Alternative Paradigms for Synthesizing Household and Person-level Attributes Jointly

Amit Ranjan Mondal
University of Connecticut - Storrs, amit.mondal@uconn.edu

Recommended Citation
https://opencommons.uconn.edu/gs_theses/1370

This work is brought to you for free and open access by the University of Connecticut Graduate School at OpenCommons@UConn. It has been accepted for inclusion in Master's Theses by an authorized administrator of OpenCommons@UConn. For more information, please contact opencommons@uconn.edu.
Alternative Paradigms for
Synthesizing Household and Person-level Attributes Jointly

Amit Ranjan Mondal

BS, Bangladesh University of Engineering & Technology, 2016

A Thesis
Submitted in Partial Fulfillment of the
Requirements for the Degree of
Master of Science
At the
University of Connecticut
2019
Copyright by

Amit Ranjan Mondal

2019
APPROVAL PAGE

Master of Science Thesis

Alternative Paradigms for
Synthesizing Household and Person-level Attributes Jointly

Presented by
Amit Ranjan Mondal, BS

Major Advisor ______________________________________________
Karthik C. Konduri

Associate Advisor ______________________________________________
Nicholas E. Lownes

Associate Advisor ______________________________________________
Jin Zhu

University of Connecticut
2019
ACKNOWLEDGMENT

I would like to extend my profound gratitude and appreciation to everyone who helped and encouraged me in one way or another in the completion of this research. I am highly indebted to my academic adviser, Dr. Karthik C. Konduri for his invaluable mentorship, generous guidance and consistent support that helped me bring this study into success. I am also grateful to my associate advisers, Dr. Nicholas E. Lownes and Dr. Jin Zhu for their insightful review and valuable suggestions.

I would like to thank my colleagues for their cooperation and encouragement and for being a part of this wonderful stay here at the University of Connecticut. I appreciate the love and support I received from the amazing human beings I met during my graduate studies.

I am thankful to my family and friends for their unconditional love and care. I am grateful to my brother for being an amazing mentor in my life. Above all, I express my heartfelt gratitude to my parents for making me who I am today.
# TABLE OF CONTENTS

ACKNOWLEDGMENTS ........................................................................................................ iv

TABLE OF CONTENTS ........................................................................................................ v

LIST OF FIGURES ............................................................................................................... vii

LIST OF TABLES ................................................................................................................ viii

ABSTRACT ........................................................................................................................ ix

1 Introduction .................................................................................................................. 1

1.1 Background ............................................................................................................. 1

1.2 Literature Review ................................................................................................ 2

1.3 Research Scope ...................................................................................................... 4

2 Study 1: Simulation-based Synthesizing Framework ...................................................... 7

2.1 Introduction ............................................................................................................. 7

2.2 Methodology .......................................................................................................... 8

2.2.1 Overview of HMM Framework .......................................................................... 8

2.2.2 Hierarchical Structure of HMM-based Population Synthesis ......................... 11

2.2.3 Geography-based Transition Probability Estimation ....................................... 15

2.3 Case Study ............................................................................................................. 16

2.3.1 Data Preparation ............................................................................................. 16

2.3.2 Results and Discussions .................................................................................. 18
Study 2: Analytical Procedure for Sample Household Weights Estimation ................. 27

3.1 Introduction ............................................................................................................. 27

3.2 Methodology .......................................................................................................... 28

3.3 Case Study ................................................................................................................ 31

3.3.1 Data Preparation ............................................................................................... 31

3.3.2 Results and Discussions .................................................................................... 32

4 Conclusion ................................................................................................................ 44

REFERENCES ................................................................................................................. 48

APPENDICES .................................................................................................................. 51
LIST OF FIGURES

Figure 2-1: Structure of a simple HMM ........................................................................................................ 9
Figure 2-2: Simple structure of a household model ...................................................................................... 12
Figure 2-3: Connection of person models with household model ................................................................. 13
Figure 2-4: Configuration of a single person (Householder) model .............................................................. 15
Figure 2-5: Comparison of marginals for BG0427002 .............................................................................. 21
Figure 2-6: Comparison of marginals for BG2531001 .............................................................................. 21
Figure 2-7: Comparison of Absolute Percent Differences (APD) for BG0427002 ................................. 22
Figure 2-8: Comparison of Absolute Percent Differences (APD) for BG2531001 ................................. 23
Figure 2-9: Comparison of synthetic attribute totals with block group marginals .............................. 24
Figure 2-10: Weighted absolute percent deviation in person total for 10 additional block groups .......................... 25
Figure 3-1: Fitting of total estimated weights using the univariate and multivariate system ........ 33
Figure 3-2: Distribution of deviations of multivariate weight totals from univariate weight totals .......................................................... 33
Figure 3-3: Deviation of joint category marginals from observed household marginals .......... 34
Figure 3-4: Deviation of joint category marginals from observed person marginals .................. 35
Figure 3-5: Distribution of regression parameters (Univariate Weights VS Multivariate Weights) for 666 block groups .................................................................................................................................. 36
Figure 3-6: Absolute differences in weight totals between HP and PH configuration for both univariate and multivariate system ........................................................................................................ 37
Figure 3-7: Absolute percent differences in weight totals between the multivariate and univariate system for both HP and PH configuration .................................................................................................. 37
Figure 3-8: Comparison of regression parameters (Univariate HP VS Multivariate HP and Univariate PH VS Multivariate PH) ................................................................. 38
Figure 3-9: Absolute Percent Deviation (APD) in synthetic person totals................................. 39
Figure 3-10: Weighted average absolute percent deviation across attribute categories .......... 41
Figure 3-11: Weighted average absolute percent deviation (APD) across block groups .......... 41
Figure 3-12: Comparison of regression parameters (observed attribute totals VS synthetic attribute marginals) for IPU, ULSQ and MLSQ................................................................. 42

LIST OF TABLES

Table 2-1: Summary of synthetic households and persons for three cases.............................. 19
ABSTRACT

