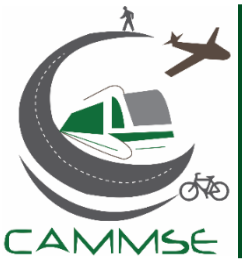

Center for Advanced Multimodal Mobility Solutions and Education

Project ID: 2017 Project 05

STOCHASTIC MULTIMODAL NETWORK MODELING: HIDDEN MARKOV MODEL BASED SYNTHETIC POPULATION GENERATION FOR USE IN MICROSIMULATION MODELS OF TRANSIT SYSTEMS

Final Report

by

Karthik C Konduri, Ph.D. (ORCID ID: https://orcid.org/0000-0003-2788-9455)
Associate Professor, Department of Civil and Environmental Engineering
University of Connecticut
261 Glenbrook Road, Unit 3037, Storrs, CT 06269-3037
Phone: 1-860-486-2733; Email: karthik.konduri@uconn.edu

and

Amit Mondal (ORCID ID: https://orcid.org/0000-0002-3691-1506)
Graduate Research Assistant, Department of Civil and Environmental Engineering
University of Connecticut
261 Glenbrook Road, Unit 3037, Storrs, CT 06269-3037
Phone: 1-917-209-2895; Email: amit.mondal@uconn.edu

for

Center for Advanced Multimodal Mobility Solutions and Education
(CAMMSE @ UNC Charlotte)
The University of North Carolina at Charlotte
9201 University City Blvd
Charlotte, NC 28223

July 2018

ACKNOWLEDGEMENTS

This project was funded by the Center for Advanced Multimodal Mobility Solutions and Education (CAMMSE @ UNC Charlotte), one of the Tier I University Transportation Centers that were selected in this nationwide competition, by the Office of the Assistant Secretary for Research and Technology (OST-R), U.S. Department of Transportation (US DOT), under the FAST Act. The authors are also very grateful for all of the time and effort spent by DOT and industry professionals to provide project information that was critical for the successful completion of this study.

DISCLAIMER

Table of Contents

List of Figures

List of Tables

EXCUTIVE SUMMARY

In order to apply microsimulation-based models of land use and travel demand, socio-economic and demographic attributes about all individuals in a region is required. This disaggregate level information is not readily available and people resort to population synthesis procedures. These procedures combine readily available information in the form of sample data and marginal distributions to generate the required inputs. In this study, a simulation-based technique for population synthesis using a Hidden Markov Model (HMM) framework is presented. An important feature of the proposed approach is the ability to generate more heterogeneous synthetic households and persons. The proposed simulation-based approach is demonstrated using a case study for Connecticut. Synthetic population is generated for two block groups in Connecticut under alternate configurations. A comparative analysis is carried out to highlight the feasibility and applicability of the proposed approach in generating consistent multilevel agents while adhering to geography-based heterogeneity. The current work is similar in spirit to other recent simulation-based generators, however, there are two important contributions. First, a hierarchical transition structure is proposed in the HMM-based model, to capture the dependencies across household and person-level attributes. Thus, the procedure ensures that both household and person level attributes are controlled simultaneously. Second, the transition matrices are estimated at the geography level incorporating the sample as well as marginal information available. This helps synthesize populations that are more accurate and consistent with available information.

Chapter 1. Introduction

1.1 Problem Statement

Over the past few decades, microsimulation models have been gaining increasing interest in land-use and transportation planning. In these models, behaviors of interest are simulated at the individual level while explicitly accounting for the environment in which they make decisions and the constraints and interactions they experience. Subsequently, these decisions are aggregated spatially and temporally to understand how a system will perform in alternate environments (*1*–*3*). Microsimulation models are better suited for assessing impacts of different policies of interest because of their focus on the individual decision maker and the underlying decision-making processes. They generate results at rich spatial and temporal resolution allowing planners to draw insights that are otherwise not possible using more aggregate model forms (*4*, *5*).

Disaggregate microsimulation models require detailed household and person level information for each individual agent. However, such information is not readily available owing to a variety of reasons including privacy issues and resource limitations among others. Instead, detailed information for a sample of the population (often referred to as sample data) and aggregate information (often referred to as marginal distribution data) about the entire population are available, typically from Census Bureaus or equivalent bodies (*6*). Analytical procedures are then applied to combine them together to create detailed records for all individuals in a region. This process is often referred to as synthetic population generation. With growing interest in microsimulation models, interest in developing synthetic population generators (SPG) has also increased. A brief overview of these approaches along with some examples is next chapter. A detailed review of synthesizers can be found in (*1*, *4*, *7*, *8*).

1.2 Objectives

The objective of this report is to (1) propose a new simulation-based population synthesis technique using a hierarchical structure of Hidden Markov Model (HMM) that can generate household and person level attributes simultaneously; (2) propose a novel procedure to estimate the transition probabilities using both disaggregate sample datasets and geography-based attribute controls to address spatial heterogeneity in synthetic population.

1.3 Report Overview

The rest of the paper is structured as follows. Chapter 2 includes a brief discussion on literature review of several popular synthesis techniques. In Chapter 3, the general HMM approach is described. In addition, the section also discusses how this approach can be adapted to perform population synthesis that controls for both household and person attributes simultaneously. Chapter 4 presents the case study including data preparation, model setup, results, and discussion of findings. Finally, concluding thoughts along with limitations and future extensions are presented in Section 5.

Chapter 2. Literature Review

2.1 Introduction

This chapter describes the most popular population synthesizing techniques that has been evolved over the years to overcome the limitations of their predecessors. This discussion is very helpful to understand the current practice in population synthesis and their use-cases in various domains. The techniques within different SPGs can be clustered into two main groups: fitting-based approaches and combinatorial optimization (CO) procedure (*9*).

2.2 Fitting-based Approaches

Fitting-based approaches focus on estimating a multiway distribution of agents' attributes. Subsequently, agents are generated from the sample based on the estimated multiway distribution, and Monte Carlo based sampling technique. Iterative Proportional Fitting (IPF) is the most dominant fitting-based technique in the literature. Deming and Stephan (1940) first introduced IPF to calculate cell values of a multiway distribution through an iterative algorithm such that observed marginal distribution are matched (*10*). Beckman et al. (1996) developed a synthetic population generator based on the IPF based procedure (*11*). This was one of the first SPGs and has been widely adopted in many operational disaggregate models in the past. A number of SPGs have been developed since to address different issues and limitations with the Beckman et al. (1996) procedure. For example, Guo and Bhat (2007) proposed an IPF-based procedure for controlling both household and person level marginal distributions (*12*). Also, addressing the same problem of household and person control matching, Arentze et al. (2007) introduced the concept of relational matrices in the IPF procedure (*13*). Ye et al. (2009) developed Iterative Proportional Updating (IPU) – a heuristic iterative procedure that also accounts for both household and person level marginal distributions (*14*). More recently, Konduri et al. (2016) extended IPU to control for marginals at multiple spatial resolutions (*8*). For high dimensional contingency table, Pritchard and Millar (2012) introduced a sparse matrix-based data structure in IPF framework to deal with memory consumption issues while controlling both household and person-level attributes simultaneously (*15*). There are several other variants of IPF implementations including hierarchical and multi-stage IPF that focus on fitting both household and person-level attributes maintaining their inter-level association (*16*, *17*).

