UC Santa Barbara

UC Santa Barbara Electronic Theses and Dissertations

Title

Enhancing Population Synthesis using Land Use Indicators

Permalink

https://escholarship.org/uc/item/3cb9k2hm

Author

McBride, Elizabeth

Publication Date

2017

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Santa Barbara

Enhancing Population Synthesis using Land Use Indicators

A Thesis submitted in partial satisfaction of the

requirements for the degree Master of Arts

in Geography

by

Elizabeth Callahan McBride

Committee in charge:

Professor Konstadinos Goulias, Chair

Professor Susan Cassels

Professor Krzysztof Janowicz

January 2018

The thesis of Elizabeth Callahan McBride is approved.

_____
Susan Cassels

_____
Krzysztof Janowicz

_____
Konstadinos Goulias, Committee Chair

January 2018

Enhancing Population Synthesis using Land Use Indicators

Copyright © 2018

by

Elizabeth Callahan McBride

ACKNOWLEDGEMENTS

I would like to thank the members of the GeoTrans Lab for their advice, support, and contributions to this project. This thesis would not be possible without Adam Davis, Dr. Jae Hyun Lee, and, of course, my advisor and committee chair Professor Konstadinos Goulias. I would also like to acknowledge the financial support of University of California Center on Economic Competitiveness in Transportation (UCCONNECT), the University of California Transportation Center (UCTC), and the California Department of Transportation (Caltrans) during the projects that went into this thesis. Finally, I would like to thank my committee members Professor Susan Cassels and Professor Krzysztof Janowicz for their thorough and thoughtful reviews of this thesis. Their commentary helped me immensely in the final stages of writing and editing.

ABSTRACT

Enhancing Population Synthesis using Land Use Indicators

by

Elizabeth Callahan McBride

This paper looks at methods of incorporating land use into population synthesis. Although it is something that has not been explored in past research, this paper will show that it is imperative for synthesizing populations that represent travel behavior patterns well. The goal of the paper is to derive a land use classification scheme that shows significant differences in travel behavior and enhances population synthesis in its ability to represent travel behavior. Three different methods were devised and implemented, then they were compared to determine the ideal method. The paper concludes that using latent profile analysis as a means for classification is an ideal method that allows for flexibility in geographical areas synthesized and variables used.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1 Introduction

Population synthesis is the generation of a synthetic population with the goal of replicating a real population of interest as closely as possible. It uses known unidimensional distributions of sociodemographic variables in given areas and estimates unknown multidimensional relationships among those variables using sample microdata of households and individuals to populate the geographical areas. The process generates a synthetic population with comprehensive data on attributes of interest that can also be correlated with the geographic context in which behavior is situated. It is the first step in activity-based microsimulation (ABM) models in travel demand modeling, in which individuals traveling across a network are modeled for an area of interest. ABM models first model the propensity of people to participate in specific activities, then derive travel among different activity locations. A key informant of this propensity is land use. In this thesis, the term Land Use means the development of land characterized by the type, distribution, and density of resident businesses (Waddell, 2002). This is important for ABMs because there is a systematic relationship between land use types (e.g., retail) and people's activities (e.g., shopping).

1.1 Problem Statement

The overarching goal of this thesis is to test whether synthetically-generated populations of California are enhanced (in terms of their ability to accurately capture travel behavior information) by the inclusion of land use measures during the synthesis process.

There is very little past work on incorporating geographic information into the population synthesis process, despite the fact that land use characteristics of an area are a

large determinant in how people travel. For example, home location influences how people travel, but people also choose to live in places that allow traveling by specific means. This is called residential self-selection, and it is a recognized factor that adds complexity in modeling and simulation (Bhat & Eluru, 2009; Cao, Mokhtarian, & Handy, 2009; Mokhtarian & Cao, 2008). For this reason, land use is a key informant that can be added to population synthesis when the aim is to transfer behaviors from a microdata sample to the entire synthetic population and/or simply improve the creation of the synthetic resident population in an area. This thesis attempts to address the lack of work on the topic by testing methods of including geographic information in population synthesis in the form of land use, and providing suggestions for the best methods to include it.

I hypothesize that including spatial information into population synthesis in the form of land use information will enhance the transferability of travel behavior traits to a synthetic population of California. I also hypothesize that determining land use classes based on statistical methods is a better approach to including land use in population synthesis than a simpler quartile-based classification. Before testing these hypotheses, a special type of regression model called a Tobit model is used to analyze the sample microdata. In this way, I demonstrate that land use is a significant factor for travel behavior prediction. Then, different techniques are employed to analyze and integrate land use into population synthesis, including a simple classification method using quartiles as well as a more complex method using latent profile analysis. The results of all the techniques I used are compared at the end.

The structure of the rest of the thesis is as follows. Chapter 2 presents a literature review that covers population synthesis generally, enhancements to synthesis methods, and latent

profile analysis. Chapter 3 introduces the methodology, including the population synthesis software used, and details about the various methods for land use classification. Chapter 4 reviews the results of analyzing the synthetic populations. Finally, Chapter 5 offers conclusions drawn from the analysis.

2 Literature Review

The following literature review will examine previous work on population synthesis.

2.1 General Topic Area: Population Synthesis

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 (Ravulaparthy & Goulias, 2011).

A family of population synthesis methods emerged after 1996 based on Beckman *et al.*, using the iterative proportional fitting (IPF) algorithm (Beckman, Baggerly, & McKay, 1996). This method uses categorical variables for which there are known values for the area of interest to populate. It selects people from a survey, providing microdata at the level of individuals and households, that is used as the "seed", and then uses them to populate the area. The method has been adopted and modified by a number of researchers over the years.

Next, the idea of multiple control levels became important. Originally, population synthesis was only run using household-level control variables to determine the selection and distribution of households into the synthetic population. Today, the incorporation of control variables at both the household- and person-level is becoming increasingly common. Although IPF is still performed without multilevel controls (Adiga et al., 2015), the use and implementation of multilevel controls is an area of interest for population synthesis researchers (Auld & Mohammadian, 2010; Konduri, You, Garikapati, & Pendyala, 2016; Pendyala, Konduri, & Christian, 2011; Zhu & Ferreira Jr., 2014).

Methods with multilevel controls that do not involve IPF have also been developed, including Markov Chain Monte Carlo approaches (Casati, Müller, Fourie, Erath, & Axhausen, 2015; Farooq, Bierlaire, Hurtubia, & Flötteröd, 2013) and a fitness function-based method (Ma & Srinivasan, 2015).

Guo and Bhat (2008) 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 (2008) 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 (2008) 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 the seed (US Census Public Use Microdata Sample in this case) 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 reviewed below.

Building on the IPF procedure for population synthesis, Auld *et al*. (2008) 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*. (2008) adopt the same method for creating a multidimensional distribution table as in other population synthesizers (Beckman et al., 1996; Guo & Bhat, 2008). 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*. (2008) 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*. (2008) 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*. (2008) 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*. (2008) 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*. 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 (Auld & Street, 2008; Beckman et al., 1996; Guo & Bhat, 2008), 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 block group. 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 any other type of survey containing microdata with persons and their households) to generate the

synthetic population. The approach Ye *et al*. (2009) adopt is similar to that of Beckman *et al*. (1996), except 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., 2012, 2013). Konduri *et al*. (2016) developed a new version of the PopGen software that is able to account for marginal distributions at different nested geographic levels.

2.2 Spatial Information in Population Synthesis

The research outlined above focuses on person and household demographics and does not consider land use as a fundamental dimension in the population synthesis. In fact, there are very few examples in which spatial elements are included as a part of the population synthesis selection process.

Ballas *et al*. (2005) use an iterative approach that is nearly identical to IPF. Their approach to improving population synthesis was to have "local" households from the sample be more likely to be used to populate the areas they were from. The method they found to be best was to only allow "local" sample households into the population that was used as the synthesis "seed". This is a similar approach to that implemented here; however, we do not employ "local" households, but rather use households that come from areas with similar spatial/land use characteristics in our population synthesis.

The program used throughout this project for population synthesis – PopGen – accounts for spatial information. As it currently works, the program uses small geographic subdivisions that are nested within larger subdivisions. The smaller subdivisions are the ones to which we want to synthesize the population. The larger subdivisions are used for the selection of the pool of eligible households from a survey that can be used for the smaller area synthesis. The larger subdivisions are necessary to avoid sparse matrices of cross-tabulated household and person attributes. However, the process can be problematic. All surveyed households residing in the coarser geographic subdivision are eligible for use in the synthesis of the smaller areas within the larger area. PopGen assumes that people living in that coarser area are all equally likely to live in any of the smaller areas within it. The smaller areas within a coarse geographic subdivision can differ greatly in their land use characteristics (e.g. low commercial density versus high commercial density environments). This means that people living in central cities will also be chosen as candidates for synthesis in more rural environments. This may lead to biases when we transfer behavioral traits to the synthetic population, because there are inherent differences in travel behavior among different residential environments. In this thesis, an attempt is made to rectify this limitation.

3 Methodology

3.1 Synthesis Program: PopGen

The software used for population synthesis here is called PopGen. It is an open-source program originally developed by the SimTRAVEL Research Initiative at Arizona State University ("PopGen: Population Generator," n.d.), but now a project of the Mobility Analytics Research Group (MARG, 2016). PopGen uses an Iterative Proportional Fitting

(IPF) and Iterative Proportional Updating (IPU) based method to perform population synthesis, as described in the literature review. 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 microdata sample (seed) – and a geographic correspondence file. This means we will input household marginal distributions, person marginal distributions, a household microdata sample, a person microdata sample, and a geographic correspondence file. Below, I describe the purpose and construction of these files.

3.1.1 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 2010 ACS, and/or the 2010 US Census.

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).

Block group–level census data was collected and used to form the marginal data in the population synthesis. The block group is a collection of Census blocks that contains between 600 and 3000 people (United States Census Bureau, 2012). They tend to be smaller in cities (where density is higher). There are 23,212 block groups in California. The marginal data is what tells the synthesizer the information about the area whose population you want to

synthesize. It tells the synthesizer the traits we want to use and the counts of people/households with those traits.

Table 3.1 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 3.1 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

3.1.2 Microdata Sample

The microdata sample 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

microdata sample 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). This survey was collected between February 1, 2012 and January 21, 2013. It spanned all 58 counties of California, and included weekdays, weekends, and holidays (NUSTATS, 2013). The 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.

CHTS collected household- and person-level demographic information about respondents. It also included a one-day travel diary from every person, and a long-distance travel log (California Department of Transportation, 2013). The people and households in the CHTS were used to populate the state of California during the synthesis process.

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. I excluded households and individuals that responded "Don't Know" or "Refused to Answer" on any of the questions that were used in this study. If an individual was excluded, so was the rest of their household. The total respondents used here were 36,925 households and 94,901 individuals. Testing revealed that removing these households did not make a significant difference overall, so I proceeded with the reduced set of respondents.

Table 3.2 shows the format for a household microdata sample file. Each row is one household, which is linked to a household ID (hhid). In the person microdata sample 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 3.2 Example of Microdata Sample 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

Figure 3.1 shows the distributions of the variables we used from the CHTS. The same sociodemographic traits are used for all 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.