In order to apply microsimulation-based models of land use and travel demand, socio-economic and demographic attributes are required for each individual in a region. This disaggregate level information is not readily available and planners resort to population synthesis procedures. This research includes two studies that are focused on developing alternative paradigms for population synthesis and for estimating sample household weights. In the first study, a simulation-based technique for multi-level population synthesis using a Hidden Markov Model (HMM) framework is presented. 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 controls and heterogeneity. As part of the second study, an analytical procedure for estimating sample household weights is proposed that helps estimate consistent weights using disaggregate information with sparse attribute categories by controlling both at the household and person level. Different configurations of the system of linear equations are formulated and evaluated for various sets of block groups as the geographical units. Finally, the synthetic population is generated for ten block groups in Connecticut using the proposed synthesizing framework and weight estimation procedure. The analysis of synthetic outputs confirms that the proposed weight estimation procedure is comparable with the heuristic-based approaches and can be used as an alternate weight estimation routine for simulating more consistent household and person level attributes or drawing households from the sample to obtain a synthetic population that closely match the available aggregate information.
1 INTRODUCTION

1.1 Background

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. Instead, the 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 presented below. A detailed review of synthesizers can be found in (1, 4, 7, 8).
1.2 Literature Review

The techniques within different SPGs can be clustered into two main groups: fitting-based approaches and combinatorial optimization (CO) procedures (9). Fitting-based approaches focus on estimating a multiway distribution of the 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 the observed marginal distributions 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).

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 selecting 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 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 attributes for multiple geographic resolutions (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.

More recently, there has been a third category of SPGs namely simulation-based approaches. The main advantage of these approaches over earlier techniques is the ability to create more diverse synthetic populations. In both fitting-based and CO approaches, records from the sample dataset 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. The simulation-based approaches use a variety of techniques to model the joint distribution of household and person attributes underlying the population. Subsequently, a synthetic population is generated by
simulating draws from the joint distribution to create agents and their attributes. Caiola and Reiter (2010) implemented Random Forest-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 into population synthesis to understand the underlying structure of the 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 a synthetic population (7). Among recent studies, Saadi et al. (2016) developed a new population synthesis technique using a Hidden Markov Model (HMM) (9). In this study, the 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 the synthetic population.

1.3 Research Scope

The primary objective of this research is to develop alternative methodologies for population synthesis and for estimating sample household weights. To this end, the thesis consists of two research studies.

In the first study, a simulation-based population synthesis approach is proposed for generating synthetic households and persons simultaneously maintaining the geography-based consistency and heterogeneity. This first study builds on the work by Saadi et al. (2016) by addressing two important limitations.
1. First, in their study, the synthetic population generation was only limited to persons; households were not generated. The study acknowledged the need for extending the work so that households and persons are both synthesized while also accounting for the available household and person level information. In this study, 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 that helps simulate both households and persons in a consistent manner.

2. Second, Saadi et al. (2016) used only 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 did not incorporate the available marginal information. In other words, their approach ignored information that could potentially be used to enhance the synthetic population. In this study, the transition matrices for the proposed models 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 the available information.

The feasibility and the applicability of the proposed model and the estimation procedure are demonstrated by generating a synthetic population using data from the US Census Bureau for 2 block groups in Connecticut. The synthetic population was generated under a variety of scenarios mimicking the existing simulation-based procedures. Results are compared across scenarios to highlight the contributions of the proposed approach.
The second study focuses on estimating sample household weights using an analytical technique to eliminate the issues generally encountered in the empirically developed fitting-based approaches. The sample household weights are estimated in forms of least square solutions of a system of linear equations formulated using the available sample information and aggregate marginals. In fitting-based approaches where the household and persons are controlled simultaneously, the presence of sparse categories in the multiway table can lead to slackness in the estimated weights resulting in a poorer match at the person level attributes. The proposed procedure is capable of dealing with the sparse disaggregate information and can impose simultaneous control both at the household and person level. The univariate and multivariate configuration of the system of linear equations is proposed and sample household weights are generated for a different set of block groups to evaluate this estimation procedure. Finally, the results of a synthetic population generated using the proposed HMM-based approach and weight estimation method for 10 block groups are analyzed to illustrate its compatibility with heuristic-based weight estimation technique.

The rest of the thesis is structured as follows. Section 2 describes the HMM-based synthesizing approach. 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. A case study is presented with necessary data preparation, model setup, results, and discussion of findings. Section 3 describes the motivation and methodology regarding the proposed sample household weight estimation procedure. A case study is designed to formulate the evaluation process and, the results and findings are presented. Finally, concluding thoughts along with limitations and future extensions are presented in Section 4.
2 STUDY 1: SIMULATION-BASED SYNTHESIZING FRAMEWORK

2.1 Introduction

The attributes of an agent (household) can be characterized as a sequence of characters. Each character in the sequence can be associated with a category for the attributes. Therefore, the length of the sequence is equal to the number of attributes of interest. Hidden Markov Models (HMM) can be used to characterize such a sequence. HMMs are probabilistic models that can be used for any sequence labeling problem (24, 25). These models are very dynamic in the sense that they can conceptualize any complex sequence analysis model using the graphical method (26). In case of population synthesis, HMMs can be implemented to generate a sequence of attributes for a particular household or person based on a transition structure that can be visualized graphically. The transition patterns can be defined based on the attributes of interest and their intra-level and inter-level relations and then available information can be used to estimate the transition probabilities. However, mapping the appropriate dependency of intra-level and inter-level attributes is necessary to generate consistent information about a household and the persons belonging to that household. In this study, a hierarchical structure is proposed to build the interconnection between the household and person level HMMs. A transition probability estimation technique is also proposed as part of this study to incorporate the geography-based aggregate information in the training stage for more accurate probability distributions. The following sections describe the general aspects of HMM and how household and person level HMMs can be defined and connected to obtain a complete hierarchical structure. Then a case study is demonstrated to illustrate the contribution of the study in generating a synthetic population that is consistent with the corresponding geographical information.
2.2 Methodology