2.3 Combinatorial Optimization (CO) Approaches

Along with fitting-based techniques, CO approaches have been emerging as a promising alternative to population synthesis. CO approaches also require both sample and marginal distributions. They also employ an iterative procedure to generate population for a geographical unit. The iterative procedure begins with a selection of a pool of agents and assessing match with the given marginal distributions. At each step of the iteration, agents may be added and/or replaced with a new agent from the sample dataset until an appropriate goodness of fit is achieved. Voas and Williamson (2000) implemented this approach by optimizing the sample weights such that the synthetic population matches the observed attributes for a geographical unit (*18*). Abraham et al. (2012) applied CO algorithm to control both household and person level controls for multiple geographic resolution (*19*). Simulated Annealing is another CO technique that follows a probabilistic reweighting procedure to pull a suitable set of agents from the sample (*18*, *20*). There have also been studies comparing these two popular approaches (*5*, *21*, *22*). While CO has been claimed to be superior in terms of performance, the fitting-based approaches are easier to implement and more scalable.

2.4 Simulation-based Approaches

More recently, there has been a third category of SPGs namely simulation-based approaches. The main advantage of these over earlier approaches is the ability to create more diverse synthetic populations. In both fitting-based and CO approaches, records from the sample are cloned to create a synthetic population. This can lead to lumpiness in the synthetic population and the synthesized results may not capture the full underlying distribution. Simulation-based approaches use a variety of techniques to model the joint distribution of household and person attributes underlying the population. Subsequently, synthetic population is generated by simulating draws from the joint distribution to create agents and their attributes. Caiola and Reiter (2010) implemented Random Forests-based synthesizer that can capture the attribute relationships effectively and performs well for high dimensional configuration (*23*). Sun and Erath (2015) proposed a probabilistic approach based on the Bayesian Network model (*4*). The study demonstrated how a Bayesian network can be incorporated in population synthesis to understand the underlying structure of population with a large set of attributes. Farooq et al. (2013) introduced a simulation-based approach for population synthesis where they implemented parametric models for conditional probability estimation and

applied Markov Chain Monte Carlo (MCMC) procedure for generating population (*7*). More recently, Saadi et al. (2016) developed a new population synthesis technique using a Hidden Markov Model (HMM) (*9*). In this study, authors note that the HMM framework is more adaptable and efficient when it comes to fusing multiple micro-samples in model training and preserving more heterogeneous composition in synthetic population. This report builds on the work by Saadi et al. (2016) by addressing two important limitations.

First, in their study, the synthetic population generation was only limited to persons; households are not generated. The study acknowledges the need for extending this work so that that households and persons are both synthesized while also accounting for the available household and person level information. In this work, a hierarchical transition structure is proposed in the HMM-based model to capture the joint distribution of both households and persons simultaneously. The model captures the dependencies across household and person-level attributes and can be used to simulate both households and persons.

Second, in their work, they only use a single model to generate a synthetic population for all geographies in a region. This approach may compromise on the heterogeneity in the population across geographies. Additionally, the model doesn't incorporate the marginal distributions information that is available. In other words, this approach ignores information that can potentially be used to enhance the synthetic population. In this paper, the transition matrices for the HMM are estimated using a novel procedure that incorporates information available from both the sample and the marginal distributions. This, in turn, helps develop populations that are more accurate and consistent with available information.

The feasibility and the applicability of proposed HMM model and the estimation procedure are demonstrated by generating a synthetic population using data from the American Census Bureau for 2 block groups in Connecticut. Synthetic population was generated under a variety of scenarios mimicking existing simulation-based procedures. Results are compared across scenarios to highlight the contributions of the proposed approach.

Chapter 3. Methodology

3.1 Overview of HMM Framework

One way to characterize the attributes of an individual (household) is as a sequence of characters. The length of the sequence is equal to the number of attributes of interest. Each character in the sequence represents a value for the attribute. Hidden Markov Models (HMM) can be used to characterize such a sequence. HMM are probabilistic models that can be used for any sequence labeling problem (*24*, *25*). HMMs are very dynamic in a sense that they can conceptualize any complex sequence analysis model using graphical methods (*26*). To explain the functional aspects of HMM, a simple toy example is presented. Consider that one is interested in understanding the educational journey for those who are currently employed. Assuming everyone employed has completed middle school, the educational journey can be represented by an HMM shown in Figure 3.1.

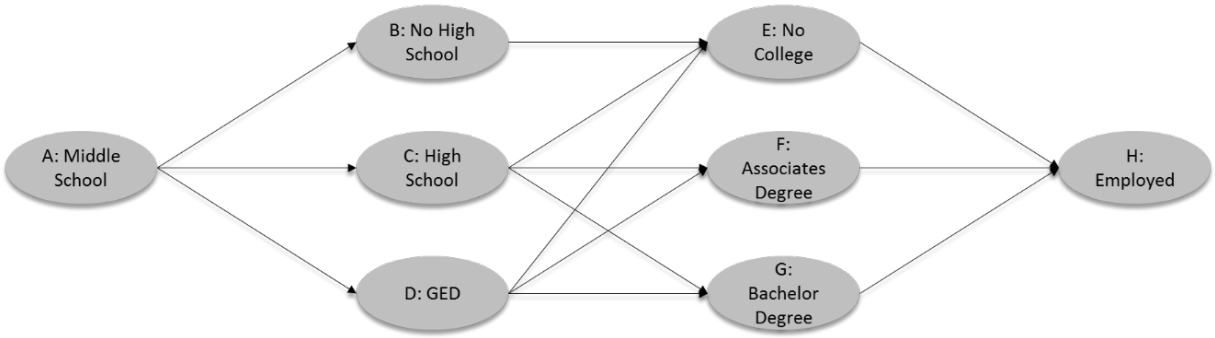

Figure 3.1 Structure of a Simple HMM

Each oval represents a state (the state is represented by a character and the associated definition is presented in the figure) and each directed link represents a potential transition from one state to the next. In this figure, a path consisting of a series of directed links beginning with the state A (i.e. "Middle School" Graduate) and ending in state H (i.e. "Employed") represents an educational journey. For example, a sequence ACEH represents an educational journey where someone completed High School after Middle School, skipped college and entered the workforce and got employed. Transitions are possible from any of the states to any other state. However, for any given use case, only a subset of transitions is reasonable and/or supported by data. This type of HMM is called a Bakis HMM where some transitions have zero probability (*27*). For example, a

transition from state A to state H is probably not supported by data. On the other hand, transition from state C to state B is inconsistent. In HMM, there are some dummy states that helps join different parts of the model without disturbing the actual transition patterns. For sequence generation problems, dummy states are very helpful for proper identification of different portions of the model. The states that emit a symbol or character are referred to as active states.

HMM comprises of three main parameters: transition probability matrix, initial probability vector, and emission probability matrix. The architecture of HMM is built with a finite set of states represented by vector $A = \{A_1, A_2, A_3, \dots, A_N\}$ where N is the total count of states. Each state is associated with a multidimensional probability distribution that regulates the transition to other possible states (*28*). Transitions from state i to state j are governed by a transition probability matrix T where $T = \{P(t_{ij})\}$ and each element in the matrix represents the transition from state i to state j. In other words, a given state k is not observable directly. Instead, state k manifests itself in the form of an outcome from an observation set $\beta_k = \{\beta_{1k}, \beta_{2k}, \beta_{3k}, \dots, \beta_{mk}\}$ where m is the size of the set. M is the set of all observation symbols corresponding to the N states. An observation symbol corresponding to state k is observed based on an emission probability vector $E_k = \{P(\beta_{mk})\}$. HMM also requires a set of initial probabilities that represents the state from which the sequence starts. The set of initial probabilities is given by a vector $\pi = \{P(i)\}$. In terms of the structure of these elements, T is a $N \times N$ dimensional matrix, E is a $N \times M$ dimensional matrix and π is N dimensional vector. In addition to the above parameters, HMM also incorporates some logical and consistency constraints.

$$0 \leq P(t_{ij}) \leq 1, \quad 1 \leq i, j \leq N \tag{1}$$

$$\sum_{j=1}^{N} P(t_{ij}) = 1, \quad 1 \leq i \leq N \tag{2}$$

$$0 \leq P(\beta_{mk}) \leq 1, \quad 1 \leq m \leq M, \quad 1 \leq k \leq N \tag{3}$$

$$\sum_{m=1}^{M} P(\beta_{mk}) = 1\,, \quad 1 \leq k \leq N \tag{4}$$

$$\sum_{i=1}^{N} P(i) = 1 \tag{5}$$

As noted above, an external observer can only see the outcome corresponding to a state and the actual states are hidden. Therefore, this configuration of the Markov model is called Hidden

Markov Model. Alternatively, if the state is observed directly then it is commonly referred to as just a Markov Model. Depending on the variant of the Markov Model that is applicable for a given situation, alternative procedures are available for estimating the parameters.