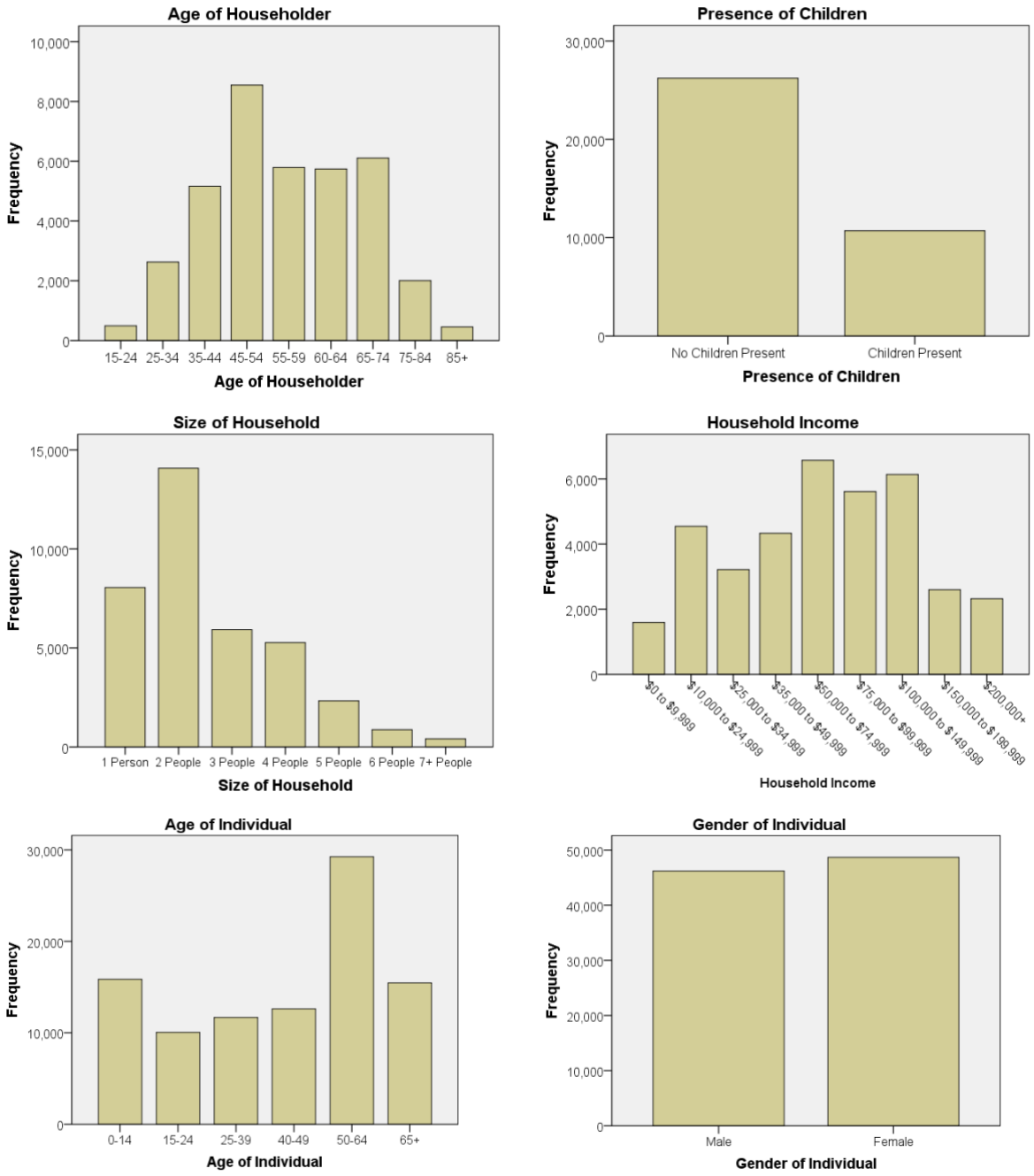


Figure 3.1 Microdata Sample Characteristic Distributions

3.1.3 Output

The PopGen output consists of two datasets: the households and the individuals. As exemplified in Table 3.3 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 3.3 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

3.1.4 IPF and IPU

Iterative proportional fitting (IPF) and iterative proportional updating (IPU) are used in PopGen 1.1 and 2.0, as described in the literature review (Chapter 2.1) and in the introduction to Section 3.1.

3.1.5 PopGen Versions 1.1 and 2.0

PopGen 1.1 can synthesize populations at the following geographic resolutions: county, census tract, census block group, and traffic analysis zone (TAZ). The software also requires a geographic correspondence file. The geographic correspondence file is the file that gives PopGen the list of areas for which it should synthesize populations. Table 3.4 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 3.4 Example of Geographic Correspondence File

county bigint	tract bigint	bg bigint	state bigint	pumano bigint	stateabb text	countyname 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

The most important difference between versions 1.1 and 2.0 is that version 2.0 now allows for multiple spatial resolutions for marginal inputs (Bar-Gera, Konduri, Sana, Ye, & Pendyala, 2008; Konduri et al., 2016; MARG, 2016; Ye et al., 2009). This means that if some variables of interest are at a coarser spatial resolution than others, it is no longer necessary to default to the coarsest scale to include all of them. Some can be at a "fine" scale, and some at a "coarse" scale. The benefit of this is that it allows the inclusion of variables from multiple data sources for the marginal distributions: income from the American Community Survey and all other variables from the U.S. Census. The U.S. Census surveys nearly the entire population, so it is a much more reliable source of data if it is possible to use the information it contains.

3.2 Land Use Classification

As mentioned earlier, incorporation of land use in population synthesis aims to account for the opportunities people have to participate in activities. For this reason, a database that contains the most elementary units of land use is ideal. The business establishment is the

elemental unit in space (a factory, plant, store) where goods are made and/or stored, and services are rendered. NETS is an annual database of business establishments in the United States (Feldman, 2017; Walls & Associates, 2012). It includes extensive information about each business establishment in the United States. The geo-coded firm-level data for this research is extracted from the 2013 NETS database to coincide with the data collection period in the CHTS. It 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*). 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. These business establishments are geolocated in each block group, then indicators of land use such as density (number of employees per square km by each industry type) are created.

The following section describes the populations synthesized using various methods of including land use.

3.3 Population: No Land Use

The synthetic population that did not include land use provides a baseline for comparison to the methods that include land use. This population was generated in the way that most synthetic populations are generated: using only sociodemographic characteristics as the basis. The distributions of the variables came from the 2013 American Community

Survey (ACS) 5-year estimates to smooth any year to year extreme variation in the ACS sample (US Census Bureau, 2016). This is because the ACS provides all the variables we want to use (the Census did not have all of them), and 2013 is the first year the US Census Bureau began making block group level data available.

3.4 Population: Coarse Land Use

The population that was created with Coarse Land Use has the same marginal specifications from the ACS as the "no land use" population. The method of including land use involved creating a land use classification scheme, dividing the areas being synthesized into groups based on the category they fall in, dividing the survey respondents based on the category their household falls in, and running the program separately for each category. This process ensures that every area is only synthesized once, and that households are only used to synthesize areas in the land use category they live in. Further details on the method can be found below.

3.4.1 How Land Use was Included

The method in this section was developed for a California Department of Transportation project (McBride, Davis, Lee, & Goulias, 2016) and published in a paper (McBride, Davis, Lee, & Goulias, 2017). First, we created a kernel density surface of employment density across all of California using the NETS (ESRI, 2016a). We chose employee density because it is a good proxy for how "urban" an area is. Smoothing with kernel density also addresses the error caused by computational artifacts and 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 for a long time and their data were never updated.

We created four categories from this density map by dividing the distribution of densities into quartiles. 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. For clarity, from now on we will call these quartiles Rural (low density), Exurban (medium-low density), Suburban (medium-high density), and Urban (high density).

Next, the state was divided into PUMA's, and the average employee density in each PUMA was used to decide which urban category a PUMA would be labeled as. There are 265 PUMA's in all of California, so this classification is quite coarse. Figure 3.2 shows an example of the difference in area between PUMAs and block groups in the city of Los Angeles. The reason we used PUMAs is because PopGen 1.1 asks for PUMA-level household locations for survey respondents to be used in the creation of the microdata sample matrix that it uses to decide which households to select for a block group.

Finally, the households in the survey were also divided using the PUMA-level classification based on their household location, PopGen was run four times (once for each land use category), and the results were combined to get a synthetic population for the entire state.

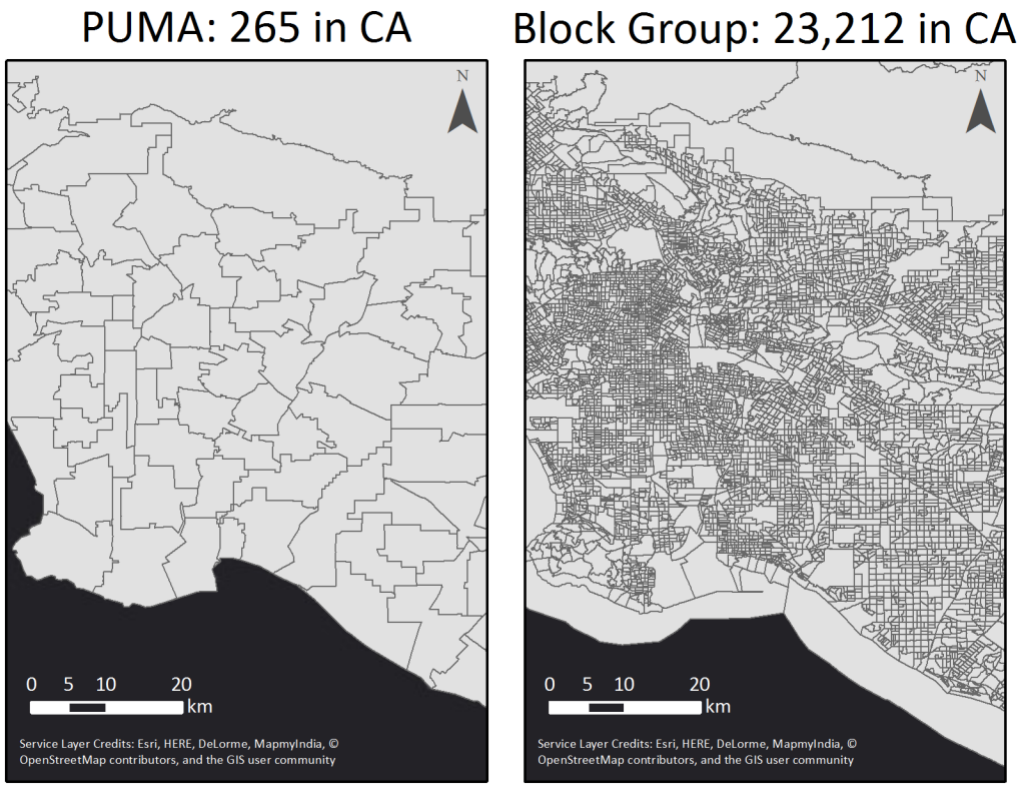


Figure 3.2 PUMA areas versus block group areas (Los Angeles)

This simple classification scheme was initially used because we want to see how coarse it can be while still showing differences in important areas. It also acts as a test of the viability of the method going forward as it gets more complex.

