University of California Transportation Center
UCTC-FR-2016-05

Spatial Transferability Using Synthetic Population Generation Methods

Elizabeth McBride, Adam W. Davis, Jay Hyun Lee, Konstadinos G. Goulais
University of California, Santa Barbra
April 2016

DISCLAIMER

Spatial Transferability Using Synthetic Population Generation Methods

FINAL REPORT

Elizabeth McBride, Adam W. Davis, Jae Hyun Lee, and Konstadinos G. Goulias

University of California Santa Barbara
Department of Geography
Geotrans Laboratory

Contract Number: 65A0528.
Principal Investigator: Konstadinos G. Goulias;
Lead Researcher: Elizabeth McBride
Support Researchers: Adam W. Davis & Jae Hyun Lee

April 19, 2016
Santa Barbara, CA

Table of Contents

1. INTRODUCTION

The emergence of individual-based travel behavior models, including discrete choice models, and activity-based microsimulation models have ushered in a new era in travel demand forecasting. These models operate at the level of the individual traveler and the household within which this individual lives. These models are regression-like equations with dependent variables – the behavior we are trying to predict (e.g., number of trips per day) – and explanatory variables – person and household characteristics such as age, employment, education, household size, and so forth. Therefore, we need household and person attribute information to inform these models and then use them for the entire regional population to predict changes in behavior. However, such detailed information is virtually unavailable at the disaggregate level for an entire region. Public Use Microdata Sample (PUMS) and travel surveys provide this detail, but they either do not include fine geographic levels or are not representing the study region. PUMS is particularly useful because it provides 1% and 5% samples from the US Census, and in this way offers dependable joint distributions of multiple variables at the level of the most elementary unit of analysis. In this way, we can replicate observed multidimensional relationships among variables and generate a synthetic (virtual) population with comprehensive data on attributes of interest.

Synthetic populations can be formed from a small random sample from which we extract key information about the relationships among a set of household and person variables. These relationships are the multidimensional (from multiple variables) distributions we want to replicate in the entire population (e.g., the cross classification between household size and number of employed persons). The sample that is used to create this multidimensional distribution is called the "seed" that starts a set of iterations. These iterations reconcile seed univariate (single variable) distributions with aggregate distributions of household and person attributes available through the US Census at small geographic units such as a block or block group. These univariate distributions are called the "marginal" distributions. In the US, Census Summary Files provide the marginal distributions of population characteristics. They can either be from the Decennial Census or the American Community Survey (ACS) that replaced the older US Census long form. The joint distributions among a set of control variables are first estimated using the seed, then their values adjusted using the Iterative Proportional Fitting (IPF) procedure

first presented by Deming and Stephan (1941). Weights created using this method are then used to generate as many households and persons we need to recreate the population of a region.

In this project we create a synthetic population of the entire State of California. In the process we also replicate travel behavior characteristics from the California Household Travel Survey allowing us to estimate the number of trips generated and the miles travelled by the synthetically created people. Key to this project is the addition of land use characteristics in the population synthesis that enable transfer of information from one region to another and to more precisely predict travel behavior of the residents. The research questions we target are:

1. Can we develop a small set of land use groups that capture behavioral heterogeneity?
2. Does the addition of the land use categories modify substantially the synthetic population generated?
3. What are the most important differences between synthetic population with and without land use?

2. LITERATURE REVIEW

In this section, we briefly review the methods that have been applied to population synthesis by Beckman et al. (1996) for the use in TRANSIMS, Guo and Bhat (2007) for Texas, Auld et al. (2007) for Illinois, and Ye et al. (2009) for Florida, Arizona, and California.

Most population synthesizers currently in use are based on the method developed by Beckman et al. (1996) for use in the TRANSIMS model. This procedure matches exact large-area multidimensional distributions of selected variables from the PUMS files to small-area marginal distributions from Census Summary files to estimate the multidimensional distributions for the small areas. The population is synthesized in two stages. First, a multidimensional distribution matrix describing the joint aggregate distribution of demographic and socio-economic variables at household and/or individual levels is constructed. This stage makes use of the Iterative

Proportional Fitting (IPF) procedure. In this procedure, the correlation structure of the large area and within it the smaller areas is assumed to be similar. In the IPF procedure, an initial seed distribution is used and fit to known marginal totals. The difference between the current total and the marginal total for each category of the variable of interest is calculated and the cells of that category are updated accordingly. This process continues for each variable until the current totals and the known marginal totals match to some level of tolerance, producing a distribution which matches the control marginal totals. In the second step, synthetic population is constructed by selecting the entire population from the PUMS in proportion to the estimated probabilities given in the multidimensional matrix obtained by the IPF technique. The number of households of each demographic type to be generated is determined from each aggregate area (or large area). For a combination of demographic characteristics, a set of probabilities is assigned to each household in the PUMS, where PUMS samples close to the combination of desired demographic characteristics are assigned higher probabilities. The households are then selected randomly according to their selection probabilities. These probabilities are computed by a weight based algorithm (Beckman et al. , 1996).

Guo and Bhat (2007) identify two issues associated with the first generation of population synthesis using the Beckman et al. (1996) algorithm. The first issue is incorrect zero cell values: this is an issue inherent to the process of integrating aggregate data with sample data, and the problem occurs when the demographic distribution derived from the sample data is not consistent with the distribution expected in the population. A second issue arises from the fact that the approach can control for either household-level or person-level variables, but not both. If these issues are left unaddressed, they may significantly diminish the representativeness of the synthesized population. Guo and Bhat (2007) propose a new population synthesizer that addresses these issues using an object-oriented programming paradigm. The issue of incorrect zero cell values is solved by providing the users the capability to specify their choice of control variables and class definitions at run time. Furthermore, the synthesizer is built with an error reporting mechanism that tracks any non-convergence problem during the IPF procedure and informs the user of the location of any incorrect zero cell values. Guo and Bhat (2007) also propose a new algorithm using an IPF-based recursive procedure, which constructs household-level and person-level multi-way distributions for the control variables. This is achieved by the

two multi-way tables for households and persons that are used to keep track of the number households and individuals belonging to each demographic group that has been selected into the target area during the iterative process. At the start of the process, the cell values in the two tables are initialized to zero to reflect the fact that no households and individuals have been created in the target area. These cells are iteratively updated as households and individuals are selected into the target area. Given the target distributions and current distributions of households, each household from PUMS is assigned a weight-based probability of selection. Based on the probabilities computed, a household is randomly drawn from the pool of sample households to be considered and added to the population for the target area. A similar idea underlines the processes developed by Pritchard and Miller, 2012, and the PopGen method we review below.

Building on the IPF procedure for population synthesis, Auld et al. (2007) propose a new population synthesizer which consists of two primary stages: creation of a multidimensional distribution table for each analysis area, and selection of households to be created for each analysis area. Auld et al. (2007) adopt the same method for creating a multidimensional distribution table as in other population synthesizers (Beckman et al. 1996, Guo and Bhat, 2007). The complete distribution for all households is fit to the marginal totals through the use of IPF procedure. This creates the regional-level multi-way table that is used to seed all the zone-level distribution tables. For each zone, the seed matrix cell values are adjusted so that the total matches the desired number of households to generate. The zone-level multi-way distribution is adjusted to match the zone marginal distributions by again running the IPF procedure. The selection probability of households from the multidimensional table is performed in a similar manner as that proposed by Beckman et al. (1996), which is a weight of household divided by the sum of the total weighted households for the category variable. Auld et al. (2007) argue that there exists large variation between control marginal totals and those generated by the process so the totals are matched exactly as desired. For this reason, Auld et al. (2007) add further constraints, such that the total number of households that have been generated for each category within each control variable represented by the demographic type. If any of the totals exceed the marginal values from the zone-level marginal by more than a given tolerance, the household is rejected. This procedure works well at keeping the generated marginal totals fairly close to the

actual totals. However, Auld et al. (2007) identify that this method might bias the final distribution. In the population synthesis procedures, aggregating control variables within range-type control variables is primarily done to allow for the use of more control variables and to reduce the occurrence of false zero-cells. With large numbers of control variables, the size of the distribution matrix can become very large and make the IPF procedure intractable. Therefore, Auld et al. (2007) introduced the category reduction option, which occurs prior to the IPF stage. The marginal values for range variables are compared to minimum allowable totals. The minimum allowable category total is defined as the total number of households in the region multiplied by a user specified percentage. The percentage forces all categories with less than the allowable number of households to be combined with neighboring categories. The category is then removed from the multidimensional distribution table. The category aggregation threshold percentage acts as a useful limiter of the total number of categories.

