needs to be exercised to avoid situations where seed distributions have an excessively large number of zeros wherein zero-cell adjustment procedures could introduce a systematic bias. Second, the sample file is the set of micro data records from which households (and all persons within each household) will be drawn to form the synthetic population.

The joint seed distributions (household and person control variable joint distributions) are adjusted iteratively until the cell values are such that marginal totals replicate the known marginal distributions. This is accomplished using the iterative proportional fitting (IPF) procedure wherein row and column totals are iteratively matched against known marginal control totals in an iterative fashion. At the end of the iterative process, one has cell values that represent the total number of households (or persons) of a particular type (as defined by the multivariate categorization of a cell). The idea behind the synthetic population generation process is to draw households from a sample file according to the cell values obtained.

However, the problem with drawing households (probabilistically) from the sample file according to the expanded household joint distribution cell values is that the drawing process does not recognize the differing household composition (person types) within households of the same cell. For example, consider a cell defined by two-person, two-worker, middle income households. While the households in this cell are all similar with respect to controlled household attributes, they may differ substantially on person attributes. One household in this cell could have a young newly married couple, while another household could have a mature couple of older adults whose children have grown up and moved away. In other words, households need to be drawn from the sample file in such a way that person attributes of interest are controlled as well.

To facilitate this, PopGen employs an additional iterative process called the iterative proportional updating (IPU) algorithm. In this procedure, weights allocated through the IPF process to households of a certain type are readjusted iteratively so that person controls are more accurately replicated in the synthetic population. Say, the IPF process indicates that there should be 100 two-person, two-worker, middle income households in a certain geography (zone, block group, block, or tract). If the sample file has 10 of these types of households, then each households gets a weight of 10. However, as mentioned earlier, not all of these households should be treated with the same weight because they differ in their composition. If the person-based IPF process suggests that this particular geography has a large number of younger individuals, then households in this cell with younger people should be weighted more heavily than households in this cell with older people. The IPU algorithm considers the IPF-generated person joint distribution cell values in reallocating weights among households of each cell (type) so that person control distributions are replicated more accurately.

After each sample household is assigned an appropriate weight that would best match household and person level control totals, appropriate rounding procedures are applied to the frequencies in the IPF-generated household attribute joint distribution so that whole numbers of households may be drawn probabilistically from the sample file into the synthetic population. The weights assigned to each household in the sample file are used to facilitate the probabilistic drawing process and a synthetic population is thus obtained. As the drawing process is probabilistic, numerous draws are performed and the synthetic population that best matches the expanded cell frequencies of the IPF-generated joint distributions is chosen (based on a χ^2 goodness-of-fit statistic).

3. SIMULATOR OF SOCIO-ECONOMIC CHOICES

The synthetic population that is obtained from PopGen includes a host of demographic and socio-economic attributes for each household. These attributes are those available in the sample file (regardless of whether they were used as control variables in the synthesis process). For example, one may have used household size, number of workers, and household income as household level control variables. In addition to these variables, there are a host of other household attributes that are likely to be available in the sample file, and all of them get carried over into the synthetic population. These may include such variables as vehicle ownership, number of children, housing unit type, family type, race of householder, age of householder, and ownership of home. Similarly, a host of person-level attributes are also carried over into the synthetic population file.

As mentioned earlier, the replication of sample records in the synthetic population results in the loss of a rich variance in population socio-economic characteristics. Moreover, many of the socio-economic choice phenomena are not explicitly modeled as a function of other demographic attributes, thus creating a system where long and medium term choice decisions are not sensitive to household and person demographic characteristics. To overcome these limitations and provide a rich set of socio-economic inputs for activity-based modeling, SimAGENT integrates a comprehensive econometric microsimulator of socio-economics, land-use, and transportation system (CEMSELTS) attributes. All of the variables that can be simulated by CEMSELTS are stripped away from the synthetic population generated by PopGen and replaced with simulated values from CEMSELTS. The resulting richer set of inputs is then fed to CEMDAP, the core activity-based modeling engine within SimAGENT to simulate complete daily activity-travel patterns for the population of the region.

Figure 1 presents the overall framework of CEMSELTS. The base year module of CEMSELTS is comprised of two components. The first component corresponds to a series of individual attributes including educational attainment, student status, school/college location, labor force participation, employment industry, work location, weekly work duration, and work flexibility. The second module corresponds to household level attributes of interest including household income, residential tenure, housing unit type, and household vehicle fleet characteristics. The model system may be considered a hierarchical system of submodels where the outputs of a model higher in the hierarchy serve as inputs to subsequent models later in the hierarchy. Virtually all of the models constitute econometric choice or duration models.