To incorporate more complex models in HMM framework, researchers have been developing extensions to HMM such as Hierarchical HMM (HHMM), Layered HMM (LHMM) and Nested HMM (NHMM) (*29–32*). Among these variants, HHMM is of particular interest given its relevance to the population synthesis approach proposed in the next subsection. HHMM allows one to organize states using a hierarchical structure. In HHMM, there are multiple root states that can each be represented as an individual HMM. These root states are stacked as layers in a hierarchical structure to form the full HHMM model. When a transition occurs to a root states, typically the corresponding underlying HMM is executed and the model then proceeds to the next root state in HHMM hierarchy. The HMMs within root states can have shared connections across root states allowing for a shared structure and recurring pass in the model. This hierarchical model structure is a key ingredient to extend the work by Saadi et al. (2016) to deal with the multi-level population synthesis i.e. synthesizing both households and persons. The basic idea is that person models can be thought of as the descendent of root states that can embedded in a hierarchical fashion within a household model which again can be a descendent of another root state. This approach allows for ensuring dependencies between person attributes and household attributes. Building and training a HHMM is computationally very expensive (*33*). HHMMs can be converted to its equivalent flat HMM without compromising the structural integrity of the model (*29*, *34*). An HHMM that has shared transition structure can be converted to flat HMM by duplicating the sub-models. Though the flattening addresses the issue of computational tractability, it comes at the expense of increase in dimension of the HMM. In the next subsection, the proposed approach to implementing multi-level population synthesis using HHMM intuition and HMM equivalency is presented. Further, since the states are observed directly in the population synthesis case, we are working with Markov Model variant of the HMM i.e. state and observed outcome are same and the emission probability vector E_k for any state k is given as {1}. In the remaining text while the term HMM will be used, it must be noted that the Markov Model variant is what is adopted in the synthesis approach.

3.2 HMM-based Population Synthesis

3.1.1. Hierarchical Configuration of Household and Person Model

The first objective of this research was to use the HMM framework to synthesize not only households but also persons within the households. While the work by Saadi et al. (2016) can be used to synthesize households and persons separately, an additional procedure is needed to tie them together. The HHMM forms the basis for incorporating both household and person synthesis jointly. The flattening of HHMM and its equivalency to HMM is adopted to estimate the model. In HMM, states are considered as attribute categories for both household and person models. Further, key household attributes are used to generate root states. Then person models consistent with the defined root states are embedded to build the hierarchical structure. Building the household model is the same as Saadi et al. (2016). Each household attribute and associated categories serve as active states in the household model. As noted earlier, a hierarchical tree structure is used to build household level HMM model and then to incorporate the person level HMM models within the household model. Subsequently, this allows synthesis of both household and person attributes together while also accounting for the consistency between the household and person level characteristics. A simple household model is shown in Figure 3.2.

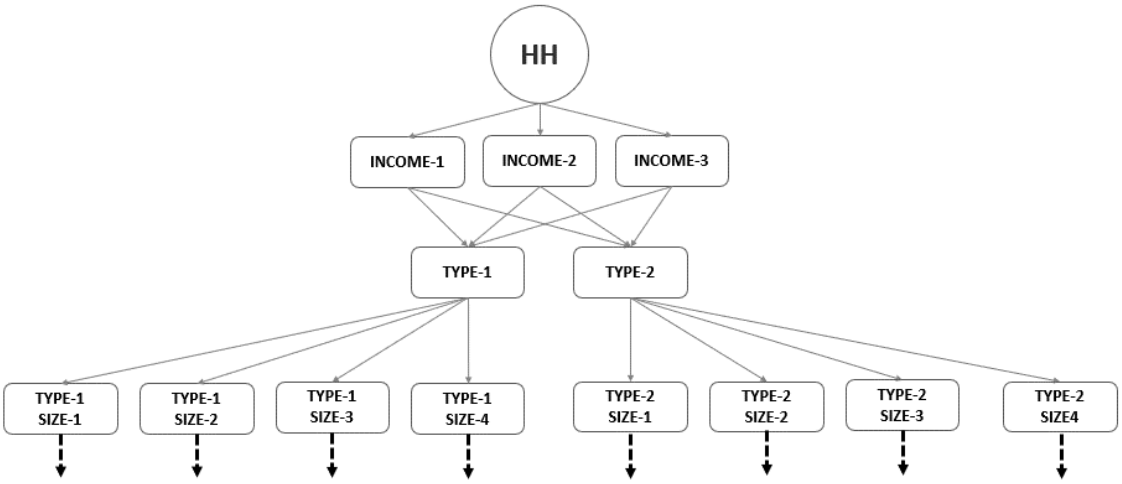

Figure 3.2 Simple Structure of a Household Model

Assuming that there are two types of households – family and non-family, the number of household members is largely influenced by the household composition in that household. Therefore, the states of SIZE attribute are branched out depending on the states of TYPE attribute in Figure 3.2.

In similar fashion, it is possible to accommodate other household attributes such as PERSON UNDER 18 YEARS. In that case, the states of PERSON UNDER 18 YEARS have separate branches originating from each of the SIZE states. The number of states for attribute PERSON UNDER 18 YEARS is governed by the originating SIZE state. For example, for a household with three persons, there will be maximum of two persons with age below 18 years (assuming that the householder's age is above 18 years). Therefore, there will be three possible transitions from the SIZE-3 state: NO PERSON UNDER 18 YEARS, 1 PERSON UNDER 18 YEARS, and 2 PERSONS UNDER 18 YEARS. Now by defining the states of PERSON UNDER 18 YEARS as a root states, person models for these household composition can be embedded based on the hierarchy. For example, in the state NO PERSON UNDER 18 YEARS, a model representing householder, a model for second person, and a model for third person are embedded. On the other hand, for the state 2 PERSONS UNDER 18 YEARS, a householder model is embedded one time and the person model for those under 18 is embedded two times. Figure 3.3 illustrates the fully embedded SIZE-3 branch based on the household composition.