Ye et al. (2009) propose a similar framework by generating synthetic populations with a practical heuristic approach while simultaneously controlling for household and person level attributes of interest. The proposed algorithm uses lessons learned from the three examples above, and it is also computationally efficient in addressing a practical requirement for agencies. The proposed algorithm by Ye et al. (2009) is termed as Iterative Proportional Updating (IPU). It starts by assuming equal weights for all households in the sample. The algorithm then proceeds by adjusting weights for each household/person constraint in an iterative fashion until the constraints are matched as closely as possible for both household and person attributes. Next, the weights are updated to satisfy person constraints. The completion of all adjustment weights for one full set of constraints is defined as one iteration. The absolute value of the relative difference between weighted and the corresponding constraint may be used as goodness-of fit-measure. The IPU algorithm provides a flexible mechanism for generating synthetic population, where both household- and person-level attribute distributions can be matched very closely. The IPU algorithm works with joint distributions of households and persons derived using the IPF procedure, then iteratively adjusts and reallocates weights across households to closely match the household and person level attributes. As mentioned in earlier works (Beckman et al. 1996; Guo and Bhat 2007; Auld et al. 2007), the problem of zero-cells is also addressed in the population synthesis by Ye et al. (2009) borrowing the prior information for the zero-cells from PUMS data

for the entire region. Moreover, due to the proposition of the IPU algorithm, Ye et al. (2009) indicate that zero-marginal problem is encountered in this context. For example, it is possible to have absolutely no low-income households residing in a particular blockgroup. If so, all of the cells in the joint distribution corresponding to low income category will be eliminated and they solve this problem by adding a small positive value to the zero-marginal categories. The IPF procedure will then distribute and allocate this small value to all of the relevant cells in the joint distribution. After the weights are assigned using the IPU algorithm, households are drawn at random from PUMS (or a survey database) to generate the synthetic population. The approach Ye et al. (2009) adopt is similar to that of Beckman et al. (1996), except that the probability with which the household is drawn is dependent on its assigned weight from the IPU algorithm. This algorithm – implemented in the software PopGen – was refined and used in a large geographical area with 18 million residents (The Southern California Association of Governments, SCAG, region). The application took a reasonably low number of hours to run with multiple dimensions at the household and person levels and performed very well in terms of its ability to replicate extremely different marginal distributions at the household and person levels (Pendyala et al., 2012a, 2012b).

Synthetic populations, in addition to providing the explanatory variables for individual and household behavioral equations, are also used to provide the baseline population for demographic microsimulators, and the population for urban economy simulators (see the review by Ravulaparthy and Goulias, 2011). There are also many extensions of the methods described here, including a two-stage IPF to add spatial information from different sources by Zhu and Ferreira (2014); a Markov Chain Monte Carlo approach by Farooq et al. (2013) to ensure uniqueness of the identified distribution, avoidance of loss of heterogeneity, and poor scalability of IPF-based methods; and extending PopGen to multiple geographical scales (Konduri et al., 2016). In this application of population synthesis, we use the PopGen software and algorithms with the addition of land use characteristics in the area of household residence. The seed data come from the California Household Travel Survey, and the land use information from NETS updated up to 2013. To match the CHTS with land use data, we use the 2012 land use characteristics of NETS.

3. POPGEN BASE METHOD

The program we are using is called PopGen. It is an open-source program developed by the SimTRAVEL Research Initiative at Arizona State University ("PopGen: Population Generator"). The program can currently synthesize populations at the following geographic resolutions: county, census tract, census block group, and traffic analysis zone (TAZ). In this project, we aim to synthesize the entire state of California at the block group level. PopGen uses an Iterative Proportional Fitting (IPF) and Iterative Proportional Updating (IPU) based method to perform population synthesis, as described in Section 2. The program can handle simultaneous household-, person-, and group quarters-level synthesis, although in this project we only synthesize households and individuals due to data availability.

To synthesize household- and person-level populations, PopGen requires five files: two input files for each level of synthesis (household and person) – called the marginal distributions and the seed – and a geographic correspondence file. This means we will input household marginal distributions, person marginal distributions, a household seed, a person seed, and a geographic correspondence file. Below, we will describe the purpose and construction of these files.

3.1 Marginal Distributions

The marginal distributions are the estimated number of households and/or people in a block group who fall under specific trait categories. In this project, the marginal distributions come from the American Community Survey (ACS) 2013 5-year summary. The ACS is the newer version of what used to be called the long-form census. A small portion of the population is asked more detailed questions, and surveying goes on year-round. In this project, we use estimates that come from five years of surveying (2009-2013).

Table 1 shows an example of two marginal characteristic distributions in some of the block groups we synthesized: household size and presence of children. This also demonstrates the format of marginal files as PopGen takes them. Each row is a block group in the state of California, and each column is the number of households in a block group that fall under a specific category. For example, there are 355 households of HHSIZE01 (one-person household)

in the first block group below. Every set of traits in one block group will add up to the same number– which is the total households in that block group. So adding the totals of every category of HHSIZE in one block group will give the same number as adding both categories of HHCHILD. The row with "bigint" in each column tells PopGen the type of data in the column (in this case, all are "big integer"). We have two marginal distribution files: one for households and one for individuals. Each will contain different traits chosen for that synthesis level.

Table 1. Example of Household Marginal Distributions

state	county	tract	bg	HHSIZE01	HHSIZE02	HHSIZE03	HHSIZE04	HHSIZE05	HHSIZE06	HHSIZE07	HHCHILD01	HHCHILD02
bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint	bigint
6	1	400100	1	355	703	112	114	9	0	0	1110	183
6	1	400200	1	142	177	90	58	0	1	0	363	105
6	1	400200	2	117	117	80	23	25	0	0	304	58
6	1	400300	1	97	234	88	5	21	0	62	351	156
6	1	400300	2	317	200	36	10	0	0	0	540	23
6	1	400300	3	265	116	88	36	38	0	0	445	98
6	1	400300	4	319	293	169	51	27	0	0	716	143
6	1	400400	1	346	266	70	57	21	13	0	613	160
6	1	400400	2	125	301	57	47	22	0	0	479	73
6	1	400400	3	224	168	117	68	0	0	0	446	131

3.2 California Household Travel Survey

The California Household Travel Survey (CHTS) is designed to support California's new transportation policy framework, building an inventory of travel behavior and taking into account possible use of new mobile technologies, as its Steering Committee clearly defined in the following paragraph during the inception of a partnership to build a consortium of agencies supporting CHTS.

"The purpose of the CHTS is to update the statewide database of household socioeconomic and travel behavior used to estimate, model and forecast travel throughout the State. Traditionally, the CHTS has provided multi-modal survey information to monitor, evaluate and make informed decisions regarding the State transportation system. The 2010 CHTS will be conducted to provide regional trip activities and inter-regional long-distance trips that will be used for the statewide model and regional travel models. This data will address both weekday and weekend travel. The CHTS will be used for the Statewide Travel Demand Model Framework (STDMF) to

develop the information for the 2020 and 2035 GHG emission rate analyses, calibrate on-road fuel economy and fuel use, and enable the State to comply with Senate Bill 391 (SB 391) implementation. The CHTS data will also be used to develop and calibrate regional travel demand models to forecast the 2020 and 2035 Greenhouse Gas (GHG) emission rates and enable Senate Bill 375 (SB 375) implementation and other emerging modeling needs."

One objective for the data collected in this household travel survey is to develop a variety of travel demand forecasting systems throughout the State and integrate land use policies with transportation policies (CALTRANS, 2016). Very important for regional agencies is the provision of suitable data that inform a variety of new model developments including activity-based models (ABM) and their integration with land use models at the state level and for each of the four major Metropolitan Planning Organizations (MPO) that surround Sacramento, San Francisco, San Diego, and Los Angeles. It is also the source of data for the many refinements of older four-step models and activity-based models in smaller MPOs and serves as the main source of data for behavioral model building, estimation of modules in other sustainability assessment tools, and the creation of simplified land use transportation models. Moreover, added details about a variety of choice contexts of households such as car ownership and car type are collected to develop a new set of prediction models to more accurately estimate emissions of pollutants at unprecedented levels of temporal and spatial resolutions.

