Chapter

A Critical Review on Population Synthesis for Activity- and Agent-Based Transportation Models

Ossama E. Ramadan and Virginia P. Sisiopiku

Abstract

Traditional four-step transportation planning models fail to capture novel transportation modes such as car/ridesharing. Hence, agent-based models are replacing those traditional models for their scalability, robustness, and capability of simulating nontraditional transportation modes. A crucial step in developing agent-based models is the definition of agents, e.g., household and persons. While model developers wish to capture typical workday travel patterns of the entire study population of travelers, such detailed data are unavailable due to privacy concerns and technical and financial feasibility issues. Hence, modelers opt for population syntheses based on travel diary surveys, land use data, and census data. The most prominent techniques are iterative proportional fitting (IPF), iterative proportional updating (IPU), combinatorial optimization, Markov-based and fitness-based syntheses, and other emerging approaches. Yet, at present, there is no clear guideline on using any of the available techniques. To bridge this gap, this chapter presents a comprehensive synthesis of practice and documents available successful studies.

Keywords: transportation planning, traffic simulation, agent-based models, population synthesis

1. Introduction

Transportation simulation models are widely used for travel demand forecasting, testing design alternatives, or predicting travel behavior. In 1992, Axhausen and Gärling [1] developed a comprehensive review of conceptualizations and approaches of activity-based transportation models with special regard to the validity of behavioral assumptions of modeled population. In the course of their review, they concluded that individual travelers and households, rather than aggregates, should be identified and considered. Nevertheless, detailed travel records for individuals have never been easily accessible for several reasons, the most important being privacy issues and cost. Hence, individual travel diaries needed to be synthesized from travel surveys, census data, and publically available records. That process has since been known as population synthesis.

Population synthesizers initially were used as feeder data avenues to travel demand models [2]; however, recent shifts toward activity- and agent-based models brought population synthesizers to the spotlight, as they became determinants to the success or failure of any transportation model of that kind. Fitting is the core of any population synthesizer, with the main focus on fitting disaggregate sample of agents (represented by tabulated demographics of a representative sample of household and individual data) to aggregate constraints (represented by available aggregate data, such as data available from census). There are several approaches for fitting including iterative proportional fitting (IPF), iterative proportional updating (IPU), combinatorial optimization, Markov-based and fitness-based syntheses (FBS), and other emerging approaches [3]. The following sections present a critical review of each approach in the chronological order by which they were introduced to illustrate the progression and evolution of each approach, with emphasis on notable and well-established efforts.

2. Iterative proportional fitting approach

Iterative proportional fitting has been first introduced in 1940 by Deming and Stephan [4]. Since then, it became the foundation of population synthesis for transportation models and sometimes referred to as the Fratar technique [5]. The most notable realization of the IPF technique is attributed to Beckman et al. [6] who pioneered population synthesis efforts through their development of a methodology for creating a synthetic baseline population of individuals and households for microscopic activity-based models. Their technique relied on using census data represented by a Census Standard Tape File and Public Use Microdata Sample (PUMS) for a given Public Use Microdata Area (PUMA) of 100,000 individuals with matching variables. In their case, the marginal totals of a multiway table were known, and a sample from the population which generated those totals was provided; thus, they applied the IPF technique to develop constrained maximum entropy estimates of the true proportions in the population multiway table. Their rationale was built upon the consensus that IPF estimates maintain the same odds ratios as those in the sample table in the absence of any marginal information which was their case. To validate the population synthesis method, they compared demographic characteristics of the synthetic population with those of the true population using variables not involved in the population synthesis. Despite their pioneer effort, Beckman et al. [6] did not provide an answer to the zero-cell problem in the PUMS; instead, they replaced it by 0.01 and imputed the corresponding household size. Müller and Axhausen [3] illustrated this as computing a series of tabulations $n_{ij}^{(k)}$, starting with the seed at $k := 0$, thus $n_{ij}^{(0)} := n_{ij}$ for all i rows and j columns. Furthermore, they illustrated how that series can be computed as represented by Eq. (1):

$$n_{ij}^{(k+1)} := n_{ij}^{(k)} \cdot \begin{cases} r_i \div n_{i\cdot}^{(k)} \\ c_j \div n_{\cdot j}^{(k)} \end{cases} \tag{1}$$

where $n_{i.}$, is the row sum; $n_{.j}$, is the column total; r_i, is the control total for row i; c_j, is the control total for column j.

Almost a decade later, Arentze et al. [7] addressed one of the limitations of the IPF methods, that is, generating synthetic households when the demographic data describes population in terms of individual counts. Their solution relied on developing a two-step IPF procedure where, first, known marginal distributions of individuals are converted to marginal distributions of households of similar attributes and, second, the resulting marginal household distributions are used as constraints of a multiway household counts. Additionally, their approach aimed to assess the relevance of spatial heterogeneity across populations. The Dutch Albatross model was used as a case study and proof of concept. The validation results yielded sample biases in the synthetic population on the dimensions of socioeconomic class, the presence of children, and the availability of transport modes. However, they were able to resolve biases in over- or underrepresentation of groups that were related to age and work status by fitting the relevant tables on these dimensions.