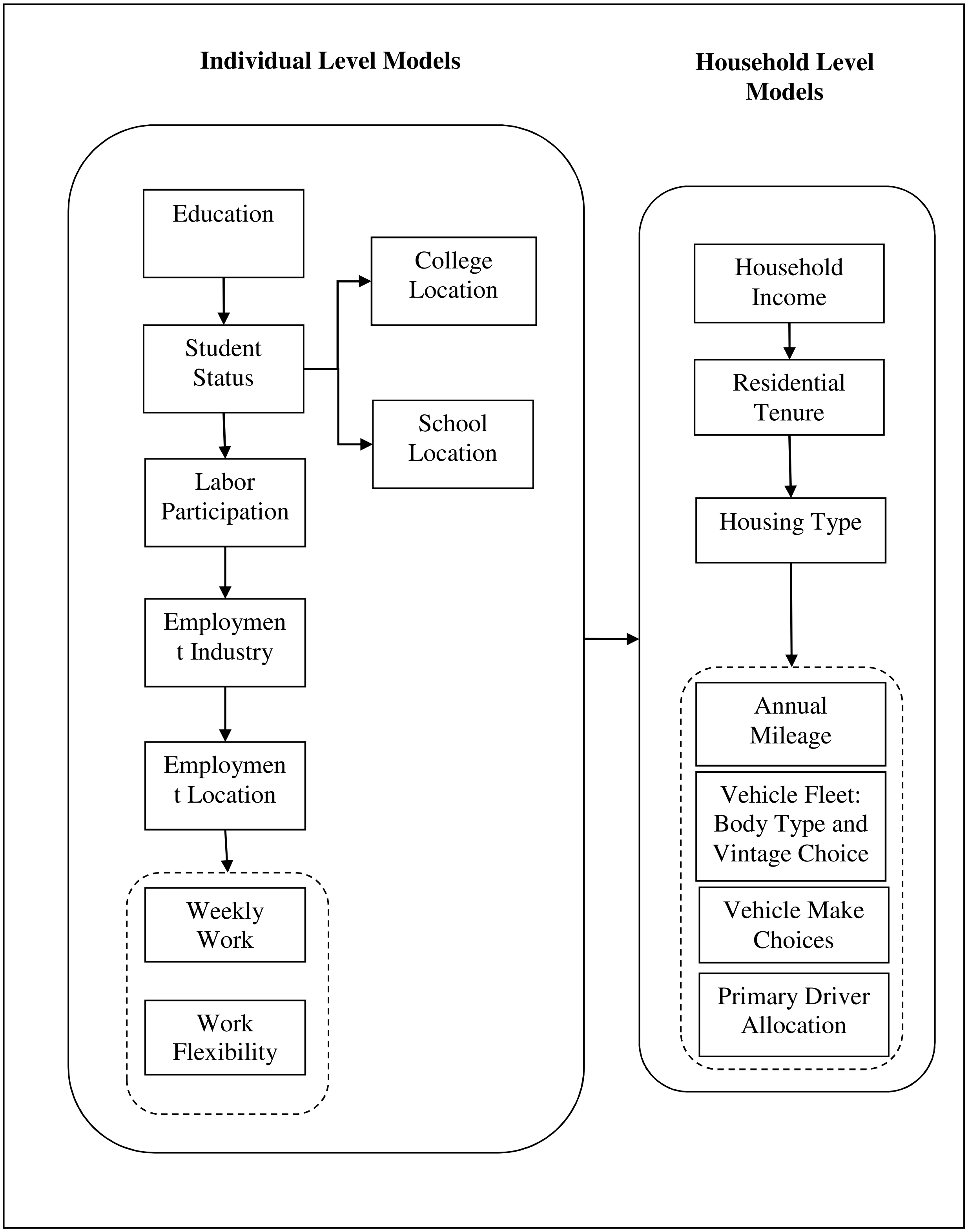

Figure 1. Basic Framework of CEMSELTS

3.1 Individual Level Models

Within the CEMSELTS model, all individuals under five years of age are assumed to not go to school (although they may go to child care facilities, such activities are modeled in CEMDAP). All individuals between 5 and 12 years of age are assumed to pursue education using a rule-based assignment to grades kindergarten through seven, based on age of the child. A rule-based probability model, constructed using look-up tables of school drop-out rates, may be used to determine the education level of individuals between 13 and 18 years of age based on such attributes as age, gender, and race. Another rule-based probability model, similarly constructed using look-up tables of educational achievement, is used within CEMSELTS to determine the education status of each individual 18 years of age or over.

Following the modeling of educational status, the school and college location of all individuals who are students are simulated. At this time, for simplicity, a simple rule-based school location model is used for individuals under the age of 18. All individuals under the age of 18 are assumed to go to school to the closest zone with a school. While it is true that many students attend schools that are not within their neighborhood or assigned school district, it is difficult to model school location choice in the absence of attributes about the various schools in the region. If such data were available, then a robust school location choice model could have been estimated. For those 18 years or age or over, a multinomial logit model of college location choice is estimated and deployed in CEMSELTS. All of the zones with colleges and universities constitute the choice set for the college location model.

A binary logit model is used to determine whether an individual is participating in the labor force. This model is estimated and applied for all individuals aged 16 years and over. The occupation industry is determined using a classic multinomial logit model with the following six alternatives – construction and manufacturing, trade and transportation, professional business, government, retail, and other. The work location of all workers is determined using a multinomial logit model. The universe of zones in the study region forms the choice set for this model. Several zonal characteristics including population, fraction of retail employment, fraction of service employment, level of service variables including travel time and travel cost, and accessibility measures capturing the number of employees (in 15 different industry types) that can be reached within different travel time windows from any given zone are included as explanatory variables in the work location model. In addition, several interaction variables that account for observed heterogeneity among individuals (due to demographic attributes, such as age and gender) are included in the work location model specification.

Finally, two additional work characteristics – weekly work duration and work flexibility – are modeled. While weekly time expenditure for work may be modeled as a continuous duration variable, CEMSELTS models weekly work duration using a multinomial logit model with a view to determine whether an individual works part-time, full-time, or over-time. The three alternatives are defined as working less than 35 hours per week, between 35 and 45 hours per week, and over 45 hours per week. Work flexibility is characterized as an ordinal variable with four levels – none, low, medium, and high degrees of flexibility (as specified by respondents to travel surveys that include such information).