3.5 Population: Finer Land Use

The method for including land use in this third population is nearly the same as for the second population, with one key difference. The land use classification is at the block group level instead of the PUMA level. This was possible because PopGen 2.0 is much more "customizable" than version 1.1. There are 23,212 block groups in California, as opposed to 265 PUMAs. The difference in precision can be visually observed above in Figure 3.2. Aside from the block group level classification, the same method was used: the state was

divided into Rural, Exurban, Suburban, and Urban areas based on the same measure of employee density, and PopGen was run four times.

For this synthetic population's marginal distributions, 2010 Census data was used for householder age, presence of children, number of household members, person age, and person gender. The 2013 American Community Survey 5-year estimates were used for household income because it is not available through the Census. As mentioned earlier, the reason we changed the data source is because PopGen 2.0 allows for multiple data sources, and since Census data is more accurate than ACS estimates (because it surveys the entire population), we used as many variables as possible from the Census.

3.6 Population: LPA Land Use

Data from the 2010 American Community Survey (ACS) was also used for the household income information that was also used as a marginal distribution variable.

3.6.1 Details of LPA

The method used to classify the land use data in this section is latent profile analysis (LPA), also known as latent class cluster analysis (LCCA) (Masyn, 2013). LPA uses a set of variables called *indicators* to determine the ideal number of classes to cluster observations into. LPA provides a way to figure out if there might be a better way to divide the block groups into land use categories with more statistical certainty for their belonging together than the previous method provided. It uncovers the most statistically significant way to divide the data. LPA is usually used to classify people from a survey. It is not commonly used to classify geospatial data. However, it is well-suited to this task due to the flexibility of the analysis method to handle all types of indicators, as long as they are continuous.

Employee densities in 17 individual categories of employment were used as indicator variables. These densities were built in ArcMap 10.2.2 (ESRI, 2013) using business data from the NETS. Businesses from the NETS were added to a block group shapefile as points based on the provided XY coordinates. A kernel smoothing process, using 2 kilometer kernels, was employed to create a density surface for each of the employee density categories. The ArcGIS function "Zonal Statistics as Table" (ESRI, 2016b) was used to calculate the mean employee density per square kilometer within each block group. The employee densities were computed using a 17-category classification of business establishments (based on a modified standard industrial classification) that include agriculture, forestry, fishing, and hunting; mining; utilities; construction; manufacturing; wholesale trade; retail trade; transportation and warehousing; information, finance, insurance, real estate, and rental and leasing; professional, scientific, management, administrative, and waste management services; educational services; health care; arts, entertainment, recreation, accommodation, and food services; other services (except public administration); public administration and armed force; and undefined.

Because of the large number of block groups with very few people in them, there are many block groups with low values for all employee density items, leading to extremely positively skewed data. Moreover, there are a small number of block groups with extremely high employee densities in city centers where the density is much higher than anywhere else. This combination creates extremely high variance for the variables. In order to mitigate this, a log transform was applied to the data. The log transform compressed the values of the 17 observed variables in each block group and eliminates extreme differences in values of the variables we use in LPA.

The LPA was conducted using Mplus 7.4 (Muthén & Muthén, 1998). The order of operations for performing an LPA are as follows: A one-class model is fit, followed by a two-class, *et cetera* until a model is run that is not well-identified (Asparouhov & Muthén, 2012; Masyn, 2013; Nylund, Asparouhov, & Muthen, 2007). With every run, a set of fit statistics are recorded. These are presented in Table 3.5. The fit statistics are used to determine whether or not the model is well-identified. It is recommended that once a model runs and has a non-significant *p*-value for either the Bootstrapped Likelihood Ratio Test (BLRT) or the Vuong-Lo-Mendell-Rubin Adjusted Likelihood Ratio Test (VLMRT), the model with one fewer classes is chosen, as long as the other fit statistics show that the model fits well (Asparouhov & Muthén, 2012; Masyn, 2013; Nylund et al., 2007). This is because a non-significant *p*-value for one of these statistics indicates that there is no longer a statistically significant improvement in model fit by adding further classes. As Table 3.5 shows, The VLMRT reached a non-significant *p*-value of 0.421 with the 5-class model. Based on fit criteria, class sizes, and interpretability, the 4-class model was chosen. An entropy value approaching one indicates clear delineation of the classes. So, the entropy value of 0.95 for the 4-class model means the indicators discriminate well between the classes (Celeux & Soromenho, 1996). Based on properties described in further detail below, the four classes will be referred to from now on by the names Rural, Exurban, Suburban, and Urban.

Table 3.5 LPA Fit Indices

Number of classes	Log likelihood	BIC	ABIC	*p*-value of BLRT	*p*-value of LMRT	Entropy	BF	cmP
1	-856163.9	1712669.6	1712395.8	-	-	-	.00	.00
2	-740942.3	1482407.4	1481988.7	< .001	< .001	.99	.00	.00
3	-688987.4	1378678.5	1378114.9	< .001	< .001	.96	.00	.00
4	-658690.7	1318266.0	1317557.4	< .001	.001	.95	.00	.00
5	-637558.4	1276182.2	1275328.7	< .001	.421	.95	.00	1.00

Note. BIC = Bayesian Information Criterion; ABIC = Sample-size Adjusted BIC; BLRT = Bootstrapped Likelihood Ratio Test; VLMRT = Vuong-Lo-Mendell-Rubin Adjusted Likelihood Ratio Test; BF = Bayes Factor; cmP = correct model probability.

The map in Figure 3.3 shows the geographic distribution of the block groups in California. The maps of the older methods are included for comparison. All three land use classifications manage to capture the main urban centers (San Francisco, Sacramento, Los Angeles, and San Diego) relatively well. It should be noted that location is not considered by the LPA, so the grouping is solely based on the most statistically significant way to group the data based on the indicator variables used. Despite this, a clear, identifiable spatial pattern in the grouping is present in the resultant map. It shows the great benefit and improvement of using LPA Land Use classification. Figure 3.3C shows that the group called Exurban does not extend as far beyond the city centers as it does in the Finer Land Use classification (Figure 3.3B). The rural area starts much closer to the urban centers. Based on empirical knowledge of the California population distribution, the LPA classification is a much more accurate depiction of the State's urban-rural landscape. The Coarse Land Use group is the oldest classification, which faced limitations imposed by the much coarser classification scheme (McBride, Davis, Lee, & Goulias, 2017).

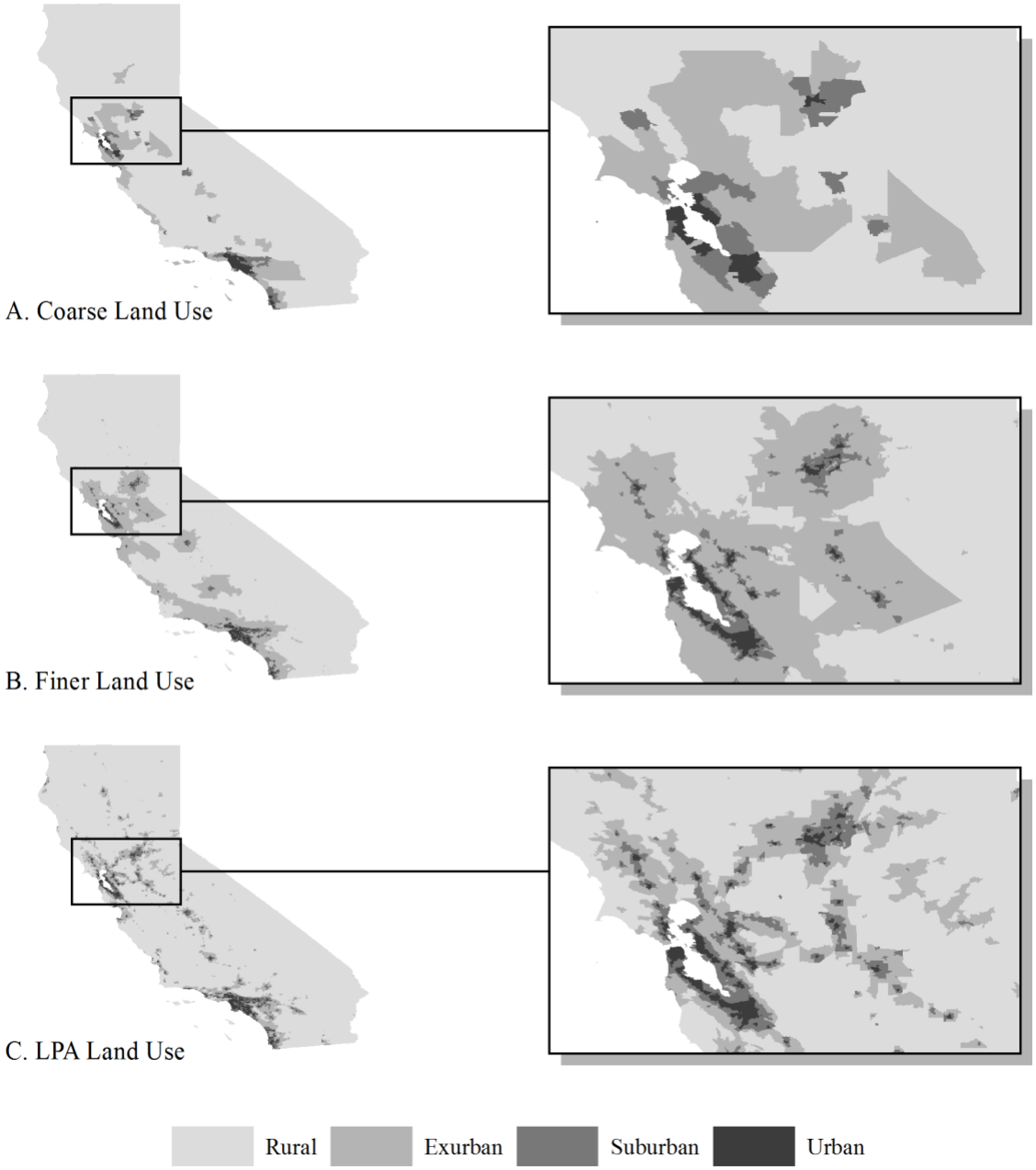


Figure 3.3 Comparison of land use classification schemes
Number of block groups in each category: (A) Rural: 3222, Exurban: 4743, Suburban: 6965, Urban: 8268. (B) Rural: 2565, Exurban: 4172, Suburban: 7656, Urban: 8803. (C) Rural: 1076, Exurban: 2582, Suburban: 10670, Urban: 8868.