CHTS meets the data needs criteria for a main core survey with satellite in-depth survey components similar in design to the ideal travel survey described at an international travel survey methods conference recently (Goulias et al., 2013). The CHTS databases include data collected by one contractor (NUSTATS) for the entire State of California and an added sample and supplement collected by another contractor (Abt-SRBI) for Southern California Association of Governments - SCAG. The databases include information about the household composition and facilities available, person characteristics of household members, and a single day place-based activity and travel diary. There are two stages in data collection: the first stage is called the recruitment and the second is called the retrieval. Sample selection was done using residential addresses and stratification to populate the final database with households that live in lower

density environments. Additional efforts were made to identify areas where response rate was expected to be low and intensify efforts to recruit residents in those areas. Details about the sampling method and efforts to make the resulting sample representative of the population in California can be found in Nustats, 2013. CHTS data were collected using paper and pencil as mail-in and mail-back survey, telephone (Computer Aided Telephone Interview, CATI), and the Internet using an interactive survey interface. CHTS also includes GPS data collection and a component administered by a different consultant for the California Energy Commission (CEC). All recruits were invited by an initial letter, TV videos (https://www.youtube.com/watch?v=h1KjCZQaDJ8). An effort was also made to contact community leaders and increase awareness of the public about the survey. A variety of monetary incentives were also used for different parts of the survey depending on the amount of time people needed to dedicate to record their responses. During the design and pretesting phases of the project, a high degree of harmonization among the three instruments of data collection was achieved using national guidelines (Goulias and Morrison, 2010, NUSTATS, 2013).

The CHTS (NUSTATS and Abt-SRBI) sample selection is a combination of exogenously stratified random and convenience sampling scheme (see NUSTATS Final Report, 2013). The final delivered databases for the statewide databases include slightly over 42,000 households (approximately 109,000 persons) for the core survey with most of their information complete. The core statewide CHTS travel days reported by respondents started on February 1, 2012 and ended in January 31, 2013 and include weekdays and weekends and spanned 58 counties of California and covering 366 days. CHTS is a joint effort among agencies to procure data collected using the same standards and funding was provided by Caltrans ($4,221,000), Strategic Growth Council ($2,028,000), Metropolitan Transportation Commission ($1,515,000), Southern California Association of Governments ($1,415,834), Council of Fresno County Governments ($49,500), Kern Council of Governments ($118,000), Association of Monterey Bay Area Governments ($183,810), San Joaquin Valley Air Pollution Control District ($150,000), Santa Barbara County Association of Governments ($33,000), Tulare County Association of Governments, ($49,500), and California Energy Commission ($250,000). This leads to approximately $240 per complete household record.

3.3 Seed

The seed is the sample that is used as the "building block" of the synthetic population. The program builds each block group's virtual population from households and individuals in the seed with the goal of matching the block group's marginal distributions as closely as possible. The sample we are using comes from the California Household Travel Survey (CHTS) reviewed above. This survey was collected between February 1, 2012 and January 21, 2013. It spanned 58 counties of California, and included weekdays, weekends, and holidays. (NUSTATS Final Report, 2013).

Table 2 shows the format for a household seed file. Each row is one household, which is linked to a household ID (hhid). In the person seed file, the household ID is also present in order to link the two together (Note: "serialno" is a placeholder that is always the same as hhid). Each column contains one characteristic (i.e. household size or presence of children), and the number corresponds to the category to which that household belongs. These categories are the same as those in the marginal distribution files. The spatial level of this data is the coarsest: we only give the program the Public Use Microdata Area (PUMA) number in which that household resides. This is to protect the privacy of the survey respondents, since there is a large amount of sensitive personal information present in the survey.

Table 2. Example of Seed Data

state	pumano	hhid	serialno	HHSIZE	HHCHILD
bigint	bigint	bigint	bigint	bigint	bigint
6	9502	1031985	1031985	2	1
6	7309	1032036	1032036	5	2
6	4702	1032053	1032053	6	2
6	8303	1032425	1032425	2	2
6	3751	1032558	1032558	1	1
6	6102	1033586	1033586	3	1
6	6506	1033660	1033660	1	1
6	7506	1033944	1033944	1	1
6	3750	1034462	1034462	2	1
6	3748	1034878	1034878	1	1

The original CHTS survey had 42,431 households, and 109,113 people. Unfortunately, not every participant responded to the questions we used as our control variables. We excluded the households and individuals that responded “Don’t Know” or “Refused to Answer” on any of the questions that we are using in this study. If an individual was excluded, so was the rest of their household. We ended up with 36,925 households and 94,901 individuals. Testing revealed that removing these households did not make a significant difference overall, so we proceeded with the reduced set of respondents.

Figure 1 shows the distributions of the variables we used from the CHTS. The same sociodemographic traits are used for both final runs of PopGen. At the household level, the traits used are householder age, presence of children, household size, and household income. At the person level, the traits used are age and gender.

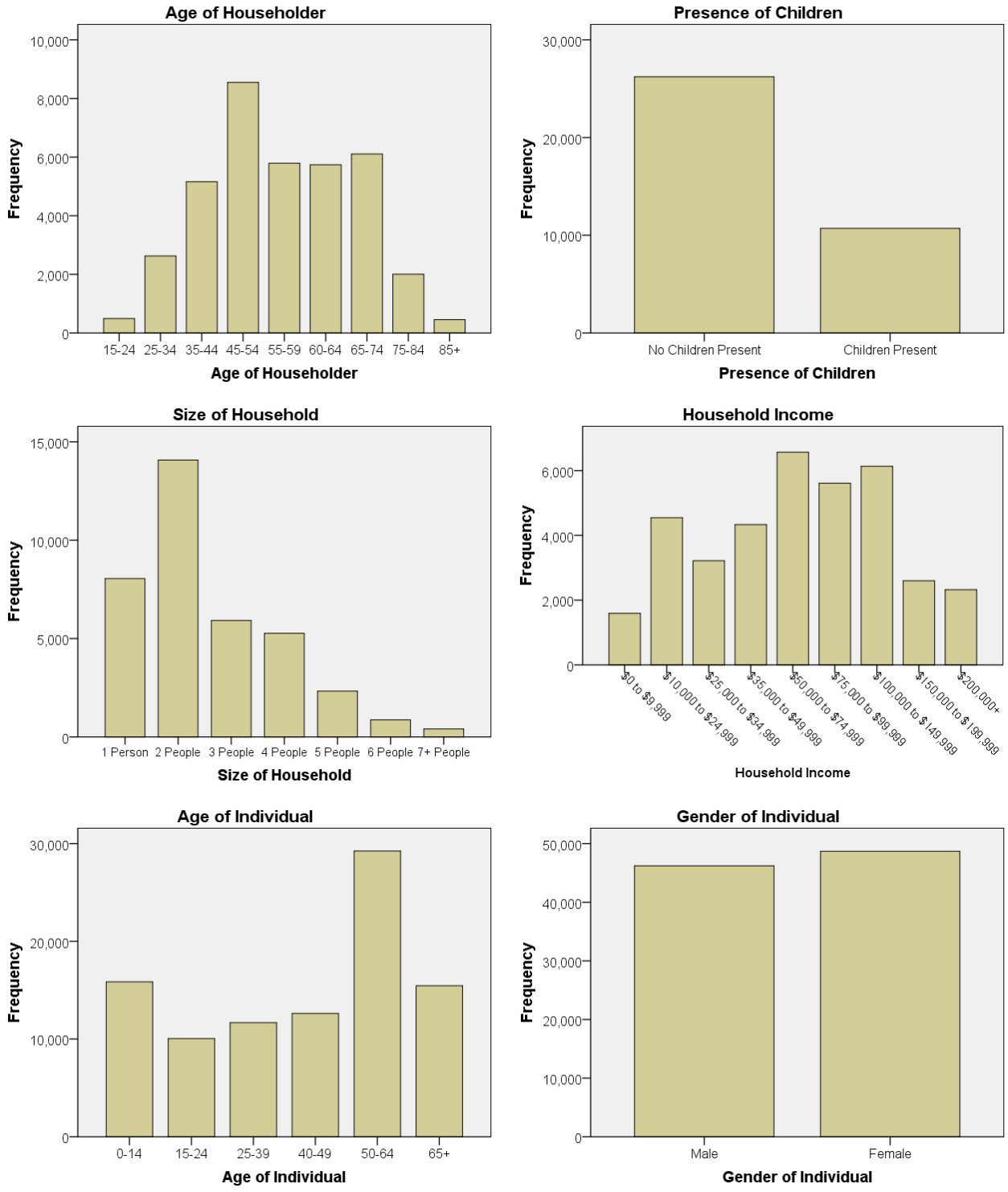


Figure 1. Seed Characteristic Distributions

3.4 Geographic Correspondence

The geographic correspondence file is the file that gives PopGen the list of areas for which it should synthesize populations. Table 3 shows how the geographic correspondence is formatted. The corresponding state, county, tract block group, and PUMA number are all listed, and the state and county names are also included.