2.2.1 Overview of HMM Framework

In order to explain the functional aspects of HMM, a simple toy example is presented in Figure 2-1. Let us 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 as shown in Figure 2-1. 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, 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. For example, a transition from state A to state H is probably not supported by data. On the other hand, the transition from state C to state B is inconsistent. In HMM, including dummy states can help join different parts of the model without disturbing the actual transition patterns. Dummy states do not have any technical implication in the model; however, in a sequence generation problem, they are very useful for the proper identification of different blocks of the model. The states that emit a symbol or character to indicate an attribute category in the sequence 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 $A = \{A_1, A_2, A_3, ..., A_N\}$ where $N$ is the total count of states. Each state is associated with a probability distribution that regulates the transition to other possible states (27). 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 probability of 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, $\beta_k = \{\beta_{1k}, \beta_{2k}, \beta_{3k}, ..., \beta_{mk}\}$ 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 $\pi$ is $N$ dimensional vector. In addition to the above parameters, HMM also incorporates some logical and consistency constraints as indicated in Equation 2-1 to Equation 2-5.
\[ 0 \leq P(t_{ij}) \leq 1, \quad 1 \leq (i, j) \leq N \] \hspace{1cm} (2-1)

\[ \sum_{j=1}^{N} P(t_{ij}) = 1, \quad 1 \leq i \leq N \] \hspace{1cm} (2-2)

\[ 0 \leq P(\beta_{mk}) \leq 1, \quad 1 \leq m \leq M, \quad 1 \leq k \leq N \] \hspace{1cm} (2-3)

\[ \sum_{m=1}^{M} P(\beta_{mk}) = 1, \quad 1 \leq k \leq N \] \hspace{1cm} (2-4)

\[ \sum_{i=1}^{N} P(i) = 1 \] \hspace{1cm} (2-5)

As noted above, an external observer can only see the outcome corresponding to a state since 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) \((28–31)\). Among these variants, HHMM is of 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 complete 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 the 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 be embedded in a hierarchical fashion within a household model which again can be descendent of another root state. This approach allows for ensuring dependencies between person attributes and household attributes. Building and training an HHMM is computationally very expensive (32). HHMMs can be converted to its equivalent flat HMM without compromising the structural integrity of the model (29, 33). An HHMM that has shared transition structure can be converted to flat HMM by duplicating the sub-models. Though the flattening process introduces computational tractability, it comes at the expense of an increase in the dimension of the HMM (34).

In the next subsection, the proposed approach for implementing multi-level population synthesis using HHMM intuition and HMM equivalency is presented. Further, since the states in the models are configured to represent different attribute categories of agents in a direct manner, each state is associated with only one outcome in the population synthesis case. In other words, the states are not technically hidden in the HMM models in this case. Therefore, this can be conceptualized as the Markov Model variant of 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.

2.2.2 Hierarchical Structure of HMM-based Population Synthesis

The first objective of this study is 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 structure. In HMM, states are considered as attribute categories for both household and person models. Therefore, the number of states in both household and person models are completely dependent on the size of attribute sets and their respective categories. Further, key household attributes are used to generate root states. Then the person models consistent with the defined root states are embedded to build the hierarchical structure. The procedure for building the household model is the same as Saadi et al. (2016). Each household attribute and their respective categories serve as active states in the household model. As noted earlier, a hierarchical tree structure is used to build the household-level HMM model and then to incorporate the person-level HMM models within the household model. Subsequently, this allows the 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 2-2.
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 2-2. In a similar fashion, it is possible to accommodate other household attributes such as persons with age under 18 years (UNDER18). In that case, the states of UNDER18 have separate branches originating from each of the SIZE states. The number of states for attribute UNDER18 is governed by the originating SIZE state. For example, for a household with three persons, there will be a 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: 0 UNDER18, 1 UNDER18, and 2 UNDER18. Now by defining the states of UNDER18 as a root states, person models for these household compositions can be embedded based on the hierarchy. For example, in the state 0 UNDER18, a model representing householder, a model for the second person, and a model for the third person are embedded. On the other hand, for the state 2 UNDER18, a householder model is embedded one time and the person model for those under 18 is embedded two times. Figure 2-3 illustrates the fully embedded SIZE-3 branch based on the household composition.

![Figure 2-3: Connection of person models with household model](image-url)
The proposed HMM structure allows the generation of household attributes in the upper level of the model and then proceeds towards the lower level to generate person level attributes. This model can also be configured to deal with open-ended categories. Choice/decision states can be placed in the model structure to decide the next transition to an embedded HMM subjected to some constraints. 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 can have root and decision states, and the concept of guaranteed pass helps build inter-person connections. This is essential to deal with inconsistent inter-personal relationship during simulation. Figure 2-4 illustrates an example of introducing decision states to deal with gender issues while simulating the householder and second person for family households. This generates consistent gender information of the second person based on householder gender information in a family household using a conditional probability distribution. The total number of states in the transition matrix is, therefore, governed by the size of household and person models and the logical way of connecting those individual models in a hierarchical fashion. It should be noted that expert knowledge and heuristics are applied to determine the interdependencies of attributes. Depending on the use case and how the hierarchy in agent’s attributes is addressed, a completely different configuration of the models can be defined to synthesize the required information. The complete hierarchical structure of the proposed HMM-based synthesis model is included in Figure A-1 as an appendix.
2.2.3 Geography-based Transition Probability Estimation

The second objective of this study is 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 the population that agrees 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 by Saadi et al (2016). The 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 does not consider accounting for both household and person marginals simultaneously. This study proposes a new procedure for estimating the transition counts using both aggregate and disaggregate information. In the proposed approach, the weights for sample households are first estimated using Iterative Proportional Updating algorithm proposed by Ye et al. (2009) controlling all available household and person level marginals for a geographic unit. The details of this heuristic-based weighting procedure are outlined in the study by Ye et al. (2009). These weights are then summed up based on the corresponding household and person attribute combinations and used directly as the exact transition frequencies for that particular geographical unit. The transition probabilities thus generated conform to the available marginal distributions. The synthetic population also accounts for this information and fewer deviations are observed with respect to available marginals.