Simultaneous to the efforts of Arentze et al. [7], Guo and Bhat [8] addressed the two main drawbacks of IPF approach, namely, the zero-cell problem and the inability to control for statistical distributions of both household- and individual-level attributes. Additionally, their study aimed to enhance the scalability and generality of the IPF method as it required code-level changes that are cumbersome and skills that are not typically found within planning agencies, who are the typical users of such approach. The algorithm developed by Guo and Bhat [8] featured generic data structures and accompanying functions to avoid the zero-cell problem, as well as revisions to the algorithm of Beckman et al. [6] to allow simultaneous control of both household- and individual-level attributes. That generic algorithm was built upon an object-oriented architecture and contained eight major steps and a recurring procedure for merging any two contingency tables with common variables. The proposed approach was used to generate synthetic population for the Dallas-Fort Worth metropolitan area in Texas, and the statistical comparison yielded results that were closer to true population than that of Beckman et al. [6]. In addition, Guo and Bhat [8] concluded that a higher percentage deviation from target size (PDTS) yielded better balance at satisfying the household- and individual-level multiway distributions than lower values of PDTS.

Srinivasan et al. [9] went a step further and attempted to fine-tune existing efforts to accommodate the household- and individual-level controls as well as assess the significance of controlling individual-level attributes. That study was performed in support of Florida Department of Transportation (FDOT) efforts to incorporate sociodemographic attributes within the Florida Standard Urban Transportation Model Structure (FSUTMS). The research was motivated by the need for reduced aggregation errors, ensuring sensitivity to demographic shifts like that of aging population, and the ability to accommodate population-specific transportation modes. That fine-tuning effort mainly aimed to address individual-level attributes of age and gender through the means of a greedy-heuristic data-fitting algorithm that was implemented in the matrix programming language GAUSS. Validation of Srinivasan et al. [9] algorithm yielded satisfactory distributions of household, size, age, gender, and employment status; however, the distributions for all other variables did not match well.

Given the limited number of attributes that can be synthesized per agent, researchers had to further improve the IPF approach to overcome this limitation. Pritchard and Miller [10] introduced a method that implements IPF approach with sparse list-based data structure that allows more attributes per agent. Additionally, they used both the conventional Monte Carlo integerization procedure and the conditional Monte Carlo to synthesize a list of individual agents from fitted tables. Despite their thorough efforts, the study of Pritchard and Miller [10] had only a minor impact on goodness-of-fit, relative to the conventional approach.

Auld and Mohammadian [11] developed a methodology to improve the basic IPF population synthesis routine in a manner that accounts for multiple levels of analysis units—control variables, which was a limitation to the population synthesizers mentioned hereinabove. Their methodology, named multilevel control, allows population characteristics to be replicated for multilevel synthetic population with one level (such as households) serving as the base level of analysis. After a runtime of 16 hours, the proposed method was able to synthesize a 7.9 million agent population for Chicago, IL, with an improved fit of the synthesized individual-level characteristics when compared with synthesis procedures that do not account for individual-level controls. The study concluded that the improved fit comes at no cost to the fit against household-level controls. However, the developed methodology was never experimented as to synthesizing commercial- or business-related agents.

Lee and Fu [12] realized that the IPF-based population synthesis approaches, specifically the original synthetic reconstruction method [6] and the complimentary combinatorial optimization method [13], are not generally applicable to all population synthesis scenarios. Based on a comparison by Ryan et al. [14], Lee and Fu [12] concluded that combinatorial optimization method produces more accurate demographic information for populations over a small area and that the population synthesis problem should be evaluated from an optimization point of view. In addition, they explored how the estimation of a multiway demographic table can be formulated and solved as a constrained optimization problem in full consideration of both household- and individual-level attributes. Accordingly, that study tackled the inconsistency problem through an approach that is based on the minimum cross-entropy theory. The validity of that model was confirmed through a case study in Singapore, through which results from a 10,641 household study area were superior to conventional IPF approaches. However, Lee and Fu [12] did not provide a full-scale application which constrains the applicability of their model to theoretical applications only.

Zhu and Ferreira [15] were intrigued by the inability of the standard IPF algorithm to fit marginal constraints on multiple agent types simultaneously. Hence, they developed a two-stage population synthesizer that utilized IPF on the first stage and then estimated the spatial pattern of household-level attributes through a second stage IPF-based approach. Their two-stage algorithm consisted of four distinctive steps. The first step involved developing an estimate joint distribution of household- and individual-level attributes. In the second step, households and individuals were drawn from microdata samples. The third step consisted of a conventional IPF with household type and parcel capacity marginal constraints. The fourth and last step included an estimated marginal distribution of other attributes from the fitted model. To validate their approach, Zhu and Ferreira [15] generated synthetic population for Singapore. Their evaluation approach involved four comparisons, namely, fitting only for households-level constraints, fitting for both household- and individual-level constraints, allocating households to buildings while constraining building capacity, and repeating the previous comparison with income level constrained. Validation results yielded realistic spatial heterogeneity while preserving some of the joint distribution of household and locational characteristics.