Table 3. Example of the Geographic Correspondence File

county	tract	bg	state	pumano	stateabb	countyname
bigint	bigint	bigint	bigint	bigint	text	text
1	420100	1	6	101	CA	Alameda
1	420100	2	6	101	CA	Alameda
1	420100	3	6	101	CA	Alameda
1	420200	1	6	101	CA	Alameda
1	420200	2	6	101	CA	Alameda
1	420200	3	6	101	CA	Alameda
1	420300	1	6	101	CA	Alameda
1	420300	2	6	101	CA	Alameda
1	420300	3	6	101	CA	Alameda
1	420400	1	6	101	CA	Alameda

3.5 Output

The PopGen output consists of two datasets: the households and the individuals. As exemplified in Table 4 for the household file, every row is a household in a block group. In the person file, every row is an individual. The frequency gives us the number of times a household was used in a specific block group. The end result is a dataset with all of the traits specified by the marginal distributions recreated as closely as possible from the respondents to the travel survey.

Table 4. Example of PopGen Household Output

state	county	tract	bg	hhid	serialno	frequency	HHSIZE	HHCHILD
6	41	101100	1	1151723	1151723	1	1	1
6	41	101100	1	2845897	2845897	1	2	1
6	41	101100	1	2100372	2100372	1	1	1
6	41	101100	1	2621834	2621834	1	1	1
6	41	101100	1	1895207	1895207	1	2	1
6	41	101100	1	1214915	1214915	1	1	1
6	41	101100	1	1425753	1425753	1	2	1
6	41	101100	1	1885797	1885797	1	5	1
6	41	101100	1	2060325	2060325	1	2	1

4. POPGEN WITH LAND USE INDICATORS

In this project, we seek to improve synthetic population generation by accounting for the effect of land use on travel behavior. To do this, we divide the state into four sections based on activity density in order to provide a relatively equal-sized seed for each of the synthetic population runs. To ensure the division made sense, we estimated models for travel behavior variables and found significant differences between the behaviors of people living in different PUMA groupings. This difference in behavior corresponds to travel behavior of people living in high density environments like a city center, medium density like a suburb, and low density like a rural environment. The corresponding cutoff points in employees per square kilometer (emp/km2) used are: 37, 360, 1090 (25%, 50%, 75% quartiles of PUMA data by HH). Below we describe the method and the final classification of PUMAs used in synthetic population generation, followed by its use in PopGen and a description of our results.

4.1 Activity Density Surface Estimation

The geo-coded firm-level data for this research is extracted from the 2013 NETS database that includes more than 6 million business establishments in California with longitudinal information about their industrial type, location, headquarters and performance over the period of 1990-2013. The NETS database is constructed by taking a series of 'snapshots' based on the Dun and Bradstreet (D&B) archival national establishment data (*Walls, 2007*). The unit of observation in the NETS database is a business establishment that produces goods or services at a single

physical location – for example, a single store. This database tracks every establishment from its birth, through any physical moves it makes, capturing any changes in ownership and recording the establishment's dissolution if it occurs. NETS records information on location of the establishment, employment, sales and industry type for each year. From the 6.7 million unique business establishments in the NETS database, we extracted a database consisting of approximately 3 million business establishments in California that were active in 2012 to coincide with the California Household Travel Survey that was collected between February 1, 2012 and the end of January 2013.

Using the NETS data, we can compute multiple possible measures of land use, as it is likely impossible to represent all aspects of land use with a single variable. Because this project ultimately aims to improve measurement of people's activities and travel behavior, we require a land use metric that is specifically targeted at variation in this sort of behavior. We choose employee density as a proxy measure for activity density, since most activities that require travel (which is to say, activities outside the home) involve trips to locations where people are employed (e.g., a restaurant, a post office, a doctor's office).

The NETS dataset is delivered in a tabular format, with each row containing the permanent characteristics and year-by-year employee counts and sales for a single business establishment with a unique DUNS number. Additional tables contain business categories – providing a 6-digit North American Industrial Classification System (NAICS) code for each year the business existed – and a record of relocation events. To extract the relevant information for this application, we performed the following tasks in R: 1) extract 2012 employee totals and final (2013) locations for each business from the main table; 2) update with 2012 business locations for businesses that moved between 2012 and 2013; and 3) export data as a shapefile for use in ArcGIS.

The next step is to convert business establishment / employment count points to a map of activity density measured consistently across the entire state. Land use is an areal property, so it can be modeled as either a continuous surface (raster) or a set of bounded units (vector polygons).

Because business locations were stored as point features, they must be converted into one of these formats to be useable for land use estimation.

The most straightforward method would be to choose a single polygon scale of aggregation such as zip codes, block groups, or public use microdata areas (PUMAs) and sum up the employees of all businesses located within each polygon, but this process has two main shortcomings: it produces edge effects, and it will perform poorly in high density areas. Edge effects become a problem when a business is located near the border between two zones. By simply aggregating to containing polygons, this business would be counted exclusively towards one, even though it should relate almost equally to the land uses of both. In high density areas, the simple aggregation process may underestimate the activity density of residential areas adjacent to a dense business zone and in areas where a small area of lower employment density doesn't represent an actual change in local land use over space. These problems are particularly important because census polygons are designed to equalize population at home locations, not the locations of other activities.

Instead of relying on simple counts, we employ the kernel smoothing process implemented in ArcMap to estimate an activity/land use density surface from business establishment locations. Kernel density functions fit a smooth, curved surface over the input points (in this case businesses). Each point's contribution to the density surface is highest at its location and diminishes with increasing distance from the point, reaching zero at the distance from the point specified by the bandwidth/maximum distance parameter. ArcGIS uses a quartic kernel function to calculate the density. The total volume under each point's kernel density surface is equal to the point's population field value (in this case, the business's number of employees in 2012). The total density in the output raster is calculated by adding the values of all the kernel surfaces at the center of each raster cell ("How Kernel Density works," 2014, see ARCGIS Resource Center, Desktop 10).

By smoothing employee/activity density over space, we seek to produce a more accurate representation of land use that can be used for statewide analysis. In addition to eliminating the issues described above, smoothing addresses the error caused by small inconsistencies in the

precision/accuracy of business establishment coordinates provided in NETS, which are more accurate for newer business locations than for ones that have existed since 1990.

To produce a final activity density map, we tested a range of kernel bandwidths (from 200m to 20km) and chose a 2km bandwidth for the final product. This kernel balances the benefits of the detailed but irregular surfaces provided by smaller bandwidths and the smooth but overgeneralized surfaces produced by larger kernels. The choice of kernel bandwidth is highly dependent on the specific application of the density surface. In final analysis, the 2km bandwidth also seemed appropriate because it represents a good portion of walking trips and the highest-density part of each point's kernel is a reasonable size for neighborhood scale. Figure 2 shows the resulting density surface.