3.2 Household Models

CEMSELTS includes a model of household income that includes a host of employment, occupation industry, and demographic variables as explanatory factors. A grouped ordered response model formulation is used for household income. The five categories in the household income model of CEMSELTS are: less than $10,000 per year, between $10,000 and $35,000 per year, between $35,000 and $50,000 per year, between $50,000 and $75,000 per year, and more than $75,000 per year. Home ownership (whether own or rent housing unit) is determined using a binary logit model that includes a series of socio-economic and demographic attributes as explanatory variables in addition to a few accessibility and built environment variables. Separate multinomial logit models are estimated and applied to the two home ownership groups (owners and renters) to determine housing unit type. The alternatives in the multinomial logit model for households that own their units are single-family detached, single-family attached, and mobile home/trailer. The alternatives in the model for those renting their home are single-family detached, single-family attached, and apartment.

Finally, CEMSELTS includes a series of four models that collectively simulate the vehicle fleet composition for each household in the synthetic population. Unlike most models that only simulate vehicle count, CEMSELTS is capable of simulating vehicle fleet composition with each vehicle characterized by body type, vintage, and make and model. In addition, each vehicle is assigned a primary driver from the household. This allows one to track vehicle usage later in the activity-travel simulation process, a critical step towards more accurately forecasting energy consumption and greenhouse gas emissions in response to alternative policies aimed at encouraging ownership and use of fuel efficient and clean vehicles.

In the vehicle fleet composition and allocation module, the total annual household mileage (including non-motorized mileage) is first determined using a log-linear regression model. The output of this model is used as input to the Multiple Discrete Continuous Extreme Value (MDCEV) model of vehicle fleet composition (Bhat and Sen, 2006). This model uses the total mileage as a travel budget which is allocated across the fleet of vehicles in the household. The MDCEV model formulation explicitly recognizes that vehicle ownership is characterized by multiple discreteness, with households free to choose multiple vehicle alternatives from among those in the market place.

At this time, each alternative in the MDCEV model is defined as a combination of body type and vintage category. Nine body types are used, namely, sub-compact car, compact car, medium car, large car, sports car, medium sports utility vehicle (SUV), large SUV, van, and pick-up truck. Six different vintage categories are used, namely, new or less than one year, two to three years, four to five years, six to nine years, 10 to 12 years, and more than 12 years. The fuel type is not yet included as a dimension in the vehicle type choice model because of the very few observations of alternative fuel vehicles in virtually all vehicle data sets of travel surveys. As additional survey data about ownership of alternative fueled vehicles becomes available, the vehicle fleet composition simulation framework in CEMSELTS can be easily expanded to include consideration of fuel type. In the current version, the total number of alternatives in the MDCEV model is 55 (54 combinations of body type and vintage categories plus one non-motorized mileage alternative).

After the vehicle type is simulated, the make and model of all vehicles in the fleet is determined. This is done using a multinomial logit model. The choice set for the multinomial logit model varies by body type and vintage category. There are a total of 759 make and model alternatives across all of the 54 combinations of body type and vintage categories. The model specifications include numerous variables that describe the attributes of each vehicle make and model. This information is obtained from the Wards Automotive Year Books and Green Vehicle Guide of the US Environmental Protection Agency (Binder, 2010; EPA, 2011). This secondary data is appended to the vehicle records in a travel survey data set to facilitate model estimation. The model is therefore able to include several key vehicle attributes such as dimensions of the vehicle, horse power, engine capacity, type of wheel drive, curb weight, greenhouse gas rating, annual fuel cost, purchase price, and vehicle manufacturer indicator variables.

Finally, a multinomial logit model is used to determine the primary driver of each vehicle owned by the household. The number of alternatives in this model is equal to the number of licensed drivers in the household. The model includes interaction terms that account for observed heterogeneity due to demographic attributes (such as gender, education, employment) that affects the allocation of drivers to vehicles. At this time, the MDCEV model of vehicle fleet composition and the multinomial logit model of primary driver allocation are independent models implemented in a sequential manner. However, in subsequent versions of CEMSELTS, a joint simultaneous equations MDCEV-MNL model of vehicle fleet composition and primary driver allocation that accounts for unobserved heterogeneity in vehicle choices and correlated unobserved factors affecting the endogenous variables will be deployed.

4. RESULTS FROM THE SYNTHETIC POPULATION GENERATION PROCESS

This section presents results from the application of PopGen in the Southern California Association of Governments model region for the year 2003. Although the base year for the activity-based microsimulation model is going to eventually be 2008, the current implementation is based on a 2003 base year. In addition, extensive comparisons against census data (to validate PopGen and CEMSELTS) have been done for 2003; hence, this chapter presents results pertaining to that year.

For the 2003 simulation year, PopGen was implemented as follows. Marginal distributions on control variables were furnished by the Southern California Association of Governments

(SCAG) at the level of the traffic analysis zone (TAZ) for a total of 4109 zones. Of these zones, 4,035 had at least one household which needed to be synthesized. Population synthesis was performed for this set of zones. While marginal distributions are obtained at the zonal level, sample joint distributions are obtained from the Public Use Microdata Sample (PUMS) of the US Census for the year 2000. The PUMS is a five percent sample for the entire State of California; although the subsample corresponding to the Southern California region could have been extracted, the entire state PUMS data was used to have a richer sample from which to draw households and upon which to derive initial joint seed distributions. As the simulation year of 2003 is rather close to the PUMS year of 2000, this sample was considered satisfactory in terms of its representation of California's population in the year 2003. Note that subsequent simulations in which 2008 is treated as the base year is using more recent American Community Survey PUMS data so that there is a reasonable temporal correspondence between the sample file and the simulation year. Regardless of the year of simulation, SCAG is providing all marginal distribution information for control variables of interest at the level of the zone.