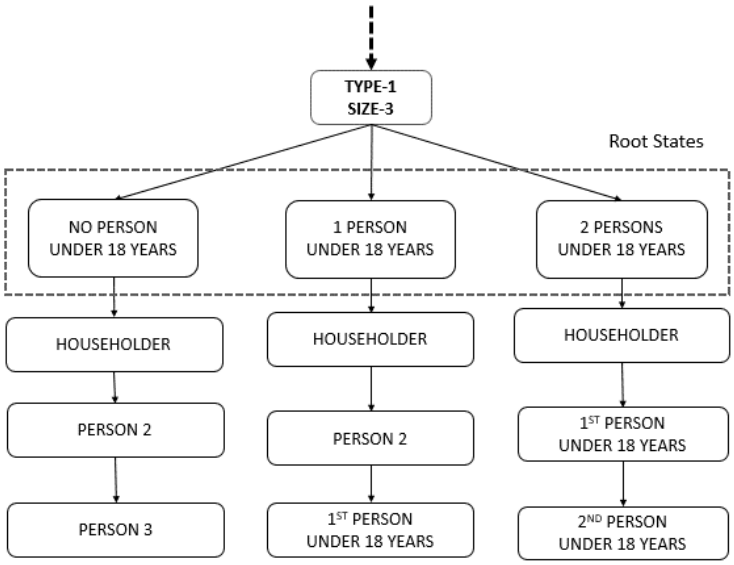

Figure 3.3 Connection of Person Model with Household Model

As can be seen, the proposed HMM structure allows the generation of household attributes in the upper level of the model and then proceeds towards lower level to generate person level attributes. This model can also be configured to deal with open ended categories. Choice states can be placed in the model structure to make a decision about the next transition to an embedded HMM subject to some constraint. As noted earlier, each person model is an individual HMM similar to Saadi et

al. (2016) that are constructed using person-level attributes. These models are duplicated as necessary within the household states. However, the idea of recurring pass allows the use of same person model without duplicating thus reducing the overall dimensions of the HMM. These individual models have a simple transition structure. Nonetheless, the order of attributes is always important to capture the conditional transitions between attributes. In order to preserve the relationship of persons belonging to a particular household, person models have root and decision states and the concept of guaranteed pass helps to build inter-person connections. This is essential to deal with inconsistent inter-personal relationship during simulation. Figure 3.4 illustrates an example of introducing decision states to deal with gender issues while simulating the householder and second person for family households. This helps to obtain consistent gender information of the second person based on householder gender information using a conditional probability distribution.

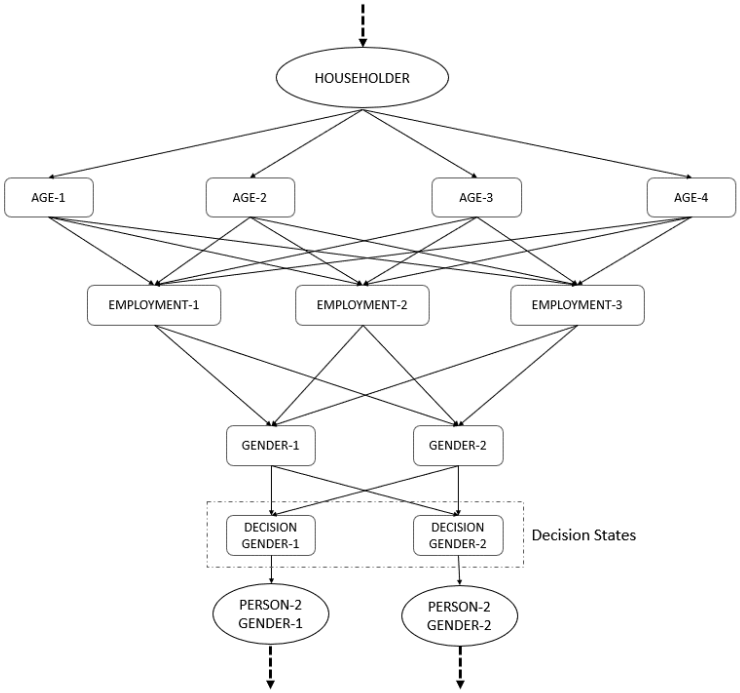

Figure 3.4 Configuration of a Single Person Model

Figure 3.5 illustrates a simple hierarchical structure of proposed HMM framework. For the purpose of illustration, only a handful of attributes is configured in the proposed hierarchical structure.

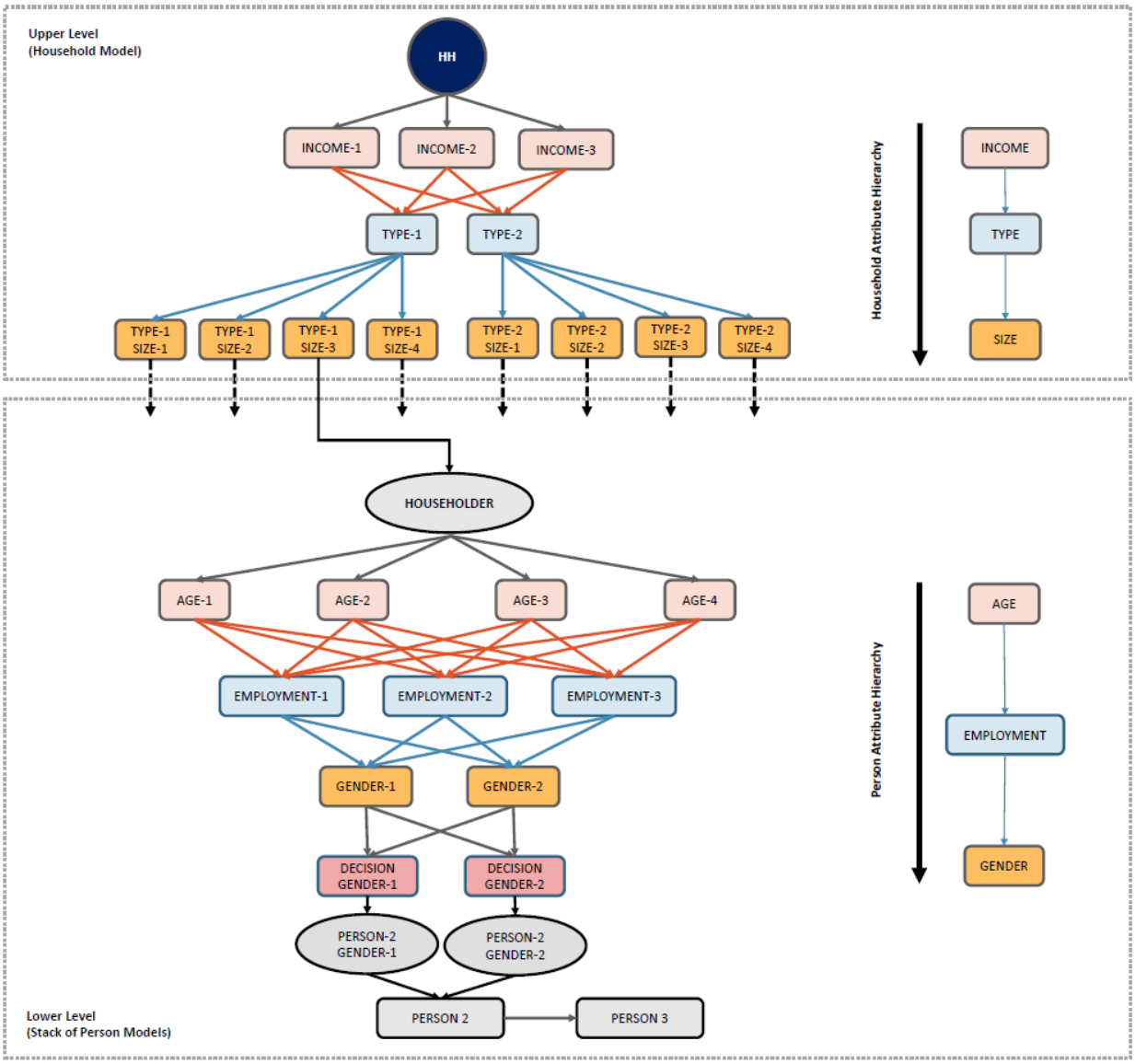

Figure 3.5 A Simple Hierarchical Structure of Proposed HMM Framework

3.1.2. Incorporating Geography-based Controls

The second objective of this research was to present an approach for estimating the transition probabilities that not only accounts for the information contained in the sample data but also accounts for the marginal distributions so that population that agree with available information can be generated. For typical HMM models, transition frequencies between the states are estimated from an observed sample. In population synthesis, this direct procedure can be used as outlined in the study conducted by Saadi et al (2016). However, this direct procedure has a major limitation in terms of matching attribute marginals for a geographic unit. The transition patterns of attributes estimated directly from sample data do not represent the real population structure for a geographic

unit. As a result, there will be large differences between synthetic population results and observed marginal distributions for a geographic unit. Saadi et al. (2018) proposed a hierarchical procedure to deal with this limitation by integrating HMM and IPF under the same framework (*35*). However, their approach doesn't consider accounting for both household and person marginal simultaneously. This study proposes a new procedure for estimating the transition counts using both aggregate and disaggregate information. In the proposed approach, weights for sample households are first estimated using Iterative Proportional Updating algorithm proposed by Ye et al. (2009) using all available household and person level marginals. These weights are then used to estimate the transition probabilities. The transition probabilities thus generated conform to the available marginal distributions. Subsequently, the synthetic population also accounts for this information and fewer deviations are observed with respect to available marginals.