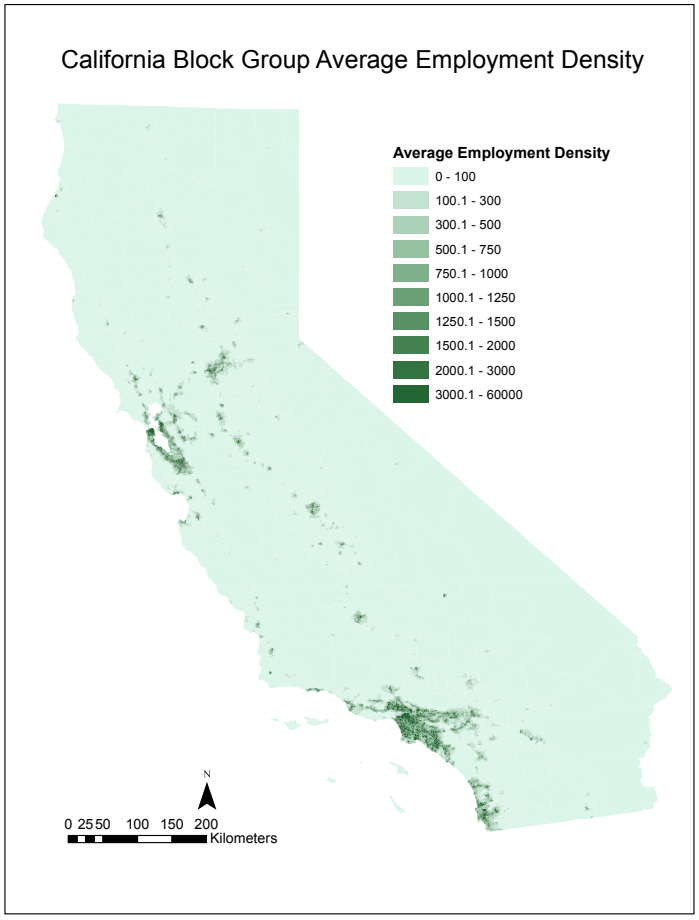


Figure 2. Average Employment Density at the US Census Block-Group Level in 2012

PopGen spatially categorizes the "seed" (CHTS) inputs at the PUMA level, so the activity density surface must be aggregated to these zones; we do this by extracting an average value of the surface over each PUMA's area. This generalization is implemented in ArcMap as follows: start with a sufficiently fine-scale raster of the 2km kernel employee density (50m pixels in this case); convert the map of polygons to a raster with cells aligned to the kernel density cells; then calculate the average value of the density raster cells that match cells in each PUMA's raster representation.

Once the average density has been calculated for each PUMA, the next step is to divide the PUMAs into groups that can be treated separately in synthetic population generation. To make the synthetic population generation process work right, it is important to ensure each population has roughly the same number of households in its seed, so the zones are classified according to quantile values for the households of the PUMA-level activity densities. The average density of each PUMA is joined to the households that are located within it. These household PUMA densities are grouped into quartiles, as shown in Figure 3, and the breaks between these quartiles are used to assign each PUMA to a land use group. Figure 4 shows the final map of the divisions. The segmentation used here is motivated by the different travel behavior of households in California. Households in rural environments travel longer distances and make fewer trips, while households in centrally located and high density environments make more trips but cover shorter distances.