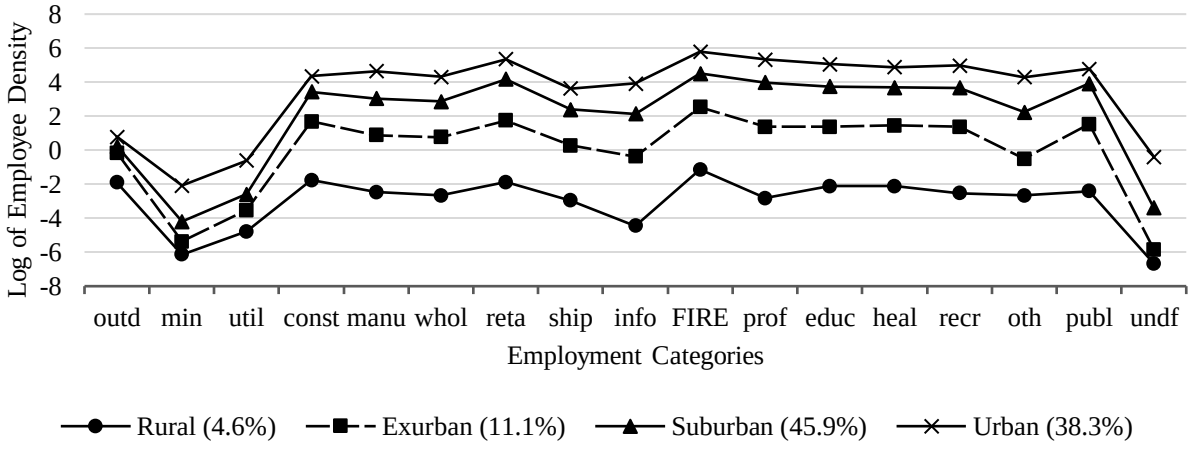
The initial idea behind the LPA method was that by dividing up the employee density into separate categories based on types of businesses, block groups could be categorized

based on the way that business establishments are located in the geographic space, and not based on proximity to an urban center. For example, the central valley of California – where most of the state's agriculture occurs – would be clustered together because of more prominent/higher number of people employed in agriculture there, along with whichever other patterns of employment are present in those areas. Regions that have some other types of employment that are most prominent, like *FIRE* (finance, insurance, real estate, and rental/leasing) in city centers, would be clustered together based on that.

Figure 3.4A is the item probability plot. It shows the mean value for each indicator variable in the four classes. The LPA did not result in major differences across the block groups for each of the types of employment within each cluster. As is made clear by both Figure 3.3 and Table 3.5, the most important factor in determining the land use classification scheme seems to still be proximity to an urban center in the LPA classification. Although if one compares the map in Figure 3.3B to Figure 3.3C, it does appear to be better overall at picking out areas that would not be considered near urban centers.

The lines run parallel to each other for all the classes. This means that the determining factor for the grouping was more about overall land use than it was about the unique regional makeup of employment types in California.

(A) Item Probability Plot



(B) Index of Variables

Item Name	Item Description
Outd	Agriculture, forestry, fishing, and hunting
Min	Mining
Util	Utilities
Const	Construction
Manu	Manufacturing
Whol	Wholesale trade
Reta	Retail trade
Ship	Transportation and warehousing
Info	Information
FIRE	Finance, insurance, real estate, and rental and leasing
Prof	Professional, scientific, management, administrative, and waste management services
Educ	Educational services
Heal	Health care
Recr	Arts, entertainment, recreation, accommodation, and food services
Oth	Other services (except public administration)
Publ	Public administration and armed force
Undf	Undefined

Note. This index of variable name descriptions provides detail about what is included in each employment category.

Figure 3.4 Items Used in LPA and their Values for each Class

Part of the reason the classification did not pick up on patterns of employment beyond the proximity to urban-ness has to do with limitations in the employment data used. The coordinates provided by NETS for the businesses are attached to business fronts (i.e. offices or storefronts). Although something like an agricultural business might have employees in the central valley, if their business front is located in a city center, those employees are

going to be "placed" there instead of in the block group that their farm is located in. The indicator variables used where this could be especially problematic are mainly the *outdoor*, *mining*, and *shipping* variables. These are industries where many of the employees of the industry would not work at the office building indicated as the "storefront" in the NETS database. Another part of this is that all the types of employment available in the NETS were used in the LPA. By including businesses that are ubiquitous (like *retail*, *education*, and *healthcare*), all groups contain substantial numbers of employees in these industries but at lower densities as we move from the center city to the outskirts of the city. In this way, the vast difference in employee density between the highest and lowest densities is the main reason for the way the categorization ended up. The highest-density areas are so much higher than the lowest-density areas that it "washes out" any sort of smaller detailed land use details that might group categories together.

Despite these issues, the LPA Land Use classification is still substantially better than previous methods at classifying the block groups of California based on land use. The previous method (Section 3.5) used an overall employee density instead of dividing them up by their type, and took proximity to an urban center into account in its classification. It also required that the number of people in the microdata sample for each group be relatively even so that the synthesizer would have a good amount of people to draw from (i.e., in population synthesis we used census univariate distributions as control totals for each block and drew households and individuals from a survey). This means that the number of block groups in each category was much more similar than what was found with LPA. LPA does not take any of this into consideration. All it considers is the values of densities given to it, and whether/how those values cluster together in a statistically significant way. The number of

block groups that went into each category was quite uneven. The number of block groups in the Suburban category was the highest (10,670). Meanwhile, in the Rural category there were only 1,076 block groups.

Rural areas are oversampled in the CHTS, so the issue of having enough observations in the microdata sample to synthesize an area properly did not become an issue, and probably would not become an issue as long as the sample is not disproportionately small compared to the number of block groups and number of people in those block groups that it is trying to synthesize.

In addition, because the groups for the LPA classification scheme ended up being based on overall employment density, the names given to the older groups can still be applied here (Urban, Suburban, Exurban, and Rural).

4 Findings

In this section, I first demonstrate the importance of including land use indicators around the residence of households to explain travel behavior. I then illustrate the findings of incorporating land use in population synthesis and transfer of behavioral data statewide.

4.1 Tobit Models

Tobit regression models are designed for limited dependent variables (Rees & Maddala, 1985). In this case, limited means a dependent variable that has a large amount of data at one value (e.g., at zero). This violates the assumption of a symmetric distribution of the linear regression random error term; furthermore, the presence of many observations at the zero

value may indicate the presence of two segments that are qualitatively different in the population (i.e., one that usually has no travel and another that just happened not to travel on the survey day). The Tobit model is a non-linear regression model that accounts for the "piling up" of observations at the zero value of the dependent variable. For this reason, the derivative of the expectation of the dependent variable with respect to an independent variable (called marginal effect herein) is not the regression coefficient as in linear regression. This can be computed using established techniques (Greene, 2003), and the estimation tables below show these derivatives. We ran these models on the CHTS for the traits of interest. We tested each population for the traits of interest in the CHTS. Number of trips and miles traveled were chosen because they are a good representation of travel behavior. All models use the same independent variables. The primary difference is the land use classification method.

4.1.1 Number of Trips per Person

Table 4.1 Tobit Model for Number of Trips (Coarse Land Use Classification)

Independent Variables		Marginal Effects	Standard Error	b/St.Er.	Prob. |z|>Z
Age of Householder (85 and older is the excluded category)	Age 15-24	1.0	0.2	5.477	0.000
	Age 25-34	1.2	0.2	7.751	0.000
	Age 35-44	1.3	0.2	8.721	0.000
	Age 45-54	1.3	0.1	9.599	0.000
	Age 55-59	1.3	0.1	9.481	0.000
	Age 60-64	1.2	0.1	9.310	0.000
	Age 65-74	1.1	0.1	8.143	0.000
	Age 75-84	0.6	0.1	4.422	0.000
Children Present in Household	Children Present	0.0	0.1	-0.759	0.448
Household Income (income between $0-$9,999 is the excluded category)	$10,000-$24,999	0.0	0.1	0.389	0.697
	$25,000-$34,999	0.0	0.1	0.320	0.749
	$35,000-$49,999	0.2	0.1	2.849	0.004
	$50,000-$74,999	0.4	0.1	5.530	0.000
	$75,000-$99,999	0.5	0.1	6.978	0.000

	$100,000-$149,999	0.6	0.1	8.751	0.000
	$150,000-$199,999	0.6	0.1	7.977	0.000
	$200,000+	0.7	0.1	8.055	0.000
Number of Females in the Household		-0.2	0.0	-8.976	0.000
Mean Age of Household		0.0	0.0	-6.766	0.000
Measurement of Land Use Around The Residence (Urban is the excluded category)	Suburban	-0.4	0.0	-11.652	0.000
	Exurban	-0.6	0.0	-15.428	0.000
	Rural	-1.0	0.0	-25.105	0.000

Table 4.1 shows the results for the Coarse Land Use model run for the number of trips. The age of the householder seems to have a bell curve-like shape to its coefficient results. The number of trips increases with age until the category for ages 35-44. Then, it decreases with age. The excluded category is age 85+. For example, the coefficient means that households with householders from age 15-24 travel on average about 1 trip more than those in the 85+ category. All categories were significantly different than zero, which means that the model suggests that the age of householder is an important determinant of the number of trips in all cases. Those with children present in their household make fewer trips than those without children according to the coefficient results, but this variable is not significantly different than zero. This means that the presence of children does not have a significant impact on the model results for the number of trips a household makes. We use this form of the variable because this is the format of the data available data from the US Census and the findings in the next (VMT) model. The next variable tested is household income. The model shows that as income increases, the number of trips increase, although it does not become a significant influence until households are making over $35,000 a year. For the number of females in the household, as the number increases, the number of trips decreases. This is a pattern that has been observed in past travel behavior research. The dichotomy of gender

roles leads women to spend more time in the private sphere, and less in the public one – and vice versa for men (M. Kwan, 2000; M. P. Kwan, 1999; Turner & Niemeier, 1997). This would lead to fewer trips in a household with more women. As households get older, they also make fewer trips. The final set of variables – land use – is the most important to this specific study, as they are the variables we hope will improve the travel behavior information retrieved from a synthetic population. The methods of creation behind the land use categories is included above in the section titled "Land Use Data". As expected, the more rural a household is, the fewer trips it tends to make, and vice versa for an urban household. The variables are all significant, showing that the land use categories we created significantly influence trip-making.

Table 4.2 Tobit Model for Number of Trips (LPA Land Use Classification)

Independent Variables		Marginal Effects	Standard Error	b/St.Er	Prob. |z|>Z
Age of Householder (85 and older is the excluded category)	Age 15-24	1.0	0.2	5.564	0.000
	Age 25-34	1.2	0.2	7.782	0.000
	Age 35-44	1.3	0.1	8.796	0.000
	Age 45-54	1.4	0.1	9.729	0.000
	Age 55-59	1.3	0.1	9.619	0.000
	Age 60-64	1.3	0.1	9.470	0.000
	Age 65-74	1.1	0.1	8.315	0.000
	Age 75-84	0.6	0.1	4.575	0.000
Children Present in Household	Children Present	0.0	0.1	-0.794	0.427
Household Income (income between $0-$9,999 is the excluded category)	$10,000-$24,999	0.0	0.1	0.221	0.825
	$25,000-$34,999	0.0	0.1	0.234	0.815
	$35,000-$49,999	0.2	0.1	2.734	0.006
	$50,000-$74,999	0.4	0.1	5.458	0.000
	$75,000-$99,999	0.5	0.1	6.938	0.000
	$100,000-$149,999	0.6	0.1	8.800	0.000
	$150,000-$199,999	0.7	0.1	8.271	0.000
	$200,000+	0.7	0.1	8.399	0.000
Number of Females in the Household		-0.2	0.0	-8.683	0.000
Mean Age of Household		0.0	0.0	-6.752	0.000
Measurement of Land Use Around the Residence (Urban is the excluded category)	Suburban	-0.5	0.0	-15.942	0.000
	Exurban	-0.7	0.0	-18.993	0.000
	Rural	-1.1	0.1	-21.632	0.000

For the LPA classification scheme Tobit model (Table 4.2), everything except the presence of children and the low-income levels were significant. Households with middle-aged householders make a higher number of trips than the younger or older householders. These households are in a lifecycle stage that requires more traveling for family members and working.