Chapter 4. Case Study and Result Analysis

4.1 Data Preparation

A case study was conducted to demonstrate the proposed HMM population synthesis framework and associated transition probability estimation routine. The study considers 4 household attributes (household type, household income, presence of persons under 18 years, and household size) and 4 person attributes (age, employment, ethnicity, and gender) to generate synthetic population for two block groups in Connecticut (IDS: 0427002, 2531001). Block groups are selected from two different Public-Use Micro Areas (PUMA). Both the aggregate and disaggregate data are collected from US Census Bureau. The aggregate data is processed from American Community Survey (ACS) 2010-2014 Summary datasets and disaggregate data is collected from corresponding Public Use Micro Sample (PUMS). The disaggregate data contains information on 70,221 households and 181,082 persons at PUMA level. Household and person attributes are defined as categorical variables, and the description of attributes and summary of aggregate marginals for two block groups are listed in Table 4.1.

A hierarchical transition structure was developed using the household and person attributes mentioned above. The order of attributes in the household model was household income, household type, household size, and presence of persons under 18 years. The household model has two major branches depending on the household types, because family and non-family households have completely different household compositions. Then based on the household size, categories for presence of persons under 18 years form the second set of branches. These categories are also set as root states to embed the person models. The order of attributes in person model is age, employment, ethnicity, and gender. The hierarchical transition matrix contains a total of 4203 states including active, dummy and decision states. Therefore, the dimension of transition matrix considered in this case study is 4203 by 4203.

Table 4.1 Description of Control Variables and Marginals for Two Block Groups

Attributes	Description	BG0427002	BG2531001
HHTYPE_1	Family Household	448	676
HHTYPE_2	Non-family Household	595	365
HHI_1	Less than 15000USD	65	5
HHI_2	15000USD - 25000USD	85	103
HHI_3	25000USD - 50000USD	217	308
HHI_4	50000USD - 75000USD	199	217
HHI_5	75000USD - 100000USD	102	96
HHI_6	100000USD - 150000USD	270	186
HHI_7	More than 150000USD	105	126
HHC_1	Presence of persons under 18 years (YES)	204	310
HHC_2	Presence of persons under 18 years (NO)	839	731
HHSIZE_1	1 person	541	289
HHSIZE_2	2 persons	288	318
HHSIZE_3	3 persons	111	302
HHSIZE_4	4 persons	103	96
HHSIZE_5	5 persons	0	17
HHSIZE_6	6 persons	0	4
HHSIZE_7	7 persons or more	0	15
PAGE_1	Less than 14 years	288	333
PAGE_2	15 years - 17 years	0	170
PAGE_3	18 years - 24 years	165	289
PAGE_4	25 years - 44 years	440	617
PAGE_5	45 years - 59 years	627	737
PAGE_6	60 years - 74 years	295	216
PAGE_7	75 years or more	98	183
PEMPLOY_1	Less than 16 years	288	368
PEMPLOY_2	Employed for last 12 months	1,172	1,618
PEMPLOY_3	Unemployed for last 12 months	453	559
PE_1	Caucasian	1,504	2,486
PE_2	Others	409	59
PG_1	Male	756	1,238
PG_2	Female	1,157	1,307

The transition structure is estimated using three different approaches to highlight the feasibility and applicability of the proposed IPU based estimation approach.

I. **Case 1**: The transition matrix is estimated directly using the entire PUMS data resulting in a general transition probability distribution for all block groups. This is similar in spirit to the approach proposed by Saadi et al. (2016).

II. **Case 2**: For each block group, the transition matrix is estimated using the sample records of only those households that belong to corresponding PUMA geographies. PUMA 2300 that is associated with block group BG0427002 has 2,000 household records and 5,196 person records. PUMA 100 associated with block group BG2531001 includes 1,868 household records and 4,546 person records. Since the block groups are selected from two different PUMA, two different transition matrices were prepared.

III. **Case 3**: In this case, the transition matrices are estimated using the proposed IPU based procedure for each of the block groups. For both block groups, the entire PUMS data is used as seed. Each block group marginals are used as controls in estimating weights for transition frequencies. Two transition matrices are prepared for two block groups in this case.

The proposed HMM model framework was implemented using a Python package named "*hmmlearn"* (*36*). This package allows the generation of as many households as needed. Households along with associated persons were generated in form of attribute sequences. Then the attribute sequences were processed using a decoding program to obtain the attribute set. For each case, 5 simulations were run to obtain a representative set of synthetic population.

4.2 Results and Findings

The total numbers of synthetic households and persons for each of the cases are summarized in Table 4.2. These totals reflect the average of the total number of households and persons from 5 simulations. In each case, the total number of synthetic households match perfectly with the observed total number of households for the block groups. Since the household model is placed at the upper level of the proposed hierarchical structure and the drawing unit is also a household, this match is not surprising. However, there are some differences in the total number of synthetic persons. As the person models are executed based on the household size distribution, the number of persons is simulated based on the probability distribution at that level. For block group BG0427002, over-synthesis of persons is observed for the first two cases with significant variation. On the other hand the total number significantly improves in Case 3 with a smaller percent difference of 1.36%. For block group BG2531001, the percent difference of the total number of

persons for each case is comparatively low compared to the other block group. However, when compared between the cases, Case 1 provides better match while Case 3 shows a variation of 4.75%.

Table 4.2 Summary of Synthetic Households and Persons for Three Cases

	BG0427002				BG2531001			
	Marginals	**Case 1**	**Case 2**	**Case 3**	**Marginals**	**Case 1**	**Case 2**	**Case 3**
Total Households	1,043	1,043	1,043	1,043	1,041	1,041	1,041	1,041
Percent Difference (%)	NA	0	0	0	NA	0	0	0
Total Persons	1,913	2,521	2,660	1,887	2,545	2,512	2,470	2,424
Percent Difference (%)	NA	-31.78	- 39.05	1.36	NA	1.30	2.95	4.75

Note: NA = not applicable

In order to understand the fitting of synthetic output with observed aggregate data, the synthetic marginals for each case are compared with corresponding observed marginals for two block groups. Figure 4.1 and 4.2 represent the comparison of marginals for block groups BG0427002 and BG2531001 respectively. In Figure 4.1, the synthetic population in Case 3 match closely with the observed category marginals both at household and person level. Table 4.1 shows that there is no household that has more than 4 persons and no person in age category 2. Synthetic results in Case 3 reflect this information completely, whereas the other two cases are unable to capture this information from aggregate marginals and they generate households and persons with these unavailable categories that is inconsistent with this block group information. Figure 4.2 also shows that the synthetic population in Case 3 fit very well with observed marginal information for this block group except for two attribute categories at the person level. That being said, for both block groups, Case 3 can incorporate the marginal distributions information of that block group to generate more reliable synthetic household and persons. On the other hand, the performances of Case 1 and Case 2 are very poor in matching the observed marginal – this is reasonable because these do not incorporate the marginal distribution information during the synthesis. Case 2 should produce better results than Case 1 as transition frequencies are estimated from corresponding PUMA records which can have more relevant information regarding the block groups. However, the analysis shows that the result is not consistent for all attribute categories and in some cases its

performance is poorer than Case 1. This may be attributed to smaller sample sizes resulting from only using the PUMA specific sample information.