Choupani and Mamdoohi [16] addressed the issue of integerization of IPF results in non-integer values instead of integers, for example, fractions of household- or individual-level attributes for zones. In doing so, they proposed a binary linear programming model for tabular rounding in which the integerized table totals and marginals perfectly fit to input data obtained from the Census Bureau. The main advantages of using tabular rounding were that it did not bias joint or marginal distributions of socioeconomic attributes of minority demographic groups and it minimized the distortion to the correlation structure of household- and individual-level non-integer tables. Furthermore, the tabular rounding approach outperformed all other eight rounding approaches. In addition, sensitivity analysis of tabular rounding demonstrated that small and large values are equally significant when it comes to integerization. Their findings were confirmed by a comprehensive literature review [17] that they performed 1 year later, which concluded that IPF is the most feasible approach for synthesizing populations for agent- and activity-based transportation models, once integer conversion and zero-cell issues were resolved. In addition, they confirmed that tabular rounding is the most efficient and feasible solution for the integerization issue.

Most recently, in an effort to further enhance the IPF approach, Otani et al. [18] identified an issue that they named the modifiable attribute cell problem (MACP) which arises from combining discrete categories of individual-level attributes or due to the contiguous nature of those attributes. The proposed solution to the MACP issue was identified as “the organized cell set” which is the best combination of categories. The procedure to identify the best organized cell set consists of five steps. The first step involves aggregation of the elemental cell set to find several cases of cell organization that generate large cells. The second step involves constructing base-year data using the conventional IPF approach. The third step focuses on forecasting using microscopic simulation. The fourth step involves identifying the statistically acceptable cell value using a Student’s t-test. The fifth and final step involves considering the case with minimum number of cells to be the best cell organization. This method is computationally complex and cannot be performed using conventional optimization algorithms. Yet, it is the sole identifiable solution to the modifiable attribute cell problem.

3. Iterative proportional updating approach

The iterative proportional updating approach is a heuristic approach that was developed by Ye et al. [19] to address the drawbacks of the IPF approach. Specifically, the IPU approach addresses the issue of control for individual-level attributes and joint distributions of personal characteristics. The IPU algorithm matches both household- and individual-level attributes in a computationally efficient manner by iteratively adjusting and reallocating weights among households of a specific type until both household- and individual-level attributes are matched. Another advantage of the IPU approach is its practicality from the implementation and computational points of view. Eq. (2) represents the mathematical optimization problem as addressed by the IPU approach. In addition, the IPU approach has been generally described in 23 computational steps that can be easily coded in most, if not all, programming languages:

$$\text{Minimize} \sum_j \left(\frac{\sum_i d_{i,j} w_i - c_j}{c_j} \right)^2 \text{or} \sum_j \frac{\left(\sum_i d_{i,j} w_i - c_j \right)^2}{c_j} \text{or} \sum_j \frac{\left| \sum_i d_{i,j} w_i - c_j \right|}{c_j} \tag{2}$$

Subject to $w_i \geq 0$

where i, denotes a household ($i = 1, 2, \ldots, n$); j, denotes the constraint or population characteristic of interest ($j = 1, 2, \ldots, m$); $d_{i,j}$, represents the frequency of the population characteristic (household/person type j in household i); w_i, is the weight attributed to the i^{th} household; c_j, is the value of the population characteristic j.

Furthermore, Ye et al. [19] proposed an alternative method to address the zero-cell problem that undermined the IPF practicality. Their method is based on borrowing the prior information for the zero cells from PUMS data for the entire region, where zero cells are not likely to exist as long as the control variables of interest and their categories are defined appropriately. However, that method has the inherent risk of overrepresenting the demographic group of interest. Despite their attempt to overcome the zero-cell problem, the researchers could not overcome the zero-marginal problem that may result due to nonexistence of a certain attribute in households of a certain geographic area, for example, having no low-income households in a certain census block or tract. Furthermore, a review by Müller and Axhausen [3] pointed to the lack of a theoretical proof of convergence.

4. Combinatorial optimization approach

The combinatorial optimization approach was materialized by Abraham et al. [20] and is a versatile approach capable of matching targets at multiple agent levels for both household- and individual-level attributes. A combinatorial optimization approach is generally simpler and more direct than IPF. Mostly, it starts by the creation of a trial population from the disaggregate sample data, and then the overall level of fit is assessed across all marginal targets. Units from the trial population are swapped with units chosen from the disaggregate samples, and when the measure of fit improves, the swap is made. This is implemented through a proprietary computer program that first identifies a list of units whose aggregate attribute values match a pre-specified set of corresponding target values and then iteratively performs one of three operations, namely, adding a unit from the sample to the list, subtracting a unit, or swapping a unit between