The two lowest income categories were not significant, meaning they were not significantly different than the $0-$9,999 category. Until households get to $35,000 or

above, there is not a significant difference in the number of trips they make in a day as compared to the lowest income category. They function similarly to the lowest income category in terms of the number of trips they make in a day. The pattern of lower income meaning fewer trips parallels the VMT model. With less money at hand, there is less flexibility to participate in activities outside of home that probably require money, like day trips, vacations, or recreational activities. A long-distance trip would not necessarily show up very strongly in number of trips measurements, since a 1-mile trip is worth the same as a 50+ mile trip. This might be why the pattern shows up more strongly in the VMT model as compared to the number of trips model.

More females in a household corresponds with fewer trips per day. Every additional female in a household corresponds with a decrease of 0.157 in the number of trips per person in a household. Past travel behavior research has shown this pattern. Gender role dichotomy leads to women spending more time in the private sphere, and men spending more time in the public sphere (M.-P. Kwan, n.d.; Turner & Niemeier, 1997). Every increase of 1 year to the mean age of household leads to a decrease in the number of trips of 0.013 as the VMT showed too.

Suburban households make 0.490 fewer trips per person than urban households. Exurban households make 0.735 fewer trips per person than Urban households. Rural households make 1.089 fewer trips per person than Urban households. This is a consistent trait of rural households that have a lower number of trips of longer distances.

When compared to the Coarse Land Use model (Table 4.1), the most defined difference is in the Exurban category. The Coarse Land Use model had -0.563 trips as compared to the Urban households, and the new model has -0.735. In the Table 4.1 model, the Rural group

had -0.952 trips as compared to the Urban group, while in this one they have -1.089. Suburban is still similar in both.

4.1.2 Vehicle Miles Traveled (VMT) per Person

Table 4.3 Tobit Model for VMT (Coarse Land Use Classification)

Independent Variables		Marginal Effects	Standard Error	b/St.Er.	Prob. |z|>Z
Age of Householder (Age 85+ is the excluded category)	Age 15-24	14.28	2.42	5.892	0.000
	Age 25-34	11.61	2.08	5.582	0.000
	Age 35-44	10.90	1.97	5.545	0.000
	Age 45-54	12.74	1.85	6.893	0.000
	Age 55-59	12.81	1.80	7.110	0.000
	Age 60-64	12.23	1.76	6.952	0.000
	Age 65-74	10.29	1.72	5.984	0.000
	Age 75-84	6.21	1.79	3.466	0.001
Children Present in Household	Children Present	-2.27	0.64	-3.531	0.000
Household Income (the excluded category is $0-$9,999)	$10,000-$24,999	5.98	1.00	5.979	0.000
	$25,000-$34,999	11.33	1.04	10.864	0.000
	$35,000-$49,999	13.64	1.00	13.664	0.000
	$50,000-$74,999	18.25	0.96	19.105	0.000
	$75,000-$99,999	19.86	0.97	20.501	0.000
	$100,000-$149,999	21.98	0.96	22.874	0.000
	$150,000-$199,999	21.63	1.07	20.173	0.000
	$200,000+	22.72	1.10	20.758	0.000
Number of Females in the Household		-0.148	-0.15	0.24	-0.629
Mean Age of Household		-0.11	0.03	-4.373	0.000
Measurement of Land Use Around The Residence (Urban is the excluded category)	Suburban	4.33	0.45	9.608	0.000
	Exurban	7.44	0.48	15.667	0.000
	Rural	8.92	0.49	18.075	0.000

Table 4.3 shows the results for the Tobit model for the Vehicle Miles of Travel (VMT). Like the number of trips, the age of householder generally follows a bell curve-shaped trend with a few exceptions. The group traveling by far the furthest is the Age 15-24. Because of the low number of under-18 householders in the microdata sample, the very high travel

numbers most likely come from the unique way in which young adults (18+) travel. Based on the fact that the number of trips they make is not much higher than any other group compared to their miles of travel, it seems that this group travels farther, but does not necessarily make more trips. The presence of children reduces the miles traveled. Unlike the number of trips model, in this case the presence of children is significantly different than zero. Households that have children travel 2.3 miles less on average than those without children. For the income variables, just like the number of trips, VMT increases with income. In this case, it is also significant at the lower income levels. The number of females in the household is not a significant determinant of the vehicle miles traveled. This is interesting because the number of trips was significant. This means that despite women making fewer trips, they travel a similar distance to men. An increase in the mean age of a household leads to a decrease in miles traveled, so as households get older, they travel fewer miles. As we expected, the land use variables are significantly different than zero. Households residing in Rural environments travel on average approximately 8.9 miles per day by vehicles more than households in Urban environments. Households in Exurban environments travel approximately 1.5 miles less than the Rural residents, whereas households in suburbs travel in vehicles for less than half the miles of Rural households. This is perfectly consistent with the spatial structure of these four different environments we identified here. Also, all these results point out that we should separate the CHTS microdata sample into four distinct groups based on these land use variables. CHTS households that live in Urban environments are used as microdata sample for Urban synthetic population, CHTS household living in suburbs are used for Suburban synthetic population, CHTS households that live in exurbs are used for Exurban synthetic population and CHTS

households that live in Rural environments are used to reproduce the Rural population. This enhances our ability to transfer household behavior statewide because it also accounts for residential self-selection.

Table 4.4 Tobit Model for VMT (LPA Land Use Classification)

Independent Variables		Marginal Effects	Standard Error	b/St.Er.	Prob. |z|>Z
Age of Householder (85 and older is the excluded category)	Age 15-24	14.00	2.42	5.777	0.000
	Age 25-34	11.42	2.08	5.491	0.000
	Age 35-44	10.62	1.97	5.406	0.000
	Age 45-54	12.46	1.85	6.737	0.000
	Age 55-59	12.56	1.80	6.969	0.000
	Age 60-64	11.93	1.76	6.782	0.000
	Age 65-74	10.03	1.72	5.830	0.000
	Age 75-84	5.98	1.79	3.336	0.001
Children Present in Household	Children Present	-2.21	0.64	-3.443	0.001
Household Income (income between $0-$9,999 is the excluded category)	$10,000-$24,999	6.04	1.00	6.039	0.000
	$25,000-$34,999	11.34	1.04	10.865	0.000
	$35,000-$49,999	13.67	1.00	13.696	0.000
	$50,000-$74,999	18.28	0.96	19.133	0.000
	$75,000-$99,999	19.91	0.97	20.553	0.000
	$100,000-$149,999	22.01	0.96	22.906	0.000
	$150,000-$199,999	21.50	1.07	20.065	0.000
	$200,000+	22.57	1.09	20.643	0.000
Number of Females in the Household		-0.16	0.24	-0.697	0.486
Mean Age of Household		-0.11	0.03	-4.380	0.000
Measurement of Land Use Around the Residence (Urban is the excluded category)	Suburban	4.29	0.40	10.686	0.000
	Exurban	7.91	0.50	15.761	0.000
	Rural	10.16	0.65	15.604	0.000

For the LPA classification model of Vehicle Miles Traveled, everything was significant except for the number of females in the household (Table 4.4). This means that the miles traveled by a person are affected by the age of the householder, the presence of children in a household, the household income level, the mean age of the household, and the LPA Land

Use classification. Categorical variables: age of householder, presence of children, household income, and land use categories. Continuous variables number of females in the household (count), mean age of household.

The younger the householder is, the more miles traveled per person in a household. The reference group is the 85+ year old householder group. A household whose householder is between 15 and 24 years old will travel about 14 miles per person more than one whose householder is 85+ years old. If the householder is between 25 and 34, the household will travel 11.4 miles more than the 85+ year old householder group. This pattern of decreasing miles traveled with increasing householder age continues. These results match with our expectations about how age should affect the amount of traveling done.

If children are present in the household, the number of miles traveled per person goes down. Households with children present travel 2.2 miles less than those without children. This is likely because the presence of children in a household tends to necessitate family members staying closer to home, because of limitations of having children such as dropping them off and picking them up from school, appointments, etc.. Although having all of these travel obligations might increase the amount of traveling to an extent, it is probably being offset by the limitations of needing to stay closer to home to deal with childcare duties.

Higher household income corresponds with higher miles traveled per person. The coefficients for the income categories are higher than any other group of dependent variables. They are one of the strongest indicators of higher VMT. The wealthier respondents may be making longer-distance trips because they have the flexibility to do so. As the mean age of household goes up, the miles traveled goes down. It has been shown that older people tend to travel less than younger people.

Suburban households travel 4.287 miles more per person than Urban households. Exurban households travel 7.909 miles more per person, and Rural households travel 10.164 miles more per person than Urban households do. Urban areas are the highest density, so distances between locations of interest tend to be shorter. This all means that people in urban areas will have much less of a distance to travel than people in Rural areas.

When compared to the Coarse Land Use classification Tobit model (Table 4.3), Suburban and Exurban categories had nearly the same values for the difference in miles traveled per person as compared to Urban. However, for the Rural group, the old method Rural households traveled 8.918 miles more, while in this one they travel 10.164 miles more. The Finer classification exacerbates the differences in VMT between the two land use classification methods.

4.2 Comparisons

After synthesis, travel behavior characteristics were transferred to the synthetic populations. These were used to study differences in travel behavior for the different populations. The purpose of this was to see whether the use of classification and various classification schemes had an effect on the travel behavior traits.

First, inclusion of land use information even when the Coarse Land Use classification is used makes a big difference in transferring and expanding data from the microdata sample to the block group synthetic population. Figure 4.1, Figure 4.2, and Figure 4.3 (reproduced from McBride at al., 2017) contain 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. The percent of trips per day walking is the best example of how land use can improve

our ability to better represent behavior in such a large scale. The figures show that walking trips are most often taken by residents of large urban environments and particularly (see the cutout for San Francisco that shows the parts of the city that are conducive to walking and inhabited by persons that are most likely to walk). These patterns also correspond to the relationship between “urban-ness" and travel behavior that we hope to see and the Tobit models above indicate we should be expecting 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 even coarsely defined land use will make these models much more valuable and reliable for modeling travel behavior.