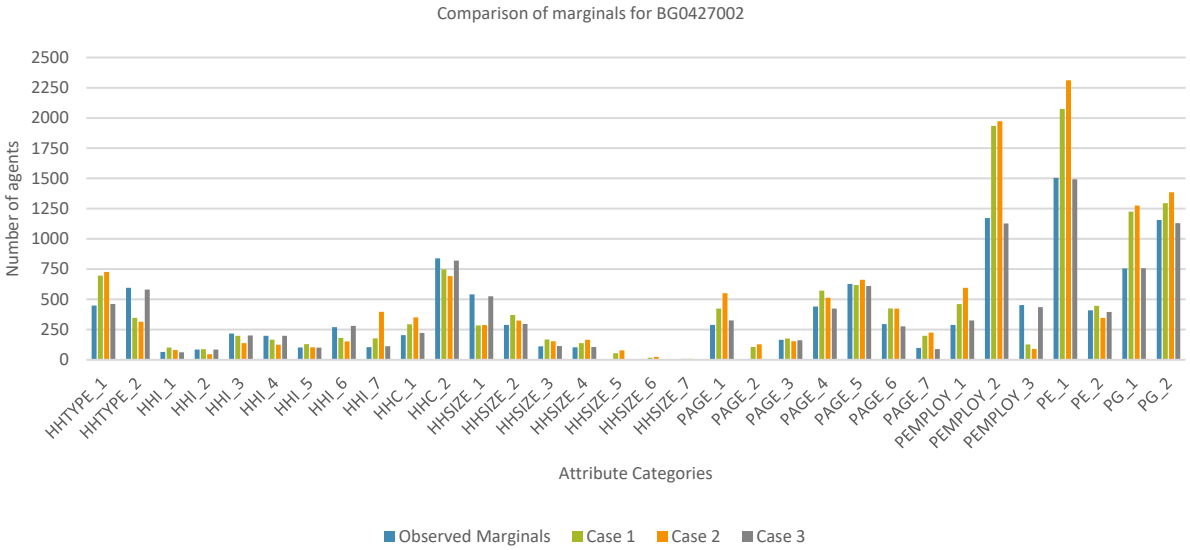

Figure 4.1 Comparison of Marginals for BG0427002

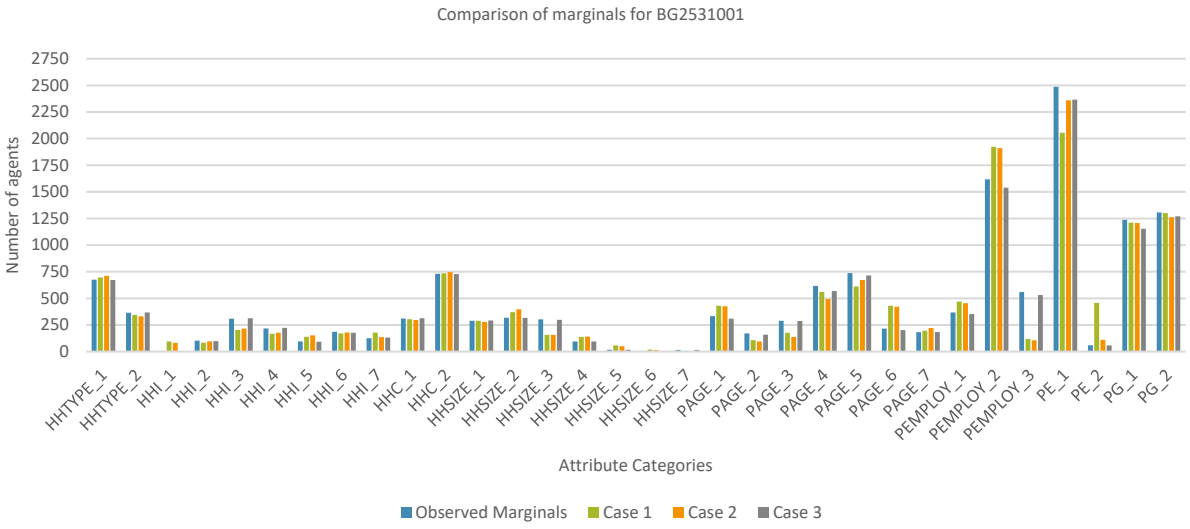

Figure 4.2 Comparison of Marginals for BG2531001

To illustrate the differences in synthesis of each case further, absolute percent difference (APD) is calculated for all attribute categories of two block groups. For block group BG0427002**,** the percent differences are very large for Case 1 and Case 2 compared to Case 3 (Figure 4.3). For Case 3, all attribute categories have APD lower than 12%. The average of APD across categories is about 55.33%, 64.56% and 3.62% for Case 1, Case 2 and Case 3 respectively.

In Figure 4.4, Case 3 also provides very lower APD compared to the other two cases. However, in the case of infrequent attribute categories, it shows comparatively large APD due to the fact that the observed marginals are very low for these categories and a small variation can result in large percent difference. For this block group, the average of APD for all attribute categories is about 42.21%, 40.56% and 3.85% for Case 1, Case 2 and Case 3 respectively. Therefore, it can be noted that in terms of matching individual household and person attribute categories, Case 3 renders less difference compared to Case 1 and Case 2.