2.3 Case Study

2.3.1 Data Preparation

A case study was conducted to demonstrate the proposed HMM population synthesis framework and the associated transition probability estimation routine. The study considers 4 household attributes (household type, household income, the 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 the US Census Bureau. The aggregate data is processed from the 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 the PUMA level. Household and person attributes are defined as categorical variables. The description of attributes and a summary of aggregate marginals for two block groups are listed in Table A-1 included in the Appendices section.

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. Based on the household size, categories for the presence of persons under 18 years form the second set of branches. These categories are also set as root states to embed person models. The order of attributes in person models is age, employment, ethnicity, and gender. The hierarchical structure contains a total of 4203 states including active, dummy and decision states. Therefore, the dimension of the transition matrix considered in this case study is 4203 by 4203. The transition probabilities are estimated using three different approaches to highlight the feasibility and applicability of the proposed sample household weight-based estimation approach.

1. **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).

2. **Case 2:** For each block group, the transition matrix is estimated using the sample records of only those households that belong to the 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.

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

For transition probability estimation, the sample household weights were calculated using IPU algorithm embedded in an open source application called PopGen (36). The proposed HMM model framework was implemented using a Python package named “hmmlearn” (37) that allows the generation of as many households as needed. Households and associated persons were generated in the 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. For every 1,000 households, each simulation took about 10 minutes to generate the sequences and about 30 seconds to process those in a test system powered by Intel Core-i7 processor with 12GB RAM.

2.3.2 Results and Discussions

The total numbers of synthetic households and persons for each of the cases are summarized in Table 2-1. In each case, the total number of synthetic households matches 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%. Another interesting point in Table 2-1 is that both block groups have almost the same number of households, but their person totals are quite different. Case 1 generates nearly equal number of persons in both block groups, because the transition probabilities come from a single transition probability matrix. Surprisingly, the person totals in BG2531001 is very close to the simulated totals resulting in a difference of 1.3%. The result may be counter-intuitive in a sense that Case 1 provides a better match than Case 3 which shows a variation of 4.75%. However, this block group may be a very average one leading to an overall better result in the first two cases. It is evident that the percent difference of the total number of persons for each case is comparatively low compared to the other block group.

Table 2-1: Summary of synthetic households and persons for three cases

<table>
<thead>
<tr>
<th></th>
<th>BG0427002</th>
<th></th>
<th>BG2531001</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Marginals</td>
<td>Case 1</td>
<td>Case 2</td>
<td>Case 3</td>
</tr>
<tr>
<td>Total Households</td>
<td>1,043</td>
<td>1,043</td>
<td>1,043</td>
<td>1,043</td>
</tr>
<tr>
<td>Percent Difference (%)</td>
<td>NA</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Total Persons</td>
<td>1,913</td>
<td>2,521</td>
<td>2,660</td>
<td>1,887</td>
</tr>
<tr>
<td>Percent Difference (%)</td>
<td>NA</td>
<td>-31.78</td>
<td>-39.05</td>
<td>1.36</td>
</tr>
</tbody>
</table>

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 2-5 and Figure 2-6 represent the comparison of marginals for block groups BG0427002 and BG2531001, respectively. In Figure 2-5, the synthetic population in Case 3 match closely with the observed category marginals both at the household and person level. Table A-1 (included as an appendix) 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 2-6 also shows that the synthetic population in Case 3 fits 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 a limited number of records in the corresponding PUMA samples.
To further illustrate the differences in the synthesis of each case, the absolute percent difference (APD) is calculated for all attribute categories of the two block groups. For block group BG0427002, the percent differences are very large for Case 1 and Case 2 compared to Case 3 in
Figure 2-7. 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 2-8, 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 since the observed marginals are very low for these categories and a small variation can result in a large percent difference. Nonetheless, the overall assessment can be made based on the average of APD for all attribute categories. The average APD is about 42.21%, 40.56% and 3.85% for Case 1, Case 2 and Case 3 respectively. This analysis shows that Case 3 offers a good improvement over the other two cases. In terms of matching individual household and person attribute categories, Case 3 renders less difference compared to Case 1 and Case 2.

![Comparison of absolute percent difference for BG0427002](image-url)
From Figure 2-7 and Figure 2-8, it can be seen that Case 3 performs better than other cases. A scatter plot helps better understand the overall fit of both household and person-level attributes that are synthesized using the proposed HMM framework. Figure 2-9 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 $R^2$ values. The observations obtained from Case 1 and Case 2 show comparatively scattered distribution. This plot helps to explain why aggregate controls are necessary to generate more fitted population in the HMM framework.
Since it is evident from the results presented so far that Case 3 is the best among other cases of estimating the transition probabilities, the further exploration is warranted to explore the results of Case 3 using more geographical units to incorporate the variations and heterogeneity in the household and person attributes. Therefore, the synthetic population is generated for 10 random block groups using the estimated transition frequencies derived from IPU-based sample household weights and the detailed results are included in the Appendices sections. As illustrated in Figure 2-10, the percent deviations in person totals in most block groups are less than 7% which is acceptable. However, there is still scope for improvement in matching the total number of synthetic persons with the observed marginals. The synthetic attribute totals are also compared to the observed marginals for each of the block group in Figure A-2 added as an appendix and a perfect $y \approx x$ relationship is observed in each case indicating that the synthetic outputs are equivalent to observed marginals and are within an acceptable range.
In order to inspect the fit at a more disaggregate level, deviations from observed marginals are calculated for both household and person-level attribute categories and are presented in Figure A-3 and Figure A-4 included as appendices. The deviations are considerably higher at the person-level attribute categories compared to those at the household level categories. This can be attributed to the heuristic-based IPU weights where the higher degree of sparsity in seed matrix may result in laxity in matching the person level attributes more closely. Therefore, the research addresses the need for developing a sample household weight estimation procedure that is based on systematic technique and is more robust in handling a higher degree of sparsity without compromising the fitness at household and person level. This motivation leads to the second study in this research that involves estimating the sample household weights using an analytical procedure.