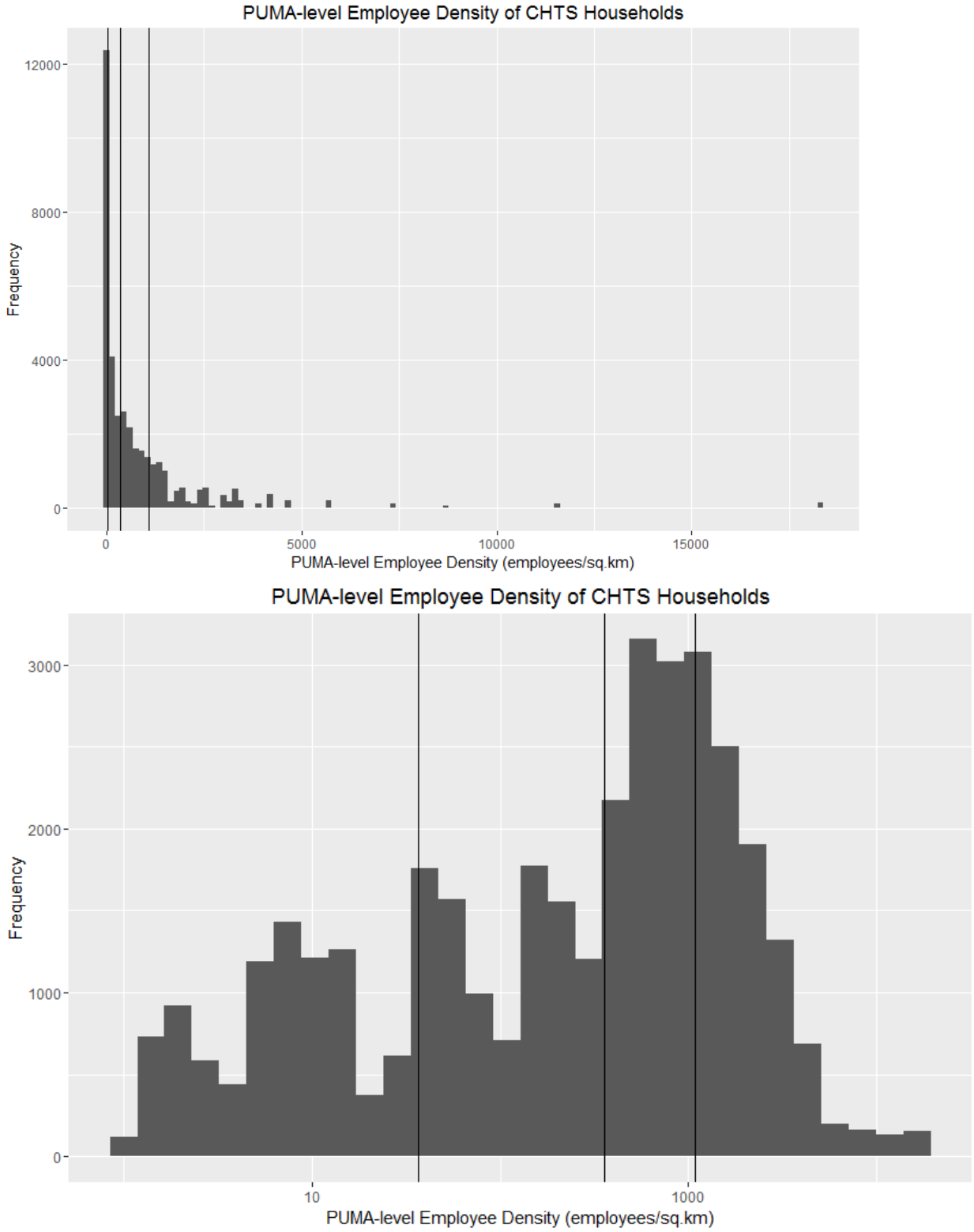


Figure 3. CHTS Households by household-specific density and logarithm of density

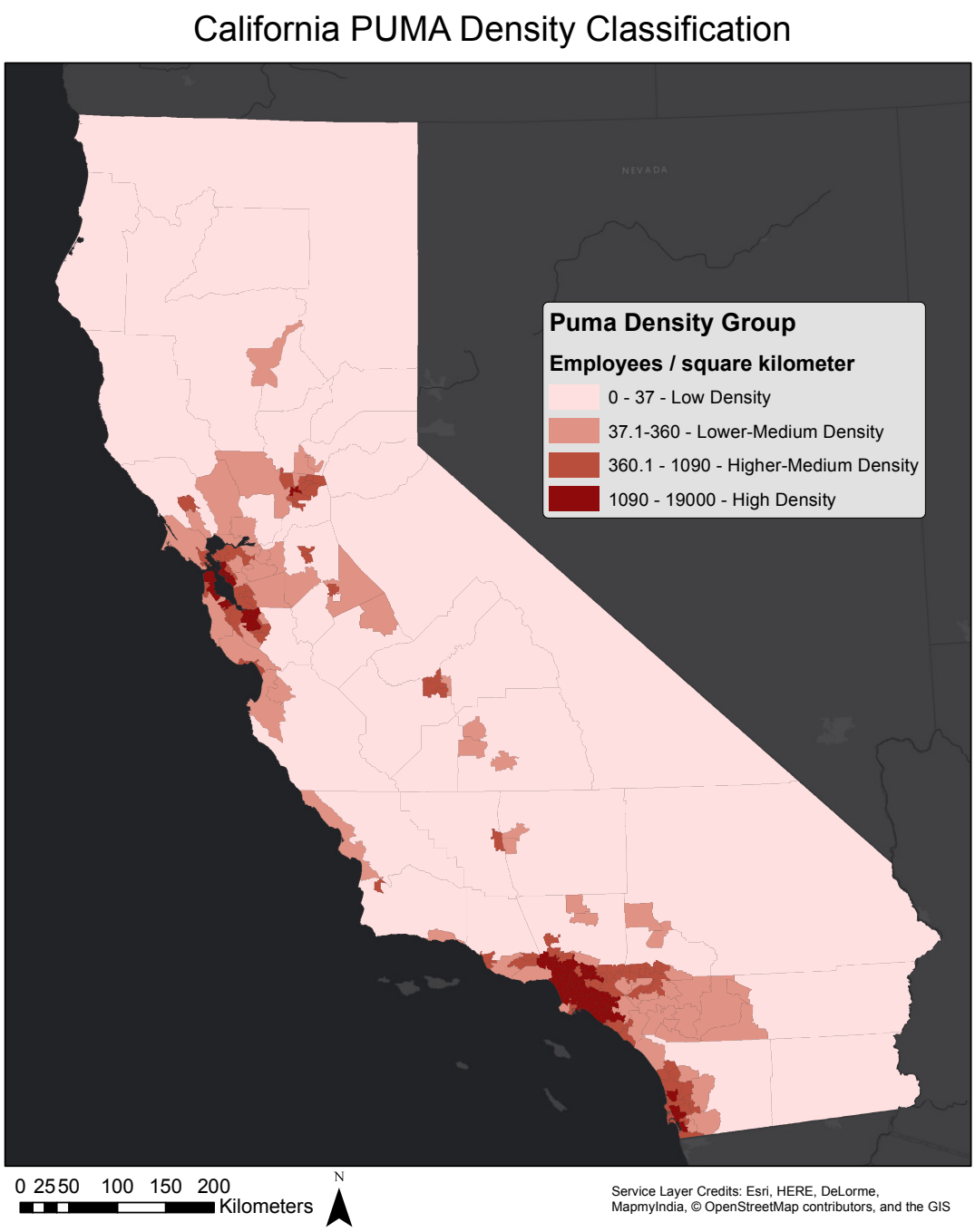


Figure 4. The Four Groups of PUMA Classification Used in Synthetic Population

Our approach has some approximations because polygon-to-raster conversion necessarily entails a loss of spatial precision, and because a spatial average of density in an area may not match the average density experienced by people in the area if the people are not uniformly distributed in space. These issues are not a major concern for this analysis because they mainly affect polygons

that are easy to classify. Spatial aggregation errors caused by producing raster representations of polygons can be significant for very small polygons, since there are fewer values to average, but the smoothing introduced by the 2km kernel means there is very little difference between the density values of adjacent 50m pixels and aggregation error is minimal. Unevenness of human activity and natural attributes over space can be substantial within larger polygons (and assuming that zone-wide values are fully representative of the entire region they cover is called ecological fallacy), but because census polygons are designed to have roughly equal populations, the largest polygons will always be located in extremely low-density parts of the state. The classification of these large, low-density polygons is not at all sensitive to slight changes in the way density is measured. It is possible that zones could be misclassified in an extremely low-population / high-employment part of a large city, but this issue did not arise in our analysis, likely in part thanks to the smoothing introduced by the kernel process. The final proof that this method produced a useful result is that the map of classified polygons is very sensible and consistent over space, which indicates that any effects of aggregation error are erased by the classification process.

4.2 Implementation in PopGen

To use these land use categories in population synthesis, we take the input files we use in the "base method" and divide them based on the groupings created using R. We divide the geographic correspondence file into four pieces. This tells PopGen which block groups in the marginal distribution files to replicate and which to ignore. There is no need to divide the marginal distributions, since PopGen will only use distributions from the block groups indicated in the geographic correspondence file. We also divide the CHTS data by land use category. This process results in four geographic correspondence files, four household seed files, and four person seed files. We run PopGen four times, synthesizing the population of a different set of block groups each time. We take the resulting output files and combine them, giving us two final files that resemble the files received from the "base method."

We hope to see that travel traits transferred over will be better replicated when land use is included. This should happen because only households from one quartile grouping (so with similar business densities) will be used to populate block groups of the same grouping. If

land use characteristics affect travel behavior, and if employee density is a good proxy for land use, then we will see better travel behavior replication.

4.3 Summary of Results

Table 5 shows a comparison of the output of PopGen with and without land use. In terms of replicating the control variables, there is very little difference between the two runs. Both are highly accurate for household-level synthesis and less accurate in person-level synthesis. This is to be expected, as it is built into the way PopGen operates. The program prioritizes replicating the households, followed by the people as closely as possible. The exclusion of group quarters also contributes to the inaccuracy at the person level since the person totals of the marginal distributions includes individuals in group quarters. We do not replicate group quarters in this study, so we do expect the person-level numbers to be lower.

5. CHTS AND HOUSEHOLD TRAVEL

Two of the most important components of the California Household Travel Survey for this project are the household and person demographics, which are what we used to extract the control variable data, and the travel diary. Respondents were asked to record their travel for one day, noting every place they went, the modes of travel they used, and with whom they traveled. For analyzing the effectiveness of the new method of PopGen, we need to compare the way it creates travel traits. The travel diary will be the way we do this.

5.1 Synthetic Travel Traits

We were able to transfer travel traits to the synthetic population using the CHTS travel diary by linking the two together by the household and person IDs. The following traits were calculated at both the household and person level for each block group: number of trips by mode, total trips by all modes, miles traveled by mode, total miles traveled by all modes, and number of people who did not travel. For the miles traveled, there were some complications because of incorrectly calculated trip distances that were impossibly large. To work around this, any non-airplane trip that was found to have an average over 85 miles per hour was selected as incorrectly calculated, and its value was replaced with the mean rate so that it would not disrupt the distribution.

Table 5. Results of PopGen

Category	Category Definition	Actual	NO LAND USE Synthesized	NO LAND USE %Diff	LAND USE Synthesized	LAND USE %Diff
			Household Level Variables			
Age of Householder						
1	15-24 years old	454876	454850	**-0.01%**	454577	**-0.07%**
2	25-34 years old	2007248	2007251	**0.00%**	2007334	**0.00%**
3	35-44 years old	2504528	2504525	**0.00%**	2504536	**0.00%**
4	45-54 years old	2735097	2735099	**0.00%**	2735145	**0.00%**
5	55-59 years old	1216284	1216284	**0.00%**	1216312	**0.00%**
6	60-64 years old	1054988	1054984	**0.00%**	1055050	**0.01%**
7	65-74 years old	1376665	1376660	**0.00%**	1376668	**0.00%**
8	75-84 years old	823979	823966	**0.00%**	823972	**0.00%**
9	85 and over	368795	368791	**0.00%**	368793	**0.00%**
Presence of Children						
1	No children in household	7928732	7936004	**0.09%**	7935523	**0.09%**
2	Children in household	4613728	4606406	**-0.16%**	4606864	**-0.15%**
Size of Household						
1	1 person	3040221	3041873	**0.05%**	3041900	**0.06%**
2	2 persons	3749732	3753438	**0.10%**	3753538	**0.10%**
3	3 persons	2048520	2049497	**0.05%**	2048860	**0.02%**
4	4 persons	1901098	1899772	**-0.07%**	1899900	**-0.06%**
5	5 persons	995789	993997	**-0.18%**	993029	**-0.28%**
6	6 persons	447064	445646	**-0.32%**	445551	**-0.34%**
7	7 or more persons	360036	358187	**-0.51%**	359609	**-0.12%**
Household Income						
1	< $10,000	714855	716716	**0.26%**	717030	**0.30%**
2	$10,000 - $24,999	1848317	1851861	**0.19%**	1851660	**0.18%**
3	$25,000 - $34,999	1137796	1138119	**0.03%**	1138172	**0.03%**
4	$35,000 - $49,999	1541102	1541308	**0.01%**	1541665	**0.04%**
5	$50,000 - $74,999	2122567	2123979	**0.07%**	2123585	**0.05%**
6	$75,000 - $99,999	1551514	1552675	**0.07%**	1552191	**0.04%**
7	$100,000 - $149,999	1870135	1869832	**-0.02%**	1869729	**-0.02%**
8	$150,000 - $199,999	848259	847809	**-0.05%**	847530	**-0.09%**
9	≥ $200,000	907915	900111	**-0.86%**	900825	**-0.78%**
			Person Level Variables			
Age						
1	Under 15 years old	7610186	7141836	**-6.15%**	7141398	**-6.16%**
2	15 to 24 years old	5552324	5019473	**-9.60%**	4971582	**-10.46%**
3	25 to 39 years old	7928889	7230698	**-8.81%**	7221969	**-8.92%**
4	40 to 49 years old	5238856	4766684	**-9.01%**	4773905	**-8.88%**
5	50 to 64 years old	6753676	6245894	**-7.52%**	6285048	**-6.94%**
6	65 and more years old	4444027	4162848	**-6.33%**	4195747	**-5.59%**
Gender						
1	Male	18611994	17231704	**-7.42%**	17218766	**-7.49%**
2	Female	18915964	17335729	**-8.35%**	17370883	**-8.17%**

5.2 Correlation of Land Use with Behavior

Land use and behavior should be correlated. It is not a new idea that a person's surroundings will influence the places that he or she goes. Numerous businesses close to the home location should

increase the likelihood of staying near the house and utilizing the nearby businesses instead of going further away. A rural household with less access to nearby businesses will probably need to travel further to get to every activity they are interested in. This could mean they travel longer, or that they do not travel as much and just skip those activities that are too far away. The following is what we expect from urban areas as opposed to rural areas: more frequent, shorter trips, more walking/bicycling, fewer car trips, and more public transport trips (Stead & Marshall, 2001).