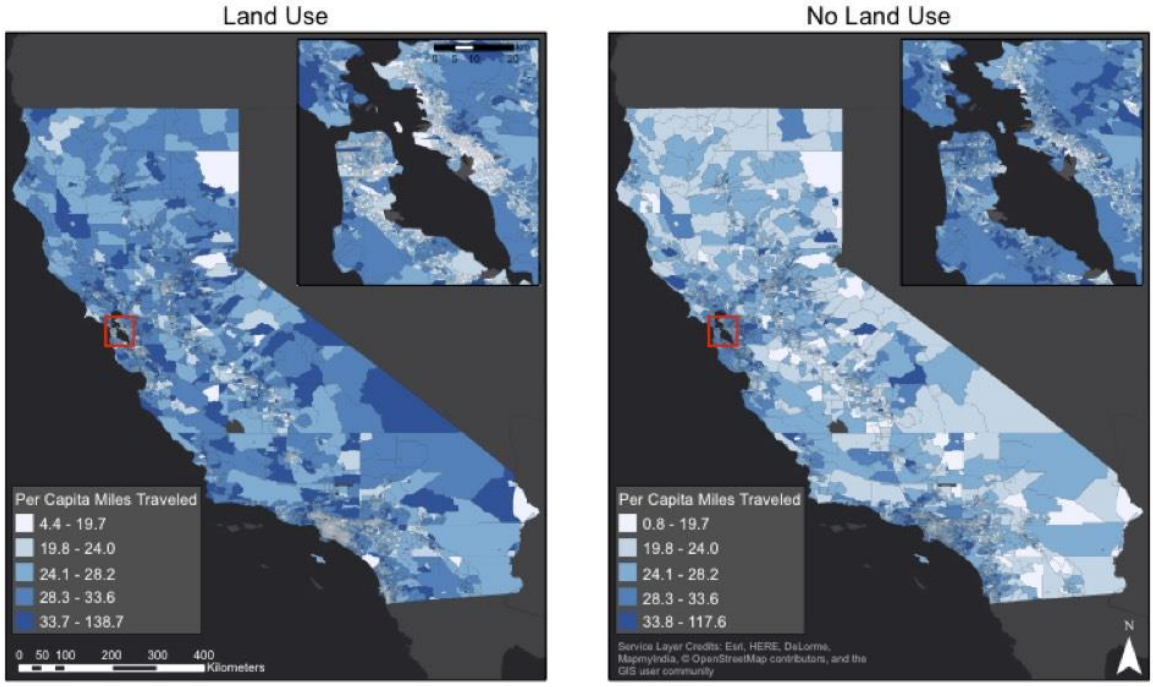


Figure 4.1 Daily miles of travel per person
(reproduced from McBride et al., 2017)

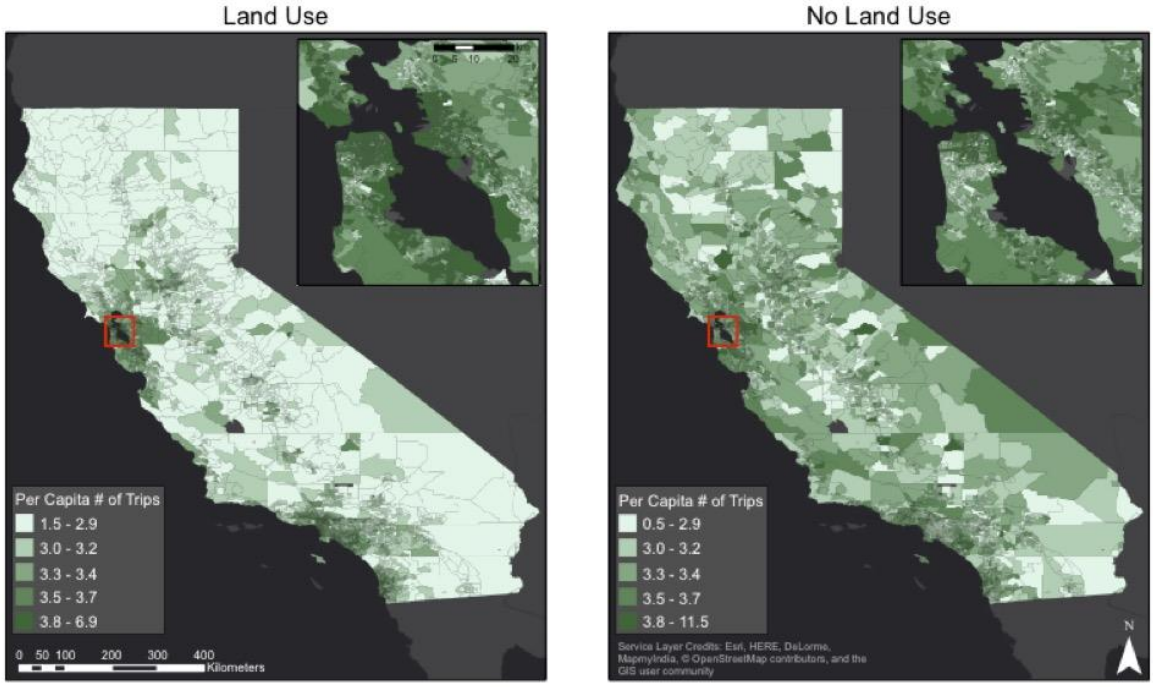


Figure 4.2 Number of trips per day per person
(reproduced from McBride et al., 2017)

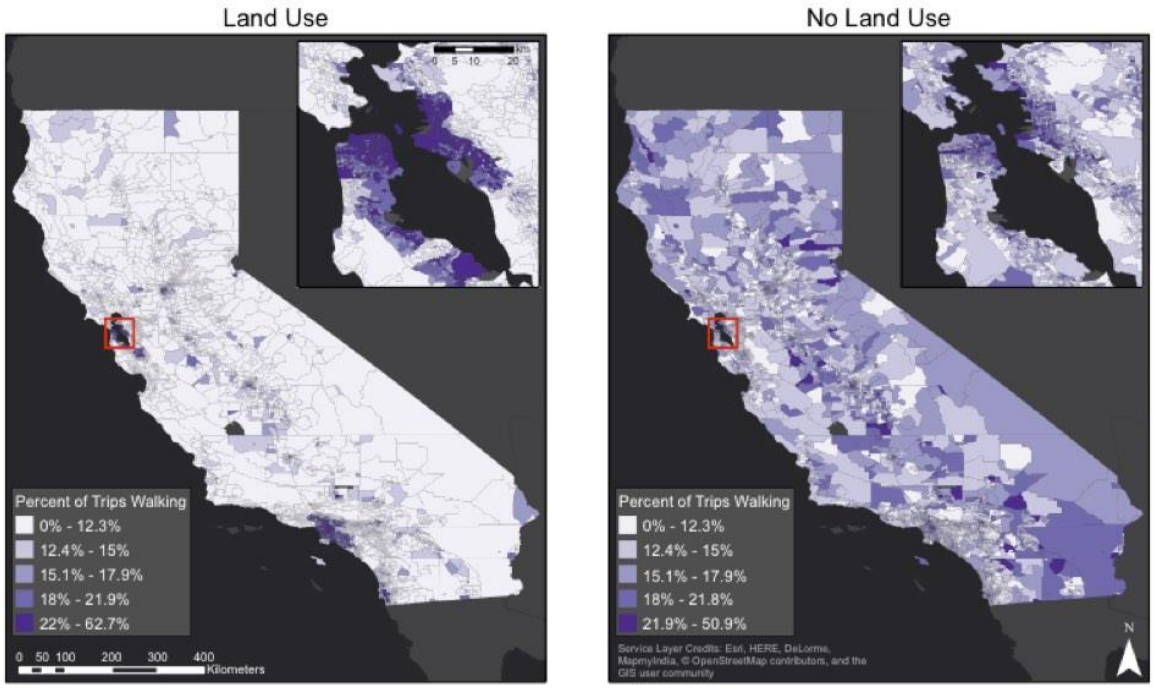


Figure 4.3 Percent of daily trips walking
(reproduced from McBride at al., 2017)

The next analysis of the maps will primarily focus on comparing the Finer Land Use to the LPA Land Use, since these are the two methods that are most similar and that the LPA method is hopefully improving upon.

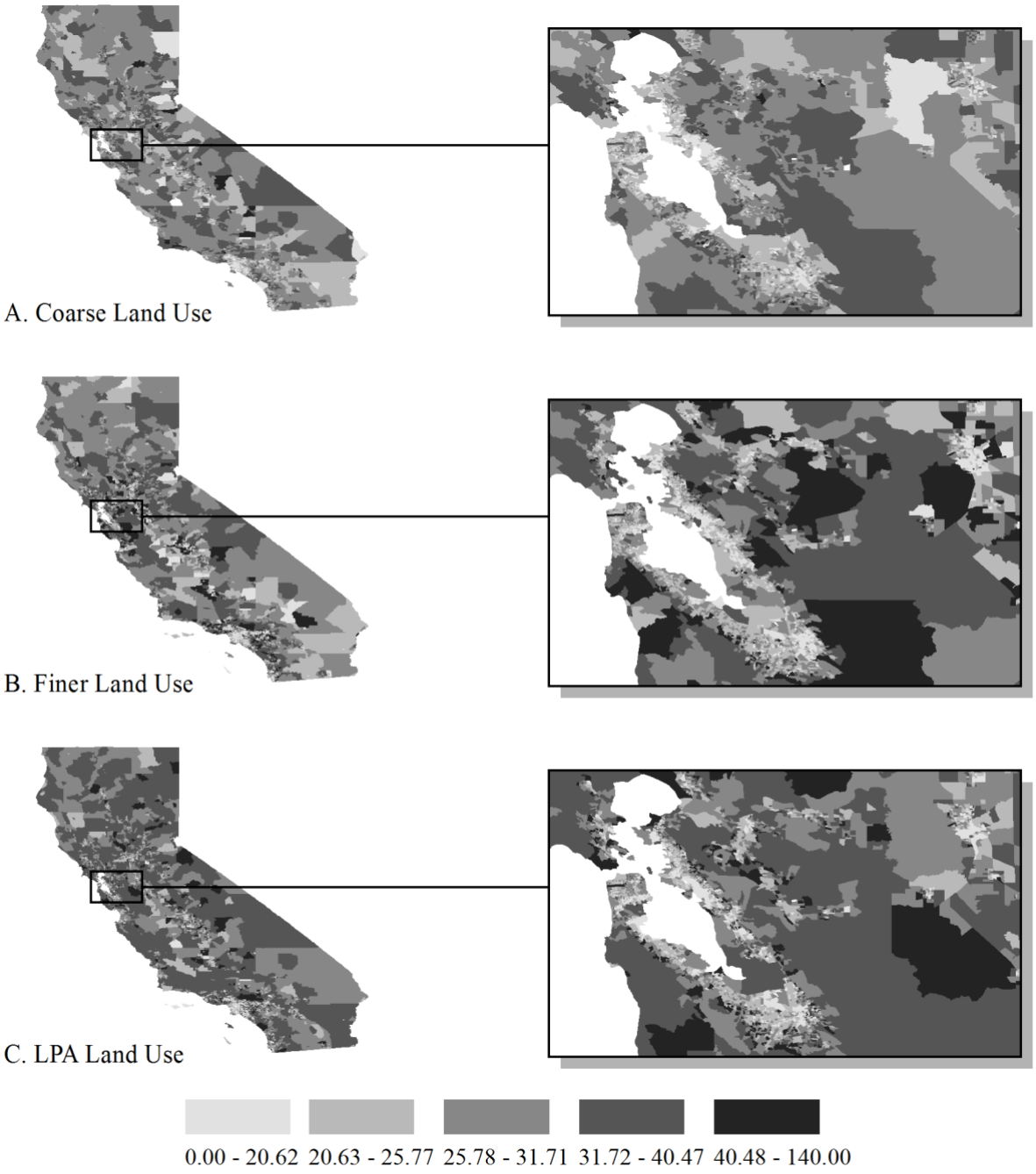


Figure 4.4 Miles Traveled per Person.