In order to facilitate the synthesis process, every zone in the model region is mapped to a PUMA or public use microdata sample area. This is because the location of each household in the PUMS file is specified at the level of the PUMA. In other words, joint seed distributions of the control variables of interest can be derived from the PUMS file only at the PUMA level. As geographical location information is available in the PUMS file only at the PUMA level, but population synthesis must be done at the zonal level, all zones that fall within a PUMA get the same sample seed joint distribution. The correspondence between zones and PUMA geographies is also provided by the Southern California Association of Governments.

The control variables used in the synthesis process and their categories are shown in detail in Table 1. Control variables were chosen based on their potential importance in influencing activity-travel patterns of individuals in the population and the availability of marginal distributions at the zonal level through the SCAG socio-demographic forecasting processes. The synthesis was conducted using a series of household level control variables, yielding a total of 280 household level constraints, and a series of person level control variables yielding a total of 140 person-type constraints. Household income is another important control variable that could have been included in the synthesis process. While household income has been added as a control for the 2008 simulation year, it was not included in the 2003 base year, partly due to concerns about the potentially large number of cells (constraints). Adding income with four categories would have increased the number of household level constraints from 280 to 1140. Although it is reasonable to accommodate such a large number of constraints in the synthesis process, the absence of income as a control variable in the 2003 simulation offers a unique opportunity to see how well the synthesis process is able to replicate the distribution of an uncontrolled variable (whose marginal distribution is known) based on the chosen set of control variables.

The synthesis was performed at the zonal level. The nature of the PopGen algorithm is such that the number of households in the synthetic population exactly matches that corresponding to the number implied by the given marginal distributions. A total of 5,549,771 households were synthesized, which is exactly the same number of households in the region. The total number of persons synthesized is 17,363,222 which is about 1.3 percent less than the actual population total (as implied by the marginal distributions) of 17,595,729. This discrepancy may, at least in part, be due to some minor inconsistencies between the person totals implied by the person control variables and the person totals implied by the household control variables.

Table 1 also presents results of the synthetic population process showing the distributions of various attributes in the synthetic population versus those used to drive the synthesis process. In general, it is found that the synthetic population generation process is able to replicate known distributions of variables in the population quite well. Among household variables, the synthetic population replicates distributions of age of householder and presence of children extremely well. It is found that the synthetic population over-represents family households and under represents non-family households. It appears that the synthetic population generation process falls somewhat short of accurately replicating non-family households. This pattern is seen both in household family type and household type. This pattern of under-synthesizing non-family households is also seen in the household size distribution where single person households are considerably under-represented while larger households are over-represented in the synthetic population. Non-family households are more likely to be single person households than multiple person households, and an under-synthesis of non-family households will naturally yield fewer single person households than desired. Additional attention needs to be paid to the controls necessary to accurately capture the presence of non-family households in the population (particularly because their presence as a proportion of all households in the population is increasing).

Table 1. Results of Population Synthesis

Category	Category Definition	Actual	Synthesized	% Diff
	Household Level Variables			
Household family type				
1	Family	3,930,319	4,040,942	2.81%
2	Non-Family	1,619,452	1,508,829	-6.83%
Householder age category				
1	15 - 64 years old	4,598,761	4,621,472	0.49%
2	65 and over	951,010	928,299	-2.39%
Household size				
1	1 person	1,260,748	1,004,031	-20.36%
2	2 persons	1,519,356	1,536,480	1.13%
3	3 persons	877,779	978,133	11.43%
4	4 persons	869,886	941,830	8.27%
5	5 persons	507,783	542,800	6.90%
6	6 persons	260,011	275,830	6.08%
7	7 or more persons	254,208	270,667	6.47%
Household type				
1	Family: married couple	2,862,133	2,937,310	2.63%
2	Family: male householder, no wife	313,016	326,636	4.35%
3	Family: female householder, no husband	755,170	776,996	2.89%
4	Non-family: householder alone	1,263,432	1,172,531	-7.19%
5	Non-family: householder not alone	356,020	336,298	-5.54%
Presence of own household children				
1	Yes	1,285,454	1,285,333	-0.01%
2	No	4,264,317	4,264,438	0.00%
Household Income (uncontrolled variable)				
1	< \$25,000	1,482,757	1,393,639	-6.01
2	≥ \$25,000 - \$50,000	1,492,578	1,494,229	0.11
3	≥ \$50,000 - \$100,000	1,673,242	1,652,769	-1.22
4	≥ \$100,000	901,194	1,009,134	11.98
	Person Level Variables			
Race				
1	White alone	9,299,723	9,299,051	-0.01%
2	African-American alone	1,305,531	1,262,273	-3.31%
3	American-Indian and Alaska Native alone	167,742	164,926	-1.68%
4	Asian alone	1,840,528	1,813,338	-1.48%
5	Native Hawaiian and other Pacific Islander alone	49,597	49,803	0.42%
6	Some other race alone	4,109,413	3,956,487	-3.72%
7	Two or more races	823,195	817,344	-0.71%
Gender				
1	Male	8,718,816	8,628,836	-1.03%
2	Female	8,876,906	8,734,386	-1.61%
Age				
1	Under 5 years	1,328,570	1,333,832	0.40%
5	35 to 44 years	2,742,378	2,684,693	-2.10%
6	45 to 54 years	2,277,766	2,243,583	-1.50%
7	55 to 64 years	1,422,660	1,408,504	-1.00%
8	65 to 74 years	910,582	924,701	1.55%
9	75 to 84 years	615,458	625,655	1.66%
10	85 and more years	217,032	215,209	-0.84%