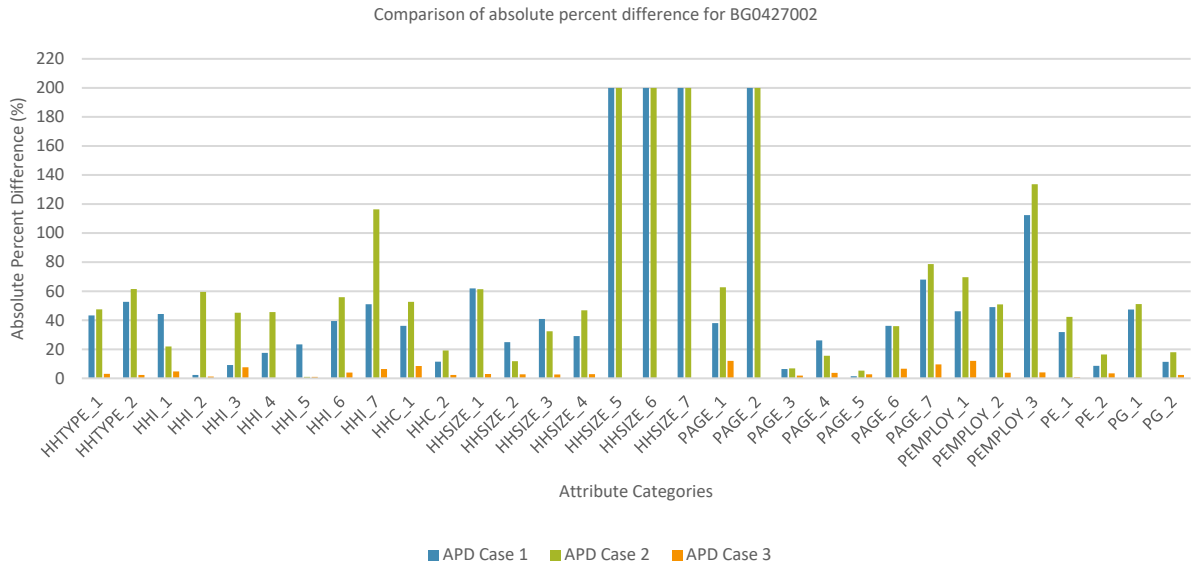

Figure 4.3 Comparison of Absolute Percent Differences for BG0427002

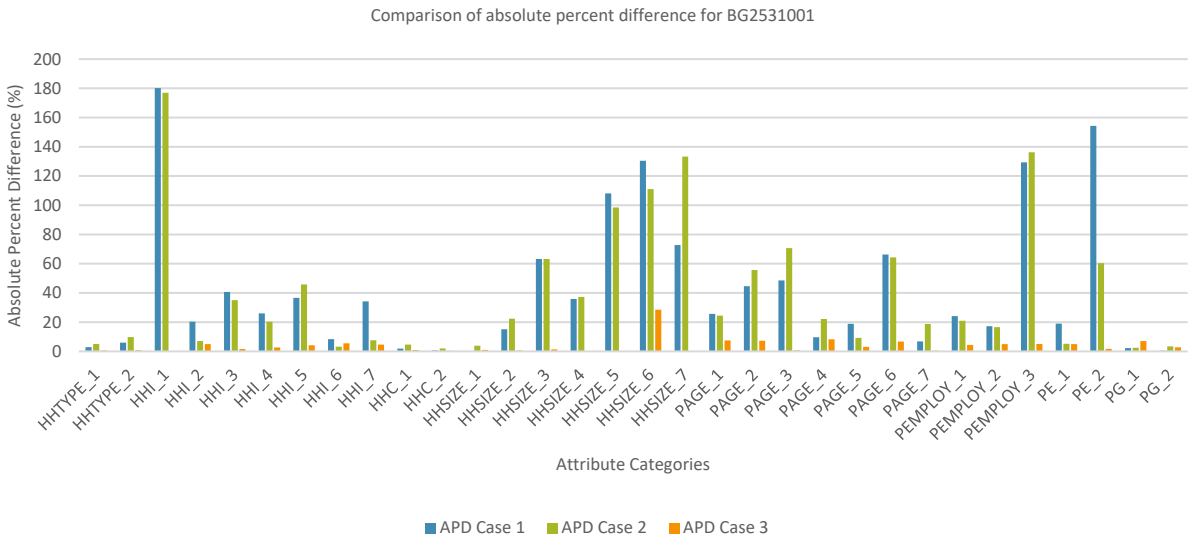

Figure 4.4 Comparison of Absolute Percent Differences for BG2531001

From Figure 4.3 and Figure 4.4, it can be seen that Case 3 performs better than other cases, however, a scatter plot helps better understand the overall fit of both household and person-level attributes synthesized using the proposed HMM framework. Figure 4.5 represents a two-dimensional plot where each observation is a particular household or person attribute categories. For both block groups, the results from Case 3 exhibit a very good fit with the observed category totals with higher R-squared values. The observations obtained from Case 1 and Case 2 show more scattered distribution. This plot helps to explain why aggregate controls are necessary to generate more fitted population in HMM framework.

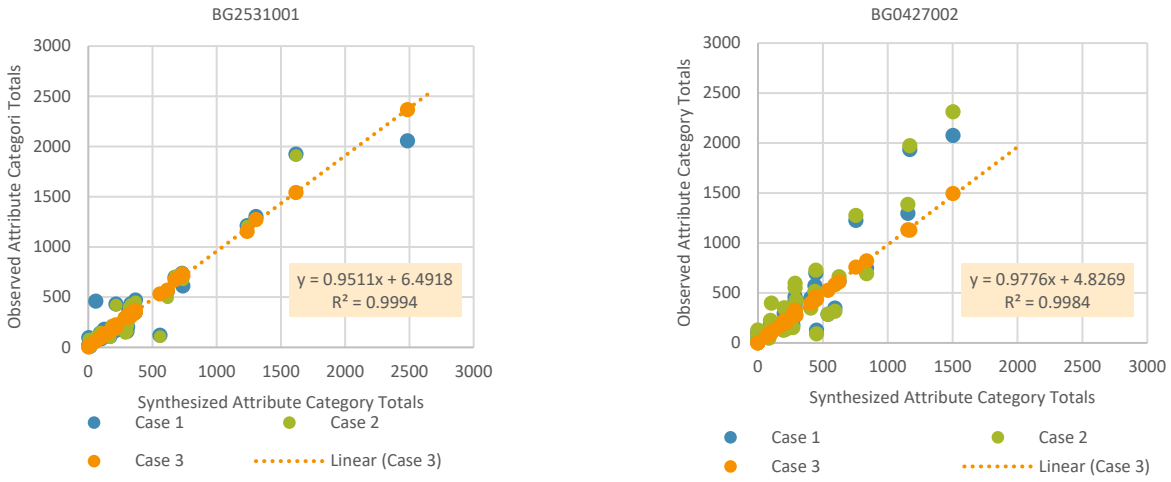

Figure 4.5 Comparison of Synthetic Attributes with Block Group Marginals

Chapter 5. Conclusion and Future Work

In order to apply microsimulation-based models of land use and travel demand, socio-economic and demographic attributes about all individuals in a region is required. This disaggregate level information is not readily available and people resort to population synthesis procedures. These procedures combine readily available information in the form of sample data and marginal distributions to generate the required inputs. With the increasing interest in disaggregate models, a number of synthetic population generators have been developed in the recent past. However, most synthesis techniques replicate the agents from sample data to generate the synthetic population. This leads to issues of lumpiness in the results and an inability to capture the true underlying distribution. Simulation-based synthesis techniques have been developed recently to resolve these issues. More recently HMM-based techniques have been proposed. The model attempts to define the process as comprising of states and achieves to capture the joint distribution of the states and transitions between states. In order to adopt HMM for population synthesis, problem of population generation is cast as a sequence labeling problem. Being a probabilistic procedure, the model can simulate agents' attributes and thus overcomes the issues associated with replication noted above.

In this research, a new HMM-based population synthesis procedure is proposed that provides two main contributions. First, the study developed a hierarchical structure of HMM to generate synthetic household and persons simultaneously. Second, in order to ensure that the synthesized information is consistent with available aggregate information, a new IPU based procedure is proposed to estimate the underlying transition probability matrix.

A case study was presented to demonstrate the feasibility and applicability of the proposed approach. Analysis from a case study confirms that the proposed hierarchical structure of HMM performs very well in generating household and person-level information simultaneously. The transition probability estimation procedure proposed in this study helps to incorporate geography-based information as controls allowing for more reliable synthetic households and persons generation.

There are some limitations of the current work that offer up avenues for future research. First, though the proposed model can generate an exact number of households in each simulation, matching the total number of persons is still an issue that needs to be explored more. In the present configuration of the model, the total number of persons normally shows a lower percent variation across simulations. Further study is needed to figure out a better configuration of the model such that the total number of persons match closely with observed totals. Second, application of the proposed approach is not as straightforward as some of the other synthesis procedures. For implementing this hierarchical structure for a different use case, a comprehensive study is required to understand the correlation between the household and person attributes of interest. Depending on the use-case, a systematic flow of attributes both at household and person-level should be established to build the hierarchical configuration. Third, the proposed model framework is developed and tested using a limited number of variables. However, the dimension of the model will increase exponentially for a larger set of attributes resulting in a large transition matrix with more complex transition patterns. One potential way to deal with the large dimensional model can be the disintegration of the model structure into several modules according to attribute hierarchy and simulation of households and persons sequentially from those modules. Again, further research is required to establish and validate this decomposition of HMM.

References

1. Müller, K., and K. W. Axhausen. Population Synthesis for Microsimulation : State of the Art. *90th Annual Meeting of the Transportation Research Board*, 2011, p. 21. https://doi.org/10.3929/ethz-a-006132973.

2. Tirumalachetty, S., K. M. Kockelman, and B. G. Nichols. Forecasting Greenhouse Gas Emissions from Urban Regions: Microsimulation of Land Use and Transport Patterns in Austin, Texas. *Journal of Transport Geography*, Vol. 33, No. Complete, 2013, pp. 220–229. https://doi.org/10.1016/j.jtrangeo.2013.08.002.