The miles traveled per person in the *LPA* population it is more strongly related to the urban-ness of a block group than in any of the other populations. As Figure 4.4 shows, San Francisco and the surrounding area are very identifiable in all three maps. However, the

block groups right outside of the city centers are where the biggest differences lie. The miles traveled seem to continually increase further away from the city centers. This pattern is present in the Coarse Land Use and Finer Land Use populations too, but not as clearly.

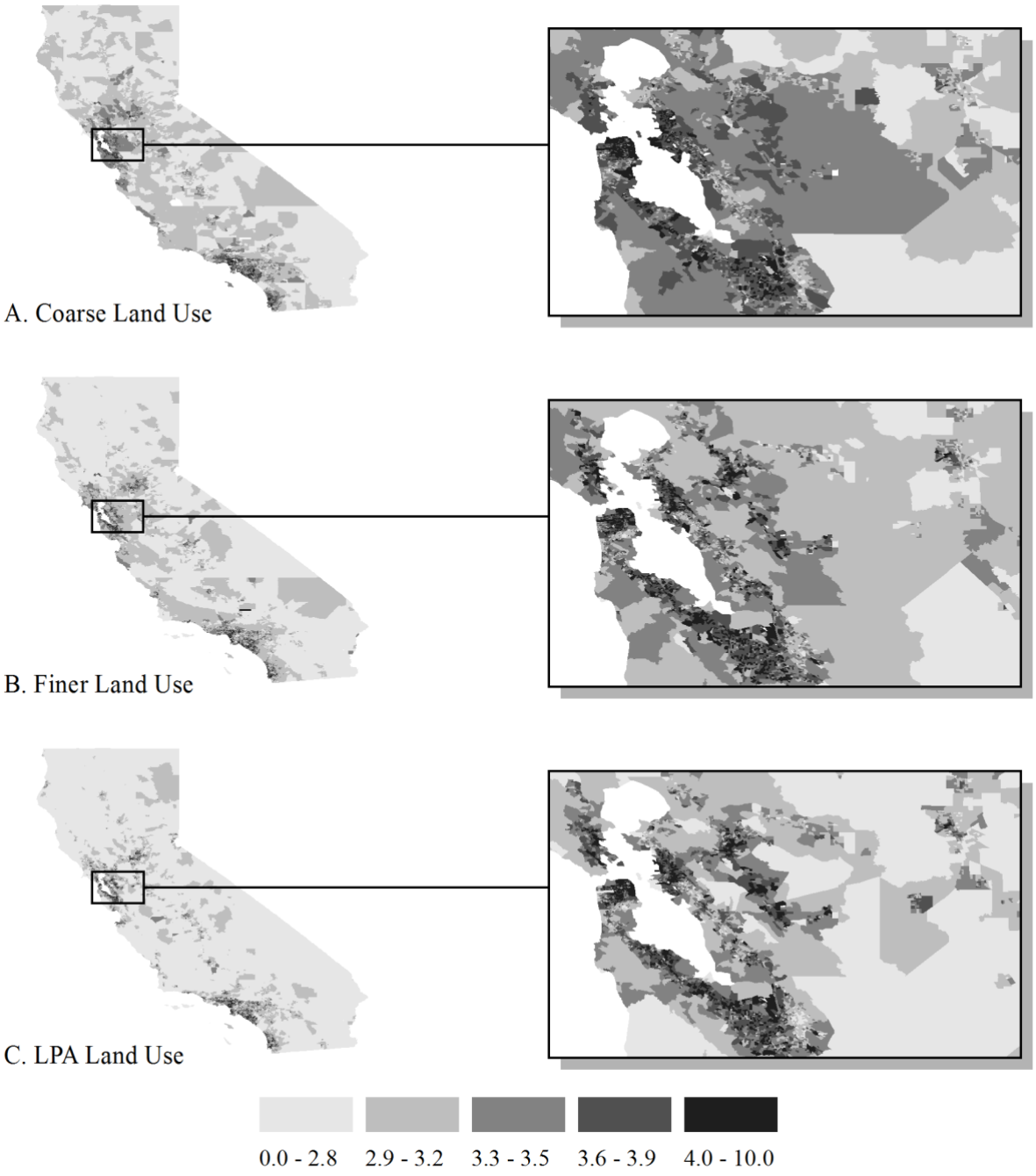


Figure 4.5 Number of Trips per Person.

It is expected that urban populations make more trips than rural populations as the Tobit model shows. The map's borders of the urban areas are much more defined in the LPA Land Use (Figure 4.5C) than they are in the Finer Land Use. The higher values for the number of trips are more restricted to urban areas in the LPA Land Use. Although the same pattern is present in the Coarse and Finer populations, it is not as distinct, and the borders of the urban areas are less clear.

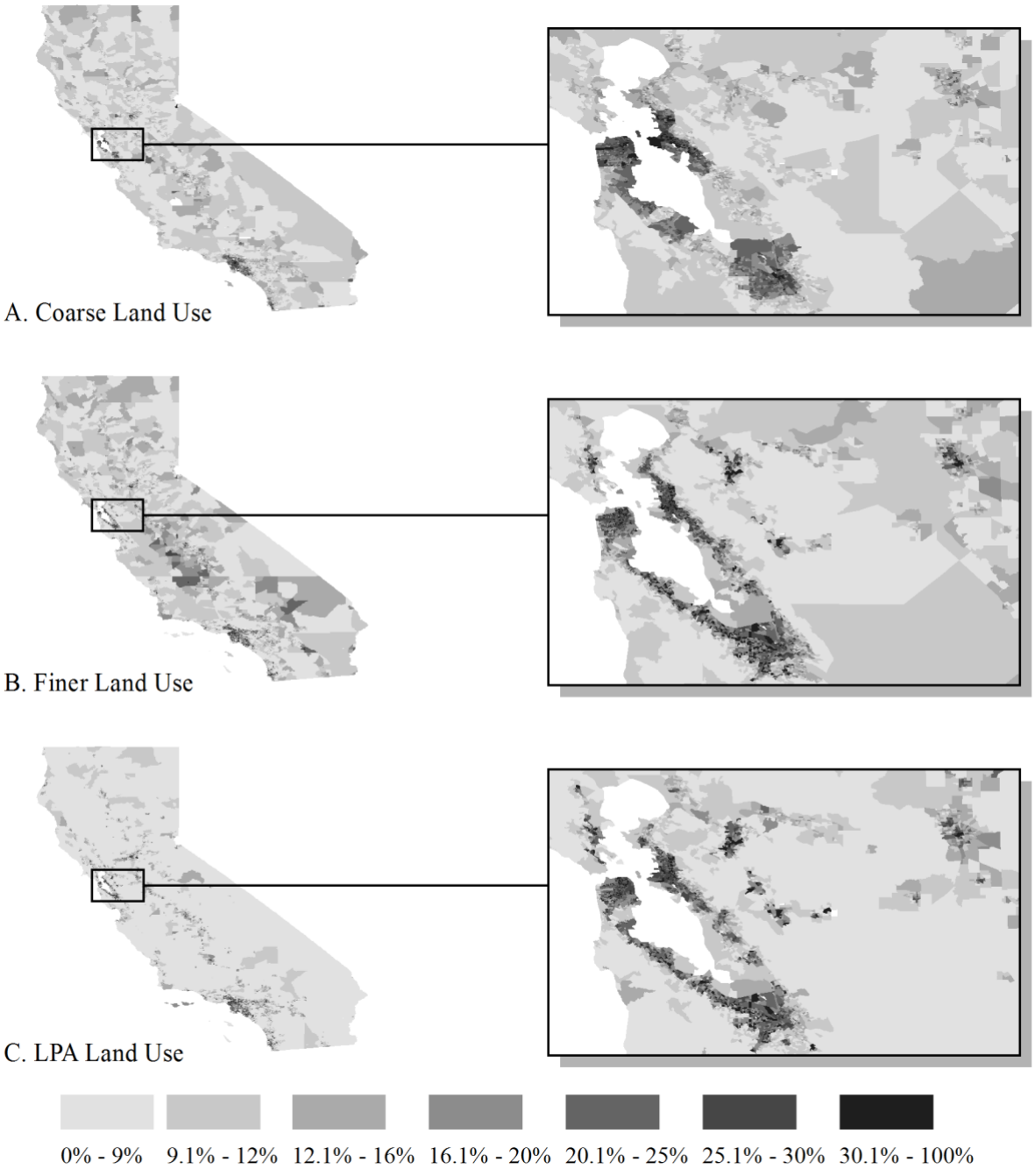


Figure 4.6 Percent of Trips on Foot

The percent of trips on foot (Figure 4.6) is a very clear way to define the urban areas. Walking for a large number of trips is only possible in "walkable areas" that are characteristic of urban areas, where the distances between points of interest are close enough

together that it is possible to walk between them. It is a good test of whether the land use classification is helping the population synthesis to populate areas with households that have travel traits that are more characteristic of the areas they are populating. The maps show that the higher percentages of walking trips are limited to the urban areas, and that that specific pattern is the most clearly shown in the LPA Land Use population (Figure 4.6C). The pattern is present in the Coarse Land Use and Finer Land Use populations (Figure 4.6A and B respectively), but it is not limited to urban areas as much. There is still some "random noise" of large, rural block groups with high percentages of walking trips, even in the Finer Land Use population that most likely should not have those numbers.

5 Conclusions

Excluding land use as a source of spatial information when performing population synthesis is neglecting an important source of information about the population, and it biases transferability of travel behavior traits. For example, when predicting the percentage of trips on foot, the highest percentages should be concentrated in the cities. These traits are more successfully transferred to the populations that included land use than the one that did not.

For the first time, land use has been explicitly and systematically used in population synthesis in this thesis and related papers. Moreover, land use has been included using a variety of classification methods that were then compared to each other, providing valuable information about what methods are best and have the best results when it comes to predicting travel behavior more accurately. The land use classification methods to create the Coarse Land Use, Finer Land Use, and LPA Land Use populations are also novel and provide guidance for many next steps in developing even better methods.

The Tobit models show that, along with sociodemographic traits that stayed constant for all models, land use is a significant contributor to the frequency of trips and miles traveled, no matter which classification method is tested for land use surrounding a household place of residence. This shows that land use is a significant contributor to those travel behavior traits of interest, and further shows that land use should be included in population synthesis.