It is found that the synthesis process yielded a population whose household income distribution closely replicates the known marginal distribution, even though income was not explicitly controlled. Although the match is quite close, it may be prudent to control for income in the synthesis process given the importance of income in shaping activity-travel behavior. When it is not controlled, the synthetic population has a slight over-representation of high income households and an under-representation of low income households. With respect to person controls, the synthetic population distributions closely mirror the given marginal distributions. All of the percent differences are quite small, and likely stem from the under-synthesis of the overall population total. By enhancing consistency between household controls and person controls, these minor discrepancies can be easily remedied. One of the issues affecting the synthesis is that the population total implied by the given marginal household size distribution is considerably less than the total population count implied by the given person control distributions. It is this discrepancy that is contributing to an under-synthesis of total population. For the 2008 base year synthesis, an adjustment process has been implemented in the synthesis process to modify the household size distribution such that the population counts from the household controls and person controls closely match one another.

5. RESULTS FROM THE APPLICATION OF CEMSELTS

This section presents a detailed discussion of the results obtained from the application of CEMSELTS to model socio-economic characteristics of the synthetic population for the Southern California region. The Southern California Association of Governments (SCAG) provided data regarding school drop-out rates for various ages so that a rule-based probability model of being in school could be constructed for 13 to 18 year old individuals based on age, gender, and race. The agency also provided data regarding educational attainment status for individuals 18 years or age or older. Much of this data is based on census information and is therefore representative of the trends in the population. Accessibility indicators which measure the number of employees that can be reached from any zone within various travel time windows were constructed using detailed micro-level land use data provided by SCAG (Chen et al, 2011). Models of work location, work flexibility, and labor force participation at the person level, and household income at the household level, were estimated using travel survey data for the region.

Finally, the MDCEV model of vehicle fleet composition was estimated using the residential component of the California vehicle survey data collected in 2008. The model to assign a primary driver for each vehicle in the household is estimated using travel survey data. In summary, a suite of models were estimated using local survey and land use data so that the model system was customized to reflect conditions in Southern California.

In order to validate CEMSELTS, the predictions from CEMSELTS were compared against regional socio-economic characteristics as reported in the American Community Survey (ACS) data of 2003 and the decennial census data of 2000. In Table 2, results from the person-level modules of CEMSELTS are compared against the census distributions for these two years. Note that the simulation year for CEMSELTS (and PopGen) is 2003. The model generally predicts characteristics of the population quite well. For children 3 to 17 years old, the model under-predicts the proportion of individuals in the higher grades and over-predicts the proportion of young children going to preschool through third grade. With regard to educational attainment status for adults, the model predicts a larger proportion of individuals as completing high school, whereas the census distributions show higher percentages of individuals having an education attainment less than high school completion. Nevertheless, the model reflects the general trend reasonably well. The labor force participation rate is replicated quite well. The occupation distribution is also reasonably consistent with census distributions except for construction and manufacturing and retail trade where the model under-predicts the proportions, and the other category here the model appears to over-predict the proportion. Overall, percent differences are not substantial.

In Table 3, a comparison of the output of the household level modules of CEMSELTS against census distributions shows that the model, with a few exceptions, is able to replicate distributions quite well. The vehicle ownership distribution is replicated very well, except for a modest over-prediction of the proportion of households falling into the highest vehicle ownership category of four or more vehicles. The distribution of households by number of workers is predicted in a satisfactory manner, with a slight over-prediction of zero-worker households and a slight under-prediction of households with two or more workers. The income distribution is also replicated well, although there is an under-prediction of the percent of households in the highest two income brackets and an over-prediction of the percent of households in the second income bracket. Home ownership and housing unit type distributions are matched very well; however, the housing unit type for renters shows considerable discrepancy. Additional work is warranted in the estimation and calibration of a renter housing unit type model. Whereas CEMSELTS predicts that renters are equally split between single units (attached and detached) and apartments, the census data suggests that nearly three quarters of renters are residing in apartments.

Table 4 offers a detailed look at census journey to work flow distributions in comparison to CEMSELTS predictions of work flows. These work-flows are based on the work locations simulated by CEMSELTS for all workers in the synthetic population. For each origin county in the Southern California model region, the table shows the percent of workers whose work location is within the origin county versus the percent of workers whose work location is outside the origin (home) county. About 85 percent of workers have a work location within the origin (home) county according to the census (American Community Survey data of 2003) and CEMSELTS replicates this number almost perfectly. Even when one examines individual counties, CEMSELTS does an excellent job of replicating journey to work patterns. Note that, consistent with expectations, just over 50 percent of all workers live and work in Los Angeles County – a statistic that is replicated by CEMSELTS.

Table 5 shows the journey to work flow distributions by county pair for the year 2000 (such information is available only in the decennial Census year of 2000) and compares the flow distributions against predictions provided by CEMSELTS. It is once again seen that the model is able to predict county to county work flow patterns remarkably well. The differences between the predicted distributions and the observed census distributions are very small for virtually all cells in the table. Overall, it appears that CEMSELTS is able to simulate socio-economic and work flow characteristics for the synthetic population such that the resulting synthetic population is representative of the true population in the region.