3. Balmer, M., K. W. Axhausen, and K. Nagel. Agent-Based Demand-Modeling Framework for Large-Scale Microsimulations. *Transportation Research Record*, Vol. 1985, No. 1, 2006, pp. 125–134. https://doi.org/10.1177/0361198106198500114.

4. Sun, L., and A. Erath. A Bayesian Network Approach for Population Synthesis. *Transportation Research Part C: Emerging Technologies*, Vol. 61, 2015, pp. 49–62. https://doi.org/10.1016/j.trc.2015.10.010.

5. Hermes, K., and M. Poulsen. A Review of Current Methods to Generate Synthetic Spatial Microdata Using Reweighting and Future Directions. *Computers, Environment and Urban Systems*, Vol. 36, No. 4, 2012, pp. 281–290. https://doi.org/10.1016/j.compenvurbsys.2012.03.005.

6. Anderson, P., B. Farooq, D. Efthymiou, and M. Bierlaire. Associations Generation in Synthetic Population for Transportation Applications: Graph-Theoretic Solution. *Transportation Research Record*, Vol. 2429, No. 1, 2014, pp. 38–50. https://doi.org/10.3141/2429-05.

7. Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd. Simulation Based Population Synthesis. *Transportation Research Part B: Methodological*, Vol. 58, 2013, pp. 243–263. https://doi.org/10.1016/j.trb.2013.09.012.

8. Konduri, K. C., D. You, V. M. Garikapati, and R. M. Pendyala. Enhanced Synthetic Population Generator That Accommodates Control Variables at Multiple Geographic Resolutions. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2563, 2016. https://doi.org/10.3141/2563-08.

9. Saadi, I., A. Mustafa, J. Teller, B. Farooq, and M. Cools. Hidden Markov Model-Based Population Synthesis. *Transportation Research Part B: Methodological*, Vol. 90, 2016, pp. 1–21. https://doi.org/10.1016/j.trb.2016.04.007.

10. Deming, W. E., and F. F. Stephan. On a Least Squares Adjustment of a Sampled Frequency

Table When the Expected Marginal Totals Are Known. *Ann. Math. Statist.*, Vol. 11, No. 4, 1940, pp. 427–444. https://doi.org/10.1214/aoms/1177731829.

11. Beckman, R. J., K. A. Baggerly, and M. D. McKay. Creating Synthetic Baseline Populations. *Transportation Research Part A: Policy and Practice*, Vol. 30, No. 6, 1996, pp. 415–429. https://doi.org/10.1016/0965-8564(96)00004-3.

12. Guo, J., and C. Bhat. Population Synthesis for Microsimulating Travel Behavior. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2014, No. 1, 2007, pp. 92–101. https://doi.org/10.3141/2014-12.

13. Arentze, T., H. Timmermans, and F. Hofman. Creating Synthetic Household Populations: Problems and Approach. *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 2014, No. 1, 2007, pp. 85–91. https://doi.org/10.3141/2014-11.

14. Ye, X., K. C. Konduri, R. M. Pendyala, B. Sana, and P. Waddell. Methodology to Match Distributions of Both Household and Person Attributes in Generation of Synthetic Populations. *Transportation Research Board Annual Meeting 2009*, Vol. 9601, No. 206, 2009, pp. 1–24. https://doi.org/10.1.1.537.723.

15. Pritchard, D. R., and E. J. Miller. Advances in Population Synthesis: Fitting Many Attributes per Agent and Fitting to Household and Person Margins Simultaneously. *Transportation*, Vol. 39, No. 3, 2012, pp. 685–704. https://doi.org/10.1007/s11116-011-9367-4.

16. Mueller, K., and K. W. Axhausen. Hierarchical IPF: Generating a Synthetic Population for Switzerland. 2011.

17. Zhu, Y., and J. Ferreira. Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation. *Transportation Research Record*, Vol. 2429, No. 1, 2014, pp. 168–177. https://doi.org/10.3141/2429-18.

18. David, V., and W. Paul. An Evaluation of the Combinatorial Optimisation Approach to the Creation of Synthetic Microdata. *International Journal of Population Geography*, Vol. 6, No. 5, 2000, pp. 349–366. https://doi.org/10.1002/1099-1220(200009/10)6:5<349::AID-IJPG196>3.0.CO;2-5.

19. Abraham, J. E., K. J. Stefan, and J. D. Hunt. Population Synthesis Using Combinatorial Optimization at Multiple Levels. *Transportation Research Record*, No. 12–3383, 2012.

20. Williamson, P., M. Birkin, and P. H. Rees. The Estimation of Population Microdata by Using Data from Small Area Statistics and Samples of Anonymised Records. *Environment and Planning A*, 1998, pp. 785–816. https://doi.org/10.1068/a300785.

21. Ryan, J., H. Maoh, and P. Kanaroglou. Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms. *Geographical Analysis*, Vol. 41, No. 2, 2009, pp. 181–203. https://doi.org/10.1111/j.1538-4632.2009.00750.x.

22. Williamson, P. An Evaluation of Two Synthetic Small-Area Microdata Simulation Methodologies: Synthetic Reconstruction and Combinatorial Optimisation. In *Spatial Microsimulation: A Reference Guide for Users*, Springer Netherlands, pp. 19–47.

23. Caiola, G., and J. P. Reiter. Random Forests for Generating Partially Synthetic, Categorical Data. *Trans. Data Privacy*, Vol. 3, No. 1, 2010, pp. 27–42.

24. Rabiner, L. R. A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE*, Vol. 77, No. 2, 1989, pp. 257–286. https://doi.org/10.1109/5.18626.

25. Durbin, R., S. Eddy, A. Krogh, and G. Mitchison. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. *Analysis*, Vol. 356, 1998. https://doi.org/10.1017/CBO9780511790492.

26. Eddy, S. R. What Is a Hidden Markov Model? *Nature Biotechnology*, Vol. 22, 2004, p. 1315.

27. Jurafsky, D., and J. H. Martin. Hidden Markov Models. In *Speech and Language Processing*.

28. Warakagoda, N. D. *A Hybrid ANN-HMM ASR System with NN Based Adaptive Preprocessing*. Norwegian Institute of Technology, 1996.

29. Murphy, K. P., and M. A. Paskin. Linear Time Inference in Hierarchical HMMs. *Advances in neural information processing systems 14: proceedings of the 2001 conference*, Vol. 833, 2002.

30. Weiland, M., A. Smaill, and P. Nelson. Learning Musical Pitch Structures with Hierarchical Hidden Markov Models. 2005.

31. Aarno, D., and D. Kragic. Layered HMM for Motion Intention Recognition. 2006.

32. L Scott, S., and I.-H. Hann. *A Nested Hidden Markov Model for Internet Browsing Behavior*. 2018.

33. Fine, S., Y. Singer, and N. Tishby. The Hierarchical Hidden Markov Model: Analysis and Applications. *Machine Learning*, Vol. 32(1), 1998, pp. 41–62. https://doi.org/10.1023/A:1007469218079.

34. Weirstra, D. *A New Implementation of Hierarchical Hidden Markov Models*. Utrecht University, 2004.

35. Saadi, I., B. Farooq, A. Mustafa, J. Teller, and M. Cools. An Efficient Hierarchical Model for Multi-Source Information Fusion. *Expert Systems with Applications*, Vol. 110, 2018, pp. 352–362. https://doi.org/https://doi.org/10.1016/j.eswa.2018.06.018.

36. HMMLEARN. https://github.com/hmmlearn/hmmlearn/blob/master/doc/index.rst. Accessed Jul. 23, 2018.