The next section proposes the analytical procedure that is based on least square solutions of a system of linear equations formed using the available aggregate and disaggregate information to estimate household weights through an iterative method by controlling the household and person-
level attributes simultaneously. This method is also analogous to IPU-based weight estimation technique in a sense that the weights are constrained by the available geography-based information and hence is capable of capturing the diversity and heterogeneity in household compositions and individual attributes. The following section also describes the methodology of the proposed weight estimation procedure and presents a case study to evaluate different aspects of the estimated weights for various configuration of the system.
3 STUDY 2: ANALYTICAL PROCEDURE FOR SAMPLE HOUSEHOLD WEIGHTS ESTIMATION

3.1 Introduction

In population synthesis domain, Iterative Proportional Fitting (IPF) is one of the most popular algorithms to obtain the joint distributions of household and person level attribute categories by fitting a contingency table at the corresponding level. The conventional implementation of this algorithm involves adjusting the household and person level attributes with the observed control totals independently. This results in two separate sets of estimated weights. IPF does not allow for simultaneous weight adjustment at multiple levels to come up with a final set of household weights. Therefore, in an IPF-based population synthesis framework, inconsistent and inaccurate distributions of person-level attributes are observed when household weights obtained by adjusting only the household level attributes are used for random drawing from the sample.

In an effort to address this limitation, Iterative Proportional Updating (IPU) was developed. IPU addresses this limitation using a heuristic iterative method that adjusts the household weights such that both household and person level marginal distributions are simultaneously satisfied. In other words, the iterative weight adjustment mechanism in IPU ensures better match at person-level without affecting the fitness at household-level (14). The use of these adjusted weights results in synthetic households and persons that are more representative. Due to the heuristic nature of IPU there are known limitations (e.g. The performance of IPU in matching the person-level attributes may not be as good as it is expected when the number of sparse categories in the household and person-level attributes may provide less flexibility in matching the person-level control totals (14)) and unanswered questions (e.g. the analytical underpinnings are yet to be established).
The primary objective of this study was to explore an analytical procedure for estimating weights that are consistent in matching household and person-level attributes simultaneously. This procedure while ensuring that the method has established analytical properties, is able to deal with sparse categories. Since the sample household weights estimated using the proposed procedure are controlled both at the household and person level, these estimated weights can be used in any fitting-based approaches to obtain synthetic households and persons that are consistent with the corresponding observed marginals. Furthermore, this analytical method is also applicable to the proposed HMM-based population synthesis framework to obtain appropriate frequencies in the transition matrix that are controlled by the observed geographical constraints.

### 3.2 Methodology

The proposed analytical procedure for sample household weights estimation is designed as a linear system of equations in the form of $Ax = b$, where $A$ is the contribution matrix containing the information about household and person level attribute categories, $b$ is the constraints array corresponding to the household and person level attribute category, and $x$ is the solution set of the linear system that represents the estimated household weights. The dimension of the contribution matrix $A$ is $(J + K) \times H$, where $J$ is the total number of household level categories, $K$ is the total number of person level categories and $H$ is the total number of households contained in the sample dataset.

The contribution matrix $A$ is formed using the information obtained from the disaggregate sample dataset. Each element of the matrix represents the frequency of household or persons that belongs to the corresponding sample household and the relevant household or person level attribute
categories. The constraints array $b$ is a vector of length $(J + K)$ that contains the observed marginals extracted from the aggregate information of the geographical unit for which the sample household weights are to be estimated. The solution set $x$ is an array having a dimension of $H \times 1$ containing the estimated household weights for a specific geographical unit. The linear system can be configured in two ways using the univariate and multivariate constraints. The univariate linear system is formulated directly using the available aggregate totals and sample information. In case of a multivariate linear system, the multiway frequency of joint distribution of household and person level attribute categories are estimated through an iterative procedure by fitting a contingency table containing the attribute categories of interests.

In a univariate linear system, two different matrices $A^H_u$ and $A^P_u$ are extracted from the sample information that contain the univariate frequencies of household and person level attribute categories respectively. $A^H_u = \{a_{jh}\}$ and $A^P_u = \{a_{kh}\}$ where, $j \in J_u$, $k \in K_u$ and $h \in H$. These two individual matrices are stacked vertically to construct the final univariate contribution matrix $A_u$ of dimension $(J_u + K_u) \times H$. $C^H = \{c_j\}$ and $C^P = \{c_k\}$ are the constraint arrays for univariate household and person level attribute categories that are also stacked vertically to form the required constraints array. Therefore, for each $j$ and $k$, the linear system corresponding to household and persons respectively are constructed using the following equations (Equation 3-1 and 3-2) and then are stacked vertically to obtain the final linear system in the form $Ax = b$ that helps estimate the household weights more consistently addressing the simultaneous control at both household and person level in the estimation routine.

$$\sum_{h=1}^{H} a_{jh}x_h = c_j \quad \text{(3-1)}$$

$$\sum_{h=1}^{H} a_{kh}x_h = c_k \quad \text{(3-2)}$$
In case of a multivariate system, the formulation technique is the same as univariate system except for the fact that each row in the household or person level contribution matrix represents a joint category that is essentially obtained from fitting a contingency table constructed using the univariate attribute categories. Therefore, the household and person level contribution matrices for the multivariate system are represented as \( A_m^H = \{a_{jh}\} \) and \( A_m^P = \{a_{kh}\} \) where, \( j \in J_m, k \in K_m \) and \( h \in H \). Since, \( J_m \) and \( K_m \) contain more elements than \( J_u \) and \( K_u \) respectively, the final contribution matrix \( A_m \) in the multivariate system has more rows compared to the univariate contribution matrix \( A_u \).