Table 2. CEMSELTS 2003 Individual Level Modules – Comparison with ACS 2003 and Census 2000

Individual Socio-demographics	Values in Percent			Values in Percent		
	ACS 2003	CEMSELTS Predicted	Difference in Percentage	Census 2000	CEMSELTS Predicted	Difference in Percentage
Enrollment of Children (3 to 17 years)						
Preschool - Grade 3	37.07	44.59	7.52	41.17	44.59	3.42
Grade 4 - Grade 8	41.64	42.16	0.52	38.76	42.16	3.40
Grade 9 - Grade 11	21.29	13.25	-8.04	20.07	13.25	-6.82
Educational Attainment (Adults)						
Less than Grade 9	11.58	2.23	-9.35	13.14	2.23	-10.91
Grade 9 - Grade 12 (no diploma)	12.05	8.28	-3.78	14.71	8.28	-6.44
Completed High School	45.70	58.48	12.78	44.00	58.48	14.48
Associate or Bachelors	22.55	22.95	0.41	20.77	22.95	2.18
Graduate Degree (Masters or Ph.D)	8.12	8.06	-0.06	7.37	8.06	0.69
Labor Participation						
Employed	59.47	59.07	-0.40	56.81	59.07	2.26
Unemployed	40.53	40.93	0.40	43.19	40.93	-2.26
Employment Industry						
Construction and Manufacturing	19.92	14.46	-5.46	20.67	14.46	-6.21
Trade and Transportation	4.94	7.32	2.38	4.86	7.32	2.46
Personal, Professional and Financial	50.63	49.42	-1.21	49.34	49.42	0.08
Public and Military	3.94	5.07	1.13	4.04	5.07	1.03
Retail Trade	15.29	10.77	-4.51	15.60	10.77	-4.83
Other	5.28	12.96	7.68	5.49	12.96	7.47

Table 3. CEMSELTS 2003 Household Level Modules – Comparison with ACS 2003 Data and Census 2000

Household Socio-demographics	Values in Percent			Values in Percent		
	ACS 2003	CEMSELTS Predicted	Difference in Percentage	Census 2000	CEMSELTS Predicted	Difference in Percentage
Number of Vehicles						
Households with no vehicles	8.29	7.27	-1.02	10.07	7.27	-2.79
Households with 1 vehicle	33.34	31.32	-2.02	34.85	31.32	-3.55
Households with 2 vehicles	37.48	34.71	-2.77	37.16	34.72	-2.44
Households with 3 vehicles	14.10	15.17	1.07	12.59	15.17	2.59
Households with 4 or more vehicles	6.79	11.52	4.74	5.33	11.52	6.19
Number of Workers						
Households with no workers	12.21	16.84	4.63	11.31	16.84	5.53
Households with 1 worker	34.23	36.80	2.58	32.98	36.80	3.82
Households with 2 or more worker	53.57	46.36	-7.21	55.71	46.36	-9.35
Household Income						
$0- $9999	8.08	8.09	0.01	8.98	8.09	-0.89
$10,000-$34,999	28.85	40.45	11.6	29.56	40.45	10.89
$35,000-$49,999	15.05	14.47	-0.58	15.24	14.48	-0.76
$50,000-$74,999	18.53	13.58	-4.95	18.89	13.58	-5.31
$75,000 and more	29.49	23.4	-6.09	27.32	23.40	-3.93
Household Tenure						
Owner	55.74	61.05	5.30	54.78	61.03	6.25
Renter	44.26	38.95	-5.30	45.22	38.97	-6.25
Household Type for Owners						
Single Unit (Attached/Detached)	88.15	93.42	5.27	54.78	61.05	6.27
Other	11.85	6.58	-5.27	45.22	38.95	-6.27
Household Type for Renters						
Single Unit (Attached/Detached)	27.87	50.49	22.62	88.32	93.42	5.10
Apartment	72.13	49.51	-22.62	11.68	6.58	-5.10

Table 4. CEMSELTS Work Flow Distribution (in Percentage) by Destination – Comparison with the ACS 2003 Data

Origin county	Within Origin County			Outside Origin County			Total		
	ACS2003 (%)	CEMSELTS 2003 (%)	Difference	ACS2003 (%)	CEMSELTS 2003 (%)	Difference	ACS2003 (%)	CEMSELTS 2003 (%)	Difference
Los Angeles	52.79	52.63	-0.16	3.86	5.29	1.43	56.65	57.92	1.26
Orange	15.61	14.28	-1.32	3.11	3.45	0.35	18.71	17.74	-0.98
Riverside	6.57	7.65	1.09	3.19	1.85	-1.35	9.76	9.50	-0.26
San Bernardino	6.88	7.58	0.70	3.18	2.60	-0.58	10.06	10.18	0.12
Ventura	3.73	3.67	-0.06	1.09	1.00	-0.09	4.82	4.67	-0.15
Total	85.57	85.81	0.24	14.43	14.19	-0.24	100	100	0.00

Table 5. CEMSELTS Work Flow Distribution (in Percent) by Destination County – Comparison with the Census 2000 Data