Many persons and households make no trips. This means we have a large number of observations at zero. To account for this in a regression model we employ a Tobit regression. Figure 5 shows the results of a Tobit model that we ran to determine whether or not the land use categories are significant in determining vehicle miles traveled, and the results show that the categorization has a very strong influence. Group 1 is what we call the urban category, group 2 is suburban, group 3 is exurban, and group 4 is rural. The model shows that Group 2 travels 14.5 miles more than group 1, group 3 travels 18.9 miles more than Group 1, and Group 4 travels 28.9 miles more than Group 1. As expected a lower employee density is correlated with a higher number of miles traveled in a car. The model results here provide support for our choice of variables to classify the State into different land use groups. The significance of the coefficients also provide evidence that the sociodemographic characteristics used in the synthetic population generation are also the right variables for transferring behavior from one place to another.

```
Pearson residuals:
             Min      1Q  Median      3Q      Max
mu       -49.450 -0.4545  0.1607  0.6623    8.048
loge(sd)  -1.011 -0.8995 -0.7733 -0.2150 1465.229

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -82.905055   4.615196 -17.963  < 2e-16 ***
(Intercept):2    4.871547   0.005009 972.479  < 2e-16 ***
HHAGE           -4.655484   0.476230  -9.776  < 2e-16 ***
HHCHILD         12.226657   2.484625   4.921 8.61e-07 ***
HHSIZE          15.314120   0.874545  17.511  < 2e-16 ***
HHINC            8.479304   0.378448  22.405  < 2e-16 ***
HHVEH           20.334277   0.938529  21.666  < 2e-16 ***
puma_gBfactor2  14.558495   2.076192   7.012 2.35e-12 ***
puma_gBfactor3  18.949949   2.072852   9.142  < 2e-16 ***
puma_gBfactor4  28.924348   2.118543  13.653  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of linear predictors:  2

Names of linear predictors: mu, loge(sd)

Dispersion Parameter for tobit family:   1

Log-likelihood: -194169.4 on 73840 degrees of freedom

Number of iterations: 12
```

Figure 5. Tobit model testing influence of land use categories on vehicle miles traveled (VMT).
Note: the "puma_gBfactor#" variables are the land use categories. puma_gBfactor1 (the control, so not included) is the urban category, and puma_gBfactor4 is the rural category.

6. TRANSFERABILITY & CONCLUSIONS

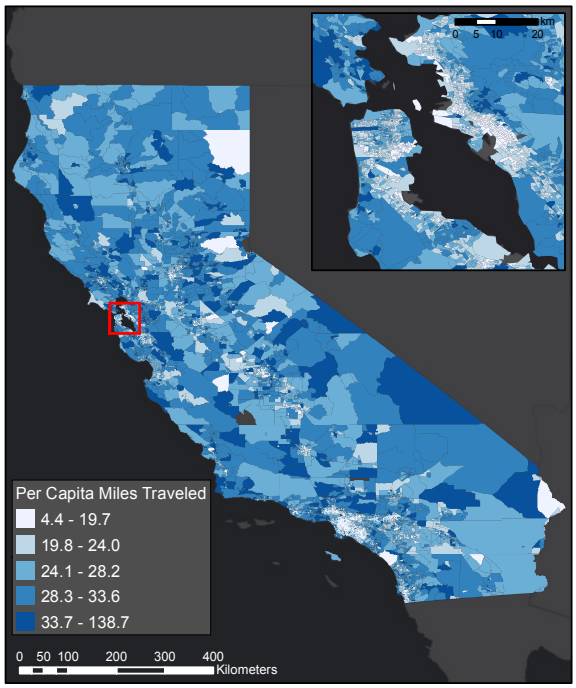
6.1 Comparison of PopGen with and without Land Use

Figure 6 contains maps comparing the transfer of travel traits to the synthetic populations created with and without land use. There is much less "random noise" in the land use maps, and the behavior patterns seem to be related to proximity to urban areas. These patterns also correspond to the relationship between "urban-ness" and travel behavior that we hope to see. The three sets of maps look at common travel traits. For rural populations, the land use population maps show fewer trips, more miles traveled, and less walking trips in rural areas – and the opposite in urban areas. These results show that there are patterns being picked up by including land use that

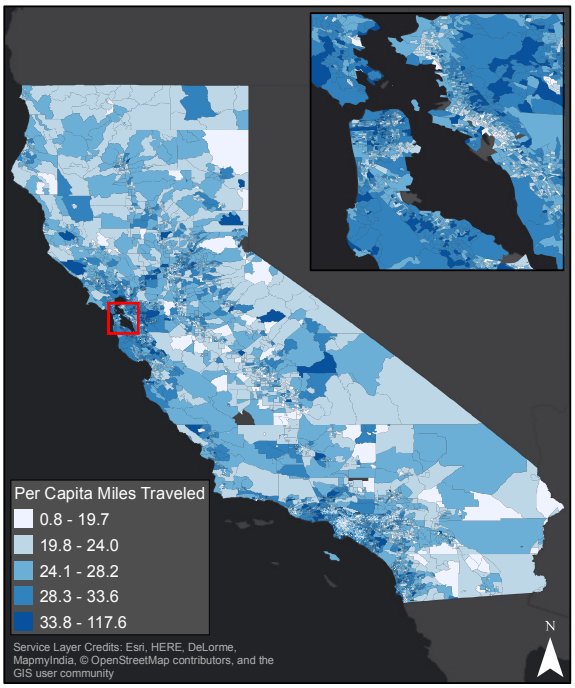
would not otherwise make it to the synthetic population. These results mean that land use will make these models much more valuable and reliable for modeling travel behavior.