Generally, the system of linear equations in case of sample household weight estimation is an underdetermined system, where \( H > (J_u + K_u) \) or \( H > (J_m + K_m) \) and the estimated weights should be non-negative and non-zero \((0 < x < \infty)\) to be consistent with the use case of population synthesis. Additionally, the contribution matrix in both univariate and multivariate configuration exhibits a higher degree of sparsity in the underlying structure. Therefore, the proposed procedure solves the system of linear equations using LSMR, an iterative algorithm for sparse least square problems that deals with the underdetermined system as well (38). LSMR implements the Golub-Kahan bidiagonalization technique (39) and monotonically reduces the value of \( \|A^T r_i\| \), where \( r_i \) is the residual \((r_i = b - Ax_i)\) in \( i^{th} \) iteration (38). Since, the solutions of the system of linear equations are lower-bounded minimization problem, Trust Region Reflective method is implemented within LSMR to efficiently explore the entire variable space and improve the convergence (40).
3.3 Case Study

3.3.1 Data Preparation

The proposed analytical procedure for sample household weight estimation is demonstrated in a case study where the sample household weights were estimated for all 666 block groups in Hartford County, Connecticut. The household and person-level attribute categories considered for this case study are the same as the previous study regarding HMM-based population synthesis. PUMS datasets with records of 70,221 households and 181,082 persons are used to obtain disaggregate information and construct the required contribution matrices. The attribute constraints are extracted from ACS (2010-2014) Summary datasets. The sample household weights are estimated using both univariate and multivariate configuration of the system of linear equations. In case of univariate system, the contribution matrix and constraints array are constructed directly using the attributes categories (18 household attribute categories and 14 person attribute categories) derived from the available disaggregate and aggregate information respectively which results in a contribution matrix with dimension \((18 + 14) \times 70,221\) and constraint array with dimension \((18 + 14) \times 1\). In terms of the multivariate system, the joint attribute categories are obtained from IPF using the same disaggregate and aggregate information which leads to 146 household-level joint attribute categories and 56 person-level joint attribute categories. Therefore, the multivariate contribution matrix had a shape of \((146 + 56) \times 70,221\) with corresponding joint constraint array which has a dimension of \((146 + 56) \times 1\). For this 666 block groups, the household-level equations are placed before the person-level equations in the system of linear equation for both univariate and multivariate system which can be denoted as HP configuration of the system.
In order to examine the impact of ordering the household and person-level linear equations on the weight estimation, a set of 23 block groups is selected randomly where the person-level equations are placed at the upper level, whereas the household-level equations are stored below the sets of person-level equations. This configuration is adopted for the univariate and multivariate system as well and can be denoted as PH configuration of the system of linear equations. The results of the PH configuration are compared with the results of the HP configuration for the same set of block groups to demonstrate the findings.

The multivariate system for each block group required about 1.2 hours to solve, whereas each of the univariate systems was solved in about 4.5 minutes in a computing system configured with Intel Xeon E5 processor with 2.8 GHz and 128GB RAM.

Finally, the performance of the proposed analytical procedure is evaluated by generating synthetic households and persons using the proposed HMM-based population synthesis framework. The results are compared to the synthetic population generated with HMM-based framework using IPU-based sample household weights to illustrate its compatibility with the proposed simulation-based population synthesis technique.

### 3.3.2 Results and Discussions

The sum of estimated weights for all 666 block groups using both univariate and multivariate systems are plotted in a two-dimensional space in Figure 3-1 to examine the fit between these two configurations. The results from simple linear regression (Slope = 1.00713, Intercept = 4.21691, \( R^2 = 0.98151 \)) indicate that these two configurations generated household weights that are very consistent with each other. In 564 block groups (85% of the selected block groups), the sums of estimated weights using the multivariate system are more than that of using the univariate system. Figure 3-2 shows the distribution of deviations of multivariate weight totals from univariate weight...
totals, where the deviation range is selected to be [-25, 25]. The plot includes deviations for 640 block groups (96% of the selected block groups). It is also evident from the distribution that most block groups exhibit a very smaller deviation ranging from -0.2 to 5.

Figure 3-1: Fitting of total estimated weights using the univariate and multivariate system

Figure 3-2: Distribution of deviations of multivariate weight totals from univariate weight totals
The deviation of multivariate weight totals from univariate weight totals can be attributed to the joint categorical constraints that are not consistent with the observed univariate marginals. To further explore this issue, 25 block groups are randomly selected and their joint distributions using IPF are compared to the actual aggregate marginals. Figure 3-3 and Figure 3-4 show the deviation of joint category marginals obtained from IPF for multivariate systems from the observed marginals for household and person-level attributes, respectively. There are noticeable deviations in some particular household and person level attributes. It is also evident that deviations at person level are higher in magnitude compared to that of at household level on account of the caveat of conventional IPF where person-level attributes are loosely controlled compared to household-level attributes. There may be some cases where these deviations are much higher leading to this issue where the sum of estimated weights using multivariate constraints becomes higher than that of using univariate constraints.

![Figure 3-3: Deviation of joint category marginals from observed household marginals](image)
In order to observe the fitting of the univariate and multivariate system at disaggregate level, simple linear regression (Univariate VS Multivariate) is applied to each of the block groups where each observation is an estimated weight for a particular household. The distributions of estimated parameters are shown in Figure 3-5. Reasonable fitness is observed for most block groups with slopes and intercepts ranging mostly from 0.55 to 0.80 and from 0.0015 to 0.0035, respectively (Figure 3-5(a) and Figure 3-5(b)). The values of correlation coefficient and standard error of estimates are acceptable as well ranging mostly from 0.70 to 0.85 and from 0.00167 to 0.0025 (Figure 3-5(c) and Figure 3-5(d)). The results also indicate that multivariate weights are larger than the corresponding univariate weights resulting in slopes less than 1. However, for most block groups univariate weights and multivariate weights for all households in the sample dataset are equivalent and positively correlated.