Origin County	Destination County													
	Imperial		Los Angeles		Orange		Riverside		San Bernardino		Ventura		Total	
	Census 2000 (%)	CEMSELTS 2003 (%)	Census 2000 (%)	CEMSELTS 2003 (%)	Census 2000 (%)	CEMSELTS 2003 (%)	Census 2000 (%)	CEMSELTS 2003 (%)	Census 2000 (%)	CEMSELTS 2003 (%)	Census 2000 (%)	CEMSELTS 2003 (%)	Census 2000 (%)	CEMSELTS 2003 (%)
Imperial	0.60	0.76	0.00	0.00	0.00	0.00	0.01	0.02	0.00	0.00	0.00	0.00	0.61	0.78
Los Angeles	0.01	0.00	53.32	52.21	2.39	3.23	0.14	0.31	0.61	1.19	0.48	0.53	56.94	57.46
Orange	0.00	0.00	2.76	2.80	16.26	14.17	0.17	0.35	0.14	0.28	0.01	0.00	19.35	17.60
Riverside	0.01	0.00	0.55	0.23	0.77	0.21	6.22	7.59	0.90	1.39	0.00	0.00	8.45	9.43
San Bernardino	0.00	0.00	1.66	1.03	0.43	0.22	0.78	1.33	6.81	7.52	0.01	0.00	9.69	10.10
Ventura	0.00	0.00	1.02	0.99	0.01	0.00	0.00	0.00	0.00	0.00	3.93	3.64	4.97	4.63
Total	0.62	0.76	59.31	57.26	19.86	17.83	7.32	9.59	8.47	10.38	4.43	4.18	100.0	100.0

6. CONCLUSIONS

The accuracy of travel forecasts is highly dependent on the accuracy of the inputs that drive the forecast. The old adage of "garbage in, garbage out" remains as true today as it has always been in the past. Although model systems are becoming behaviorally more realistic, statistically more rigorous, and econometrically more theoretical and robust, the fact remains that the quality and accuracy of socio-economic input data is of paramount importance in any traditional or emerging transportation modeling system.

In the context of activity-based travel model systems which are capable of microsimulating daily activity-travel patterns of individual travelers, it is necessary to generate a synthetic population with a rich set of explanatory variables (socio-economic and demographic characteristics) that can be used to drive the activity-travel simulation process. This chapter focuses on the generation of such a synthetic population with a rich set of attributes. In particular, this chapter describes the socio-economic model system that has been implemented for the Southern California Association of Governments in conjunction with its activity-based travel demand model implementation effort. The socio-economic model system, which is responsible for generating a representative synthetic population with a rich set of demographic variables, is comprised of primarily two components. The first component is a synthetic population generator capable of simultaneously controlling for known household-level and person-level control distributions. The second component is a comprehensive econometric microsimulator of socio-economics, land-use, and transportation system (CEMSELTS) that is comprised of a series of submodels capable of simulating various medium- and long-term choices of individuals. These include such dimensions as school status, educational attainment, labor force participation, occupation industry, household housing unit type, household income, and household vehicle fleet composition.

The process employed begins with the generation of a synthetic population based on known distributions of control variables. The synthetic population is comprised of households probabilistically drawn from a sample file such that the known marginal control distributions are replicated in the synthetic population. However, as the size of the population is far greater than the size of the sample file, many records get replicated in the synthetic population resulting in a loss of rich variance in socio-economic and demographic attributes that is desirable in a representative population. Many of the medium- and long-term choice attributes are deleted from the synthetic population obtained from the population synthesizer, and are instead simulated using the series of choice models embedded in CEMSELTS. This results in a representative synthetic population with a set of explanatory attributes that vary across the population. The entire model system has been calibrated for the Southern California region and applications of the model system to the 2003 base year simulation show that the process is able to replicate known distributions of attributes in the population very well. Except for the occasional deviation (e.g., housing unit type distribution for renters), the models produce a synthetic population with distributions on socio-economic attributes and journey-to-work flows that closely resemble those in census data.

The contributions of this study are noteworthy on several counts. First, it demonstrates that an enhanced socio-economic modeling system that includes both a population synthesizer and a microsimulator of demographic attributes can effectively produce a representative population for a model region. While the application of a population synthesizer by itself may yield desirable results, the application of a comprehensive econometric microsimulator of socio-economic characteristics in conjunction with a population synthesizer will help provide the rich variance in input variables desired for travel forecasting. This chapter offers real-world empirical evidence that known census distributions can indeed be replicated by a socio-economic modeling system such as that deployed for the Southern California Association of Governments. Second, the chapter demonstrates that microsimulation model systems can be applied in large scale settings such as the Southern California region that encompasses a population of nearly 18 million people. Although there were initial concerns about the ability of a microsimulation model system to replicate patterns of population distributions in such a large and diverse region, it has been shown that a synthetic population generator combined with a socio-economic microsimulator can be successfully deployed in large scale simulation contexts. Finally, the model system includes a novel multiple discrete continuous extreme value (MDCEV) model combined with a multinomial logit model to simulate vehicle fleet composition by type of vehicle and the allocation of vehicles to drivers in the household. This component of the simulator will undoubtedly be useful in addressing emerging planning issues related to energy sustainability and greenhouse gas emissions.

Additional work is ongoing to migrate the model system to a new 11,000+ zone system and examine the computational feasibility of implementing a socio-economic microsimulation model system for such a large number of spatial units. In addition, some of the components of CEMSELTS that are currently implemented sequentially are being combined into joint model systems to simultaneously simulate multiple attributes while accounting for unobserved heterogeneity and correlated unobserved factors across dimensions of interest.

Note: This report is heavily based on the 2012 Annual Transportation Research Board meeting paper with title: THE APPLICATION OF A SOCIO-ECONOMIC MODEL SYSTEM FOR ACTIVITY-BASED MODELING: EXPERIENCE FROM SOUTHERN CALIFORNIA by the same authors of this report.