Per Capita Miles Traveled

Land Use

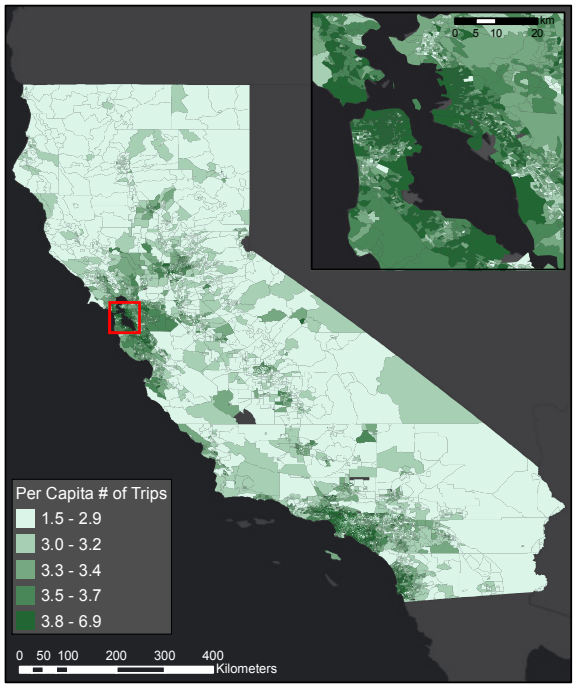


No Land Use

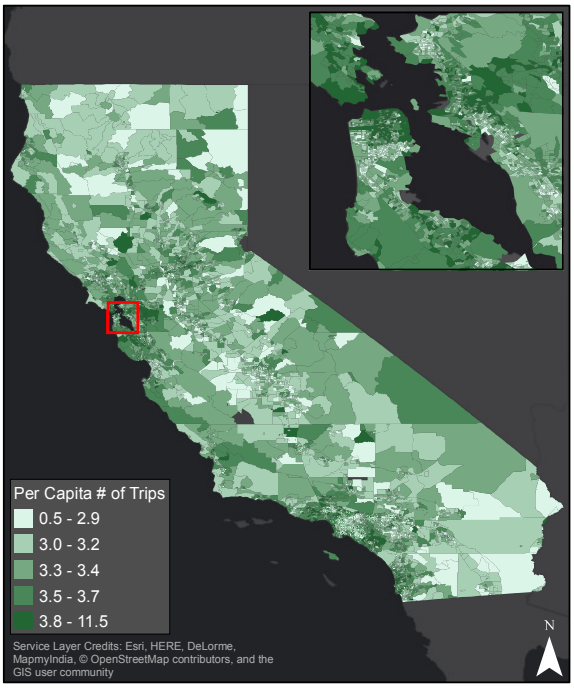


Per Capita Number of Trips

Land Use



No Land Use



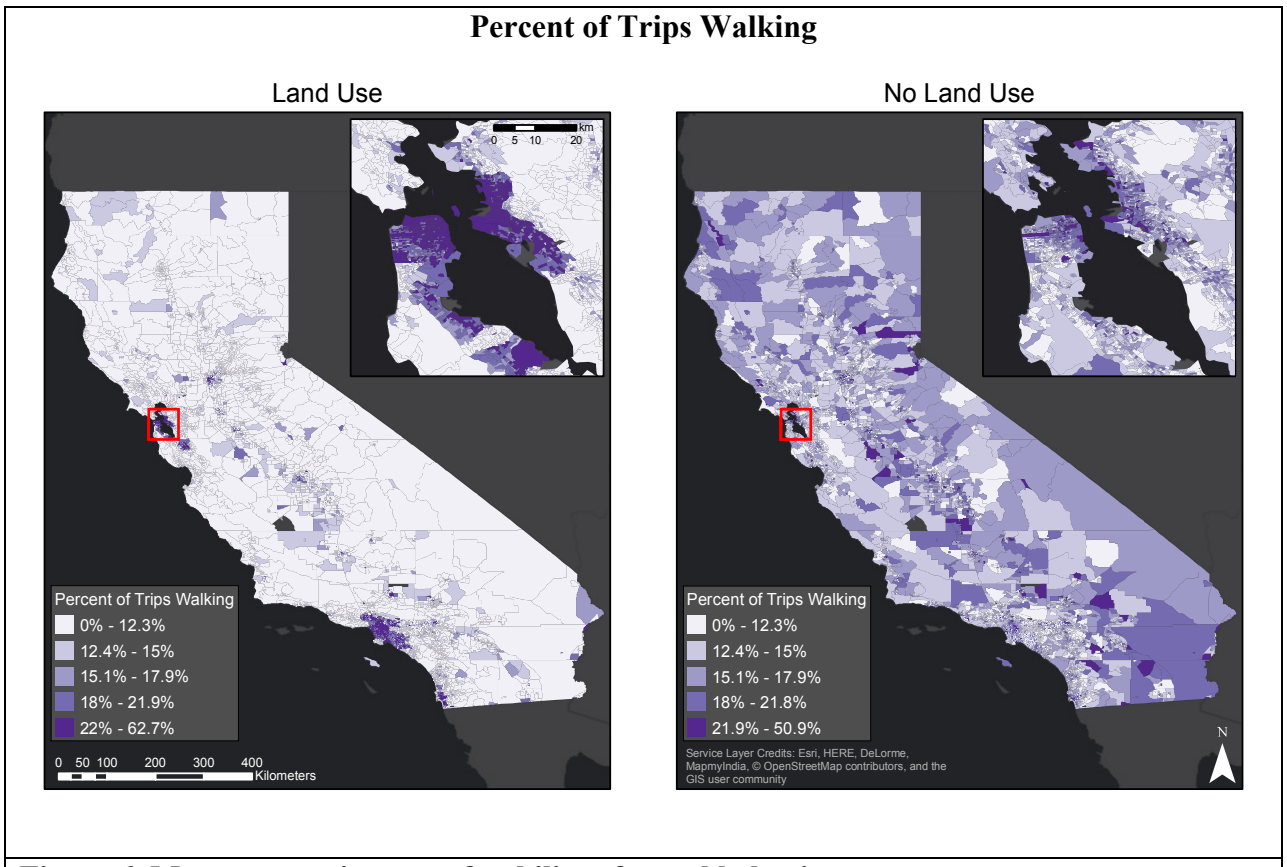


Figure 6. Maps comparing transferability of travel behavior

6.2 Examples and Recommendations for Transferability

Numerous travel traits can be transferred to a synthetic population – as we did. Miles traveled by mode, number of trips by mode, and number of people that do not travel. We have aggregated these traits to the block group level, but this is not necessary. The output of PopGen is at the individual household or person level. If it is desired for future research, the traits can be kept at the individual household or person level. The only limitation of this is computational. For the entire state of California, working with a database of 12 million households and/or 36 million individuals presents challenges for most computers and programs. For this reason, aggregation to the highest spatial resolution of value to future research is recommended.

6.3 Concluding Remarks

With respect to the research questions we posed in the introduction, we were able to develop a simple method based on a small set of land use groups that capture behavioral heterogeneity.

This was tested with regression models, with one example reported here for VMT and illustrated with maps. We also verified the effectiveness of this approach showing substantial spatial differences in the synthetic population and the behavior of the synthetic households created here. The most important differences emerge from the ability of the method to distinguish between residents of urban environments and rural environments while accounting for their social and economic differences.

The method here worked well with just four groups of geographic subdivisions using existing open source software (PopGen). However, there are many improvements that can be developed. First, we can add more social and demographic variables at the household and person levels to account for behavior of special groups of people. Second, we can develop more detailed land use indicators based on different industries (e.g., retail, education, health) to better represent travel behavior. Third, we can develop regional synthetic populations to capture traits that are not reflected in sociodemographics and/or the land use indicators we use here. Moreover, we can also use a variety of indicators for the supply of transportation infrastructure to further improve transferability.

7. References

Auld, J., Mohammadian, A., and Wies, K. Population Synthesis with Regional-Level Control Variable Aggregation. Paper presented at the 87th Annual Transportation Research Meeting, Washington D.C., January 2008.

Beckman, R.J., K.A. Baggerly, and M.D. McKay (1996) Creating Synthetic Baseline Populations. Transportation Research Part A: Policy and Practice, 30(6), pp. 415-429.

Konduri K.C., D. You, V.M Garikapati, and R.M. Pendyala (2016) Application of an Enhanced Population Synthesis Model that Accommodates Controls at Multiple Geographic Resolutions. Paper 16-6639 Presented at the 2016 Annual Transportation Research Board Meeting, Washington D.C.

Pritchard, D. R., & Miller, E. J. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. *Transportation*, *39*(3), 685-704.

Zhu, Y., & Ferreira, J. (2014). Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. *Transportation Research Record: Journal of the Transportation Research Board*, (2429), 168-177.

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243-263.

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.

NUSTATS (2013). 2010-2012 California Household Travel Survey Final Report: *Version 1.0. June 14. Submitted to the California Department of Transportation. Austin, TX*.

Pendyala, R., Bhat, C., Goulias, K., Paleti, R., Konduri, K., Sidharthan, R., ... & Christian, K. (2012a). Application of Socioeconomic Model System for Activity-Based Modeling: Experience from Southern California. *Transportation Research Record: Journal of the Transportation Research Board*, (2303), 71-80.

Pendyala, R.M. , C. R. Bhat, K. G. Goulias, R. Paleti, K. Konduri, R. Sidharthan , and K. P. Christian. (2012b) SimAGENT Population Synthesis. Phase 2 Final Report 3 Submitted to SCAG, March 31, 2012, Santa Barbara, CA.

PopGen: Population Generator. Retrieved from http://urbanmodel.asu.edu/popgen.html

Ravulaparthy S. and K.G. Goulias (2011) Forecasting with Dynamic Microsimulation: Design, Implementation, and Demonstration. Final Report on Review, Model Guidelines, and a Pilot Test for a Santa Barbara County Application. University of California Transportation Center (UCTC) Research Project. Geotrans Research Report 0511-01, May, Santa Barbara, CA

Stead, D., & Marshall, S. (2001). The relationships between urban form and travel patterns. An international review and evaluation. European Journal of Transport and Infrastructure Research, 1(2), 113-141.

Walls, D. (2007). National Establishment Time Series Database: Data Overview. Presented at 2007 Kauffman Symposium on Entrepreneurship and Innovation Data.

Ye, X., K. Konduri, R. M. Pendyala, B. Sana and P. Waddell. A Methodology to match Distributions of both Household and Person Attributes in Generation of Synthetic Populations. Paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C., January 2009.