Figure 3-4: Deviation of joint category marginals from observed person marginals
Figure 3.5: Distribution of regression parameters (Univariate Weights VS Multivariate Weights) for 666 block groups

The reordering of household and person-level equations in both univariate and multivariate system is a good exploration to determine if there are any significant differences in the estimated weights for both HP and PH configuration of defining the contribution matrices. Figure 3-6 shows the absolute differences in the sum of weights between HP and PH configuration of contribution matrices for both univariate and multivariate system. It is evident that changing the order from HP to PH results in noticeable differences in the multivariate system compared to the univariate system, however, the level of differences is very less compared to their actual sum of weights. Additionally, for both HP and PH configuration, the absolute percent differences in the sum of weights between univariate and multivariate system are comparable for almost all the block groups considered for this exploration as illustrated in Figure 3-7. In other words, changing the order of household and person level equations does not necessarily improve the differences in the sum of weights between univariate and multivariate system.
To explore at a disaggregate level, a simple linear regression is applied between two pairs of configurations: univariate HP VS multivariate HP, and univariate PH VS multivariate PH where each observation denotes a weight associated with a particular sample household. The comparison of regression parameters obtained from these two regressions is presented in Figure 3-8. For each of the parameters considered for this comparison, a $y \approx x$ relationship is observed with the values distributed in an acceptable range. Therefore, it can be concluded that the order of household and
person-level equations does not have any significant impact on the estimation procedure of sample household weights at a disaggregate level.

![Comparison of regression parameters (Univariate HP VS Multivariate HP and Univariate PH VS Multivariate PH)](image)

**Figure 3-8: Comparison of regression parameters (Univariate HP VS Multivariate HP and Univariate PH VS Multivariate PH)**

Finally, the sample household weights estimated using the proposed analytical procedure are compared with heuristic-based IPU weights by implementing in a practical use case to generate household and person-level attributes for additional 10 block group considered in the extended analysis in Study 1. The proposed HMM-based population synthesis framework is used to generate synthetic household and persons. Both the univariate and multivariate weights are used to estimate the transition probability matrices following the procedure described in Section 2.2.3. The synthesis outputs using the least square-based weights (Univariate Least Square and Multivariate Least Square) are compared side by side with those using IPU-based weights to better understand
the similarities and differences. Therefore, the comparisons are essentially for the three weight estimation procedures (ULSQ, MLSQ, IPU) and are denoted as such in the comparison and result analysis.

![Absolute Percent Deviation (APD) in Synthetic Person Totals](image)

*Figure 3-9: Absolute Percent Deviation (APD) in synthetic person totals*

As illustrated in Figure 3-9, the absolute percent deviations (APD) in total synthetic persons are below 4% in most block groups when univariate and multivariate weights are used instead of IPU-based sample household weights to define the estimated transition frequencies in HMM models. There are some cases as well where univariate weights performed better to generate a more accurate number of persons in the synthetic population compared to multivariate least square weights and IPU-based weights. Overall, the weights estimated using the proposed analytical
procedure are more consistent in getting more accurate person totals compared to the heuristic-based IPU approach.

The weighted average APD calculated across the attribute categories for univariate and multivariate LSQ are mostly below 6% for most block groups and they are also comparable to IPU for the fact that the average APD in most cases is very low as shown in Figure 3-10. The weighted average APDs across the block groups for each household and person attribute categories are shown in Figure 3-11. For most attribute categories, the average APD for all three methods are below 10% except for the categories where the observed numbers of households or persons belonging to those categories are very low resulting in larger percent deviations (i.e. HHSIZE_6, HHSIZE_7, HHI_1, HH_2 etc.). In person level categories, the ULSQ shows fewer deviations in most person attribute categories compared to the other two methods leading to the evidence that univariate weights are more reliable in matching person level attributes. For household level categories, there is no consistent pattern to compare the deviations among the three methods. However, the lower deviations in household level attribute categories ensure the compatibility of all three methods, and IPU seems to perform comparatively better than the other two methods.
Figure 3-10: Weighted average absolute percent deviation across attribute categories

Figure 3-11: Weighted average absolute percent deviation (APD) across block groups
For each block group, the observed attribute totals are compared with synthetic attribute marginals to examine the $y \approx x$ relationship among the attribute categories. The ranges of regression parameters for all three methods are compared in Figure 3-12. In terms of slopes and intercepts, the ranges for the three methods are acceptable though IPU shows a relatively shorter range. The 25th percentile and 75th percentile values of correlation coefficient and standard error indicate that MLSQ and ULSQ have a condensed range close to the desired values compared to IPU. However, the magnitudes of the parameters ensure that all three methods are equivalent and acceptable in creating synthetic households or persons that closely match the observed marginals for that particular geographical unit.

*Figure 3-12: Comparison of regression parameters (observed attribute totals VS synthetic attribute marginals) for IPU, ULSQ and MLSQ*
Based on the results of this study, it can be seen that the proposed least square-based weight estimation procedure can be a feasible alternative to heuristic-based IPU algorithms for sample household weight estimation. The univariate and multivariate systems are both equivalent to each other with a reasonable range of variability. The sum of multivariate weights can be larger than the sum of univariate weights due to the slackness in the calculating joint attribute marginals for constraints used in the system of linear equations which may result in inconsistent synthetic population when the deviation is very high. The computation time is also higher than the univariate system because of the larger dimension of contribution matrices. However, the univariate system takes less time to solve and is also consistent with the available aggregate information as the constraints in the system of linear equations are directly obtained from the readily available information without the need for any intermediate estimation process. The analysis also confirms that the reordering of household and person-level equations does not impact the estimation routine significantly. The practical implementation indicates that the proposed univariate and multivariate weight estimation techniques can perform better in matching the synthetic person totals with observed marginals. The proposed techniques are also comparable to IPU at a disaggregate level and the univariate system is more reliable in matching the person level attributes along with household level variables. The performance of multivariate system is as good as univariate system. However, special caution is warranted to inspect the appropriateness in estimating the joint category constraints.
4 CONCLUSION

In order to apply microsimulation-based models of land use and travel demand, socio-economic and demographic attributes are required for each individual in a region. 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, the 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 sample household weights-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 concurrently. The transition probability estimation procedure proposed in this study helps incorporate geography-based information as controls allowing for more reliable synthetic household and person generation.

There are some limitations of this study 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, the 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. Furthermore, for the same set of attributes, it is possible to come up with various configurations of the models. Therefore, a comprehensive analysis is required to evaluate the performance of each of the configurations. 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. This dimensionality issue is attributed to the general principle of HMM as well as the multilevel structure proposed in this study. 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.