7. REFERENCES

Arentze T., H.J.P. Timmermans, and F. Hofman (2007) Creating Synthetic Household Populations: Problem and Approach., *Transportation Research Record: Journal of the Transportation Research Board*, 2014, pp. 85-91.

Auld, J. and A. Mohammadian (2010) Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels. *Transportation Research Record: Journal of the Transportation Research Board*, 2175, pp. 138-147.

Bhat, C.R. and S. Sen (2006) Household Vehicle Type Holdings and Usage: An Application of the Multiple Discrete-Continuous Extreme Value (MDCEV) Model. *Transportation Research Part B*, 40(1), pp. 35-53.

Bhat, C.R., J.Y. Guo, S. Srinivasan, and A. Sivakumar (2004) Comprehensive Econometric Microsimulator for Daily Activity-Travel Patterns. *Transportation Research Record: Journal of the Transportation Research Board*, 1894, pp. 57-66.

Binder, A.K. (2010) *Ward's Automotive Yearbook*. Wards Communications, 72nd Edition.

Bowman, J.L. (2009) Population Synthesizers. *Traffic Engineering and Control*, 49(9), p 342.

Chen, Y., S. Ravulaparthy, K. Deutsch, P. Dalal, S.Y. Yoon, T. Lei, K.G. Goulias, R.M. Pendyala, C.R. Bhat, and H-H. Hu (2011) Development of Opportunity-Based Accessibility Indicators. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2255, 2011, pp. 58-68

Eluru, N., A.R. Pinjari, J.Y. Guo, I.N. Sener, S. Srinivasan, R.B. Copperman, and C.R. Bhat (2008) Population Updating System Structures and Models Embedded in the

Comprehensive Microsimulator for Urban Systems. *Transportation Research Record: Journal of the Transportation Research Board*, 2076, pp. 171-182.

EPA (2011) Green Vehicle Guide. Website at http://iaspub.epa.gov/greenvehicles/Index.do, accessed July 28, 2011.

Goulias, K.G. and R. Kitamura (1992) Travel Demand Forecasting with Microsimulation. *Transportation Research Record*, 1357, pp. 8-17.

Goulias, K.G., C.R. Bhat, R.M. Pendyala, Y. Chen, R. Paleti, K.C. Konduri, T. Lei, D. Tang, S.Y. Yoon, G. Huang, and H-H. Hu (2011) Simulator of Activities, Greenhouse Emissions, Networks, and Travel (SimAGENT) in Southern California. Paper submitted to the 91st Annual Meeting of the Transportation Research Board. Working Paper, University of California at Santa Barbara, CA.

Guo, J. Y., and C.R. Bhat (2007) Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 2014, pp. 92-101.

Morand, E., Toulemon, L., Pennec S., Baggio R., and Billari F. (2010) Demographic Modelling: The State of the Art, SustainCity Working Paper, 2.1a, Ined, Paris. Available at http://www.sustaincity.org/publications/WP_2.1a_-_Demographic_models.pdf, accessed July 28, 2011.

Mueller, K. and K.W. Axhausen (2011) Population Synthesis for Microsimulation: State of the Art. DVD Compendium of Papers of the 90th Annual Meeting of the Transportation Research Board, TRB, Washington, D.C.

Pendyala, R.M., K.P. Christian, and K.C. Konduri (2011) *PopGen 1.1 User's Guide*. Lulu Publishers, Raleigh, North Carolina.

Pritchard, D.R. and E.J. Miller (2009) Advances in Agent Population Synthesis and Application in an Integrated Land Use and Transportation Model. DVD Compendium of Papers of the 88th Annual Meeting of the Transportation Research Board, TRB, Washington, D.C.

Purvis, C.L. (1994) Using 1990 Census Public Use Microdata Sample to Estimate Demographic and Automobile Ownership Models. *Transportation Research Record*, 1443, pp. 21-29.

SCAG (2010) SB 375/SCS Technical Methodology and Related Processes for Estimating GHG Emissions. Southern California Association of Governments, Los Angeles, CA. Available at: http://www.scag.ca.gov/sb375/pdfs/CEHD-TechMethodolgy032510_strikethrough.pdf, accessed July 28, 2011.

Shiftan, Y. and J. Suhrbier (2002) The Analysis of Travel and Emission Impacts of Travel Demand Management Strategies Using Activity-Based Models. *Transportation*, 29(2), pp. 145-168.

Srinivasan, S., L. Ma, and K. Yathindra (2008) Procedure for Forecasting Household Characteristics for Input to Travel Demand Models. Final Report TRC-FDOT-64011-2008. Florida Department of Transportation, Research Center, Tallahassee, Florida.

Sundararajan, A. and K.G. Goulias (2003) Demographic Microsimulation with DEMOS2000: Design, Validation, and Forecasting. In K.G. Goulias (ed) *Transportation Systems Planning: Methods and Applications*, CRC Press, Boca Raton, FL, pp. 14-1 – 14-23.

Vovsha, P. and M. Bradley (2006) Advanced Activity-Based Models in Context of Planning Decisions. *Transportation Research Record: Journal of the Transportation Research Board*, 1981, pp. 34-41.

Ye, X., K.C. Konduri, R.M. Pendyala, B. Sana, and P. Waddell (2009) A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations. DVD Compendium of Papers of the 88th Annual Meeting of the Transportation Research Board, TRB, Washington, D.C.