The use of LPA for land use classification provides benefits that make it the most valuable method of those tested. It produces land use indicators that explain travel behavior and shows a finer-grain ability to transfer behavioral data to an entire population. The method also allows for inclusion of all kinds of variables in the classification. This makes it easily transferrable to whatever region researchers might be interested in synthesizing. If employment variables are not available, but other variables that could contribute to a land use classification are available, then those can be used. Compared to the previous methods, the LPA method is also more statistically sound: It reduces the statistical uncertainty associated with the previous method, and it demonstrates the potential to explain two key behavioral traits—number of trips and vehicle miles traveled. Most importantly, the LPA method reproduces a spatial distribution that most closely resembles the spatial distribution that is expected for California.

In this application of the LPA classification, we used the entire suite of 17 types of business establishments, enumerating their number of employees and producing block group-level counts that were then used as indicators for land use LPA. One possibility for future applications is to use fewer employee categories that are more location-based, being selective about what to choose based on substantive theories about the types of employment that are more prevalent in certain regions of California than in others. Along with including

employee density categories in the classification, one can also include network density, public transportation services, and/or distances to major employment centers. This expanded repertory of data will properly place the employee density where it should be for categories where they might not be present at the business "storefront." This will also enable comparison with somewhat more traditional accessibility indicators (Chen et al., 2011).

In addition, taking into account the market size of each area (e.g., using something like number of employees divided by the number of residents in a block group) may also allow to account for any jobs-housing imbalance. This would mean it would be normalized by the population of an area, and we would be comparing ratios of different employment types instead of just the number of people employed per square kilometer. Finally, NETS is a longitudinal record of business establishments that can be used further to study the evolution of land use, and, if combined with a longitudinal version of LPA, enables the study of spatial transitions in land use and include in the land use classification history of development. All these are left as future tasks.

6 References

Adiga, A., Agashe, A., Arifuzzaman, S., Barrett, C. L., Beckman, R., Bisset, K., … Xie, D. (2015). *Generating a synthetic population of the United States. Technical Report NDSSL 15-009*.

Asparouhov, T., & Muthén, B. O. (2012). Using Mplus TECH11 and TECH14 to test the number of latent classes. *Mplus Web Notes*, (14), 1–17. Retrieved from https://www.statmodel.com/examples/webnotes/webnote14.pdf

Auld, J., & Mohammadian, A. (2010). Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels. *Transportation Research Record: Journal of the Transportation Research Board*, *2175*, 138–147. https://doi.org/10.3141/2175-16

Auld, J., & Street, W. T. (2008). Population Synthesis With Region-Level Control Variable Aggregation. *Transportation Research Record*, *531*, 1–17.

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., & Rossiter, D. (2005). SimBritain: A spatial microsimulation approach to population dynamics. *Population, Space and Place*, *11*(1), 13–34. https://doi.org/10.1002/psp.351

Bar-Gera, H., Konduri, K. C., Sana, B., Ye, X., & Pendyala, R. M. (2008). Estimating survey weights with multiple constraints using entropy optimization methods. In *Proceedings of 88th Annual Meeting of the Transportation Research Board*. Washington, D.C.: National Research Council. Retrieved from http://rampendyala.weebly.com/uploads/5/0/5/4/5054275/popsynentropyapproach_bargerakondurisanayependyala_transportationrev.pdf

Beckman, R. J., Baggerly, K. A., & McKay, M. D. (1996). Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*. https://doi.org/10.1016/0965-8564(96)00004-3

Bhat, C. R., & Eluru, N. (2009). A copula-based approach to accommodate residential self-selection effects in travel behavior modeling. *Transportation Research Part B: Methodological*, *43*(7), 749–765. https://doi.org/10.1016/j.trb.2009.02.001

California Department of Transportation. (2013). *2010-2012 California Household Travel Survey Final Report Appendix*.

Cao, X. (Jason), Mokhtarian, P. L., & Handy, S. L. (2009). *Examining the impacts of residential self selection on travel behaviour: A focus on empirical findings*. *Transport Reviews* (Vol. 29). https://doi.org/10.1080/01441640802539195

Casati, D., Müller, K., Fourie, P. J., Erath, A., & Axhausen, K. W. (2015). Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record: Journal of the Transportation Research Board*, *2493*, 107–116. https://doi.org/10.3141/2493-12

Celeux, G., & Soromenho, G. (1996). An entropy criterion for assessing the number of clusters in a mixture model. *Journal of Classification*, *13*(2), 195–212. Retrieved from https://doi.org/10.1007/BF01246098

Chen, Y., Ravulaparthy, S., Deutsch, K., Dalal, P., Yoon, S. Y., Lei, T., … Hu, H.-H. (2011). Development of Indicators of Opportunity-Based Accessibility. *Transportation Research Record: Journal of the Transportation Research Board*, *2255*(2), 58–68. https://doi.org/10.3141/2255-07

ESRI. (2013). ArcGIS Desktop: Release 10.2. *Redlands CA*.

ESRI. (2016a). How Kernel Density works. Retrieved from

http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/how-kernel-density-works.htm
ESRI. (2016b). Zonal Statistics as Table—Help | ArcGIS for Desktop. Retrieved from http://desktop.arcgis.com/en/arcmap/10.3/tools/spatial-analyst-toolbox/zonal-statistics-as-table.htm
Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. *Transportation Research Part B: Methodological*, *58*, 243–263. https://doi.org/10.1016/j.trb.2013.09.012
Feldman, M. (2017). National Establishment Time-Series. Retrieved from http://maryannfeldman.web.unc.edu/data-sources/longitudinal-databases/national-establishment-time-series-nets/
Greene, W. H. (2003). *Econometric analysis*. Pearson Education India.
Guo, J. Y., & Bhat, C. R. (2008). Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record*, *2014*(1), 92–101. https://doi.org/10.3141/2014-12
Konduri, K. C., You, D., Garikapati, V. M., & Pendyala, R. M. (2016). Enhanced synthetic population generator that accommodates control variables at multiple geographic resolutions. *Transportation Research Record, Journal of the Transportation Research Board*, *16–6639*(16), 40–50.
Kwan, M. (2000). Gender differences in space-time constraints. *Area*, *32*, 145–156. https://doi.org/10.1111/j.1475-4762.2000.tb00125.x
Kwan, M.-P. (n.d.). Space-Time and Integral Measures of Individual Accessibility: A Comparative Anal ysis Using a Point-based Framework. https://doi.org/10.1111/j.1538-4632.1998.tb00396.x
Kwan, M. P. (1999). Gender and individual access to urban opportunities: A study using space–time measures. *Professional Geographer*, *51*(2), 211–227. https://doi.org/10.1111/0033-0124.00158
Ma, L., & Srinivasan, S. (2015). Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations. *Computer-Aided Civil and Infrastructure Engineering*, *30*(2), 135–150. https://doi.org/10.1111/mice.12085
MARG. (2016). PopGen: Synthetic population generator. Retrieved from http://www.mobilityanalytics.org/popgen.html
Masyn, K. E. (2013). Latent class analysis and finite mixture modeling. In *The Oxford handbook of quantitative methods (Vol 2): Statistical analysis.* (pp. 551–611). Retrieved from http://0-search.ebscohost.com.library.unl.edu/login.aspx?direct=true&db=psyh&AN=2013-01010-025&site=ehost-live
McBride, E. C., Davis, A. W., Lee, J. H., & Goulias, K. G. (2016). Spatial Transferability Using Synthetic Population Generation Methods (Contract Number: 65A0528). University of California, Santa Barbara. Department of Geography, GeoTrans Laboratory. Contract Number: 65A0528. Submitted to the California Department of Transportation, May 2016.
McBride, E. C., Davis, A. W., Lee, J. H., & Goulias, K. G. (2017). Incorporating land use in synthetic population generation methods and transfer of behavioral data. *Transportation Research Record: Journal of the Transportation Research Board, 2668*, 11–20.

Mokhtarian, P. L., & Cao, X. (2008). Examining the impacts of residential self-selection on travel behavior: A focus on methodologies. *Transportation Research Part B: Methodological*, *42*(3), 204–228. https://doi.org/10.1016/j.trb.2007.07.006

Muthén, L., & Muthén, B. (1998). *Mplus user's guide (7th ed.)*. *Los Angeles: Author*. https://doi.org/10.1111/j.1600-0447.2011.01711.x

NUSTATS. (2013). *2012 California Household Travel Survey Final Report*. Austin, TX. Retrieved from http://www.dot.ca.gov/hq/tpp/offices/omsp/statewide_travel_analysis/Files/CHTS_Final_Report_June_2013.pdf

Nylund, K. L., Asparouhov, T., & Muthen, B. O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A Monte Carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*(4), 535–569. https://doi.org/10.1080/10705510701575396

Pendyala, R. M., Bhat, C. R., Goulias, K. G., Paleti, R., Konduri, K. C., Sidharthan, R., … Christian, K. P. (2012). 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. https://doi.org/10.3141/2303-08

Pendyala, R. M., Bhat, C. R., Goulias, K. G., Paleti, R., Konduri, K., Sidharthan, R., & Christian, K. P. (2013). *SimAGENT Population Synthesis*. Santa Barbara, CA.

Pendyala, R. M., Konduri, K. C., & Christian, K. P. (2011). *PopGen 1.1 User's Manual*. USA.

PopGen: Population Generator. (n.d.).

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. https://doi.org/10.1007/s11116-011-9367-4

Ravulaparthy, S., & Goulias, K. G. (2011). *Forecasting with Dynamic Microsimulation: Design, Implementation, and Demonstration*. UC Berkeley.

Rees, H., & Maddala, G. S. (1985). Limited-Dependent and Qualitative Variables in Econometrics. *The Economic Journal*, *95*(378), 493. https://doi.org/10.2307/2233228

Turner, T., & Niemeier, D. (1997). Travel to work and household responsibility: New evidence. *Transportation*, *24*(4), 397–419. https://doi.org/10.1023/a:1004945903696

United States Census Bureau. (2012). Geographic Terms and Concepts - Block Groups. Retrieved from https://www.census.gov/geo/reference/gtc/gtc_bg.html

US Census Bureau. (2016). American Community Survey: When to Use 1-year, 3-year, or 5-year Estimates. Retrieved from https://www.census.gov/programs-surveys/acs/guidance/estimates.html

Waddell, P. (2002). Urbansim: Modeling urban development for land use, transportation, and environmental planning. *Journal of the American Planning Association*, *68*(3), 297–314. https://doi.org/10.1080/01944360208976274

Walls & Associates. (2012). *National Establishment Time - Series (NETS) Database : 2012 Database Description*. Denver, CO. Retrieved from http://exceptionalgrowth.org/downloads/NETSDatabaseDescription2013.pdf

Ye, X., Konduri, K. C., Pendyala, R. M., Sana, B., & Waddell, P. (2009). A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In *Proceedings of 88th Annual Meeting of the Transportation*

Research Board. Washington, D.C.: National Research Council. Retrieved from http://rampendyala.weebly.com/uploads/5/0/5/4/5054275/populationsynthesizerpaper_trb2009.pdf

Zhu, Y., & Ferreira Jr., 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. https://doi.org/10.3141/2429-18