The second study in this research proposes a least square-based procedure to estimate the sample household weights to better match the household and person-level attribute simultaneously. The proposed procedure can deal with sparse disaggregate information and estimate the household weights that are constrained both at the household and person level. The univariate and multivariate approaches of weight estimation are equivalent at the individual geography-level as demonstrated by their overall fitness with each other. Though the multivariate approach performs reasonably well, this procedure is still based on IPF and may lead to some degree of lumpiness due to this intermediate step of estimating joint category constraints. From the analysis of case study perform in this research, it is evident that the proposed approaches are comparable to IPU-based reweighting procedure; however, they perform better in terms of matching the total number of persons synthesized for a particular geographical unit. The results also confirm that univariate weights can be a good alternative to obtain a better match in the household and person-level attributes simultaneously.

The proposed sample household weight estimation procedure accounts for the geographical constraints only and does not consider any regional controls. In other words, the proposed technique is not configured to include any household or person-level attribute that are available at larger geography or regional level. However, it is very important to consider multi-level controls to get a comprehensive synthetic population for a region. Therefore, there is scope for further improvement in defining the system of linear equations by including the regional controls to get sample household weights for different geographical units that are also consistent at the regional
level. Further exploration should be carried out to determine an analytical estimation procedure that imposes this multi-level controls in sample household weight estimation.
REFERENCES


Figure A-1: The complete hierarchical structure showing the connection between upper and lower levels containing the household and person-level HMMs respectively.
Table A-1: Description of Control Variables and Marginals for Two Block Groups

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Description</th>
<th>BG0427002</th>
<th>BG2531001</th>
</tr>
</thead>
<tbody>
<tr>
<td>HHTYPE_1</td>
<td>Family Household</td>
<td>448</td>
<td>676</td>
</tr>
<tr>
<td>HHTYPE_2</td>
<td>Non-family Household</td>
<td>595</td>
<td>365</td>
</tr>
<tr>
<td>HHI_1</td>
<td>Less than 15000USD</td>
<td>65</td>
<td>5</td>
</tr>
<tr>
<td>HHI_2</td>
<td>15000USD - 25000USD</td>
<td>85</td>
<td>103</td>
</tr>
<tr>
<td>HHI_3</td>
<td>25000USD - 50000USD</td>
<td>217</td>
<td>308</td>
</tr>
<tr>
<td>HHI_4</td>
<td>50000USD - 75000USD</td>
<td>199</td>
<td>217</td>
</tr>
<tr>
<td>HHI_5</td>
<td>75000USD - 100000USD</td>
<td>102</td>
<td>96</td>
</tr>
<tr>
<td>HHI_6</td>
<td>100000USD - 150000USD</td>
<td>270</td>
<td>186</td>
</tr>
<tr>
<td>HHI_7</td>
<td>More than 150000USD</td>
<td>105</td>
<td>126</td>
</tr>
<tr>
<td>HHC_1</td>
<td>Presence of persons under 18 years (YES)</td>
<td>204</td>
<td>310</td>
</tr>
<tr>
<td>HHC_2</td>
<td>Presence of persons under 18 years (NO)</td>
<td>839</td>
<td>731</td>
</tr>
<tr>
<td>HHSIZE_1</td>
<td>1 person</td>
<td>541</td>
<td>289</td>
</tr>
<tr>
<td>HHSIZE_2</td>
<td>2 persons</td>
<td>288</td>
<td>318</td>
</tr>
<tr>
<td>HHSIZE_3</td>
<td>3 persons</td>
<td>111</td>
<td>302</td>
</tr>
<tr>
<td>HHSIZE_4</td>
<td>4 persons</td>
<td>103</td>
<td>96</td>
</tr>
<tr>
<td>HHSIZE_5</td>
<td>5 persons</td>
<td>0</td>
<td>17</td>
</tr>
<tr>
<td>HHSIZE_6</td>
<td>6 persons</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>HHSIZE_7</td>
<td>7 persons or more</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>PAGE_1</td>
<td>Less than 14 years</td>
<td>288</td>
<td>333</td>
</tr>
<tr>
<td>PAGE_2</td>
<td>15 years - 17 years</td>
<td>0</td>
<td>170</td>
</tr>
<tr>
<td>PAGE_3</td>
<td>18 years - 24 years</td>
<td>165</td>
<td>289</td>
</tr>
<tr>
<td>PAGE_4</td>
<td>25 years - 44 years</td>
<td>440</td>
<td>617</td>
</tr>
<tr>
<td>PAGE_5</td>
<td>45 years - 59 years</td>
<td>627</td>
<td>737</td>
</tr>
<tr>
<td>PAGE_6</td>
<td>60 years - 74 years</td>
<td>295</td>
<td>216</td>
</tr>
<tr>
<td>PAGE_7</td>
<td>75 years or more</td>
<td>98</td>
<td>183</td>
</tr>
<tr>
<td>PEMPLOY_1</td>
<td>Less than 16 years</td>
<td>288</td>
<td>368</td>
</tr>
<tr>
<td>PEMPLOY_2</td>
<td>Employed for last 12 months</td>
<td>1,172</td>
<td>1,618</td>
</tr>
<tr>
<td>PEMPLOY_3</td>
<td>Unemployed for last 12 months</td>
<td>453</td>
<td>559</td>
</tr>
<tr>
<td>PE_1</td>
<td>Caucasian</td>
<td>1,504</td>
<td>2,486</td>
</tr>
<tr>
<td>PE_2</td>
<td>Others</td>
<td>409</td>
<td>59</td>
</tr>
<tr>
<td>PG_1</td>
<td>Male</td>
<td>756</td>
<td>1,238</td>
</tr>
<tr>
<td>PG_2</td>
<td>Female</td>
<td>1,157</td>
<td>1,307</td>
</tr>
</tbody>
</table>
Figure A-2: Fitting of synthetic attribute totals (Case 3) with observed attribute marginals for 10 random block groups
Figure A-3: Deviations in each attribute categories for the first 5 of 10 block groups
Figure A-4: Deviations in each attribute categories for the last 5 of 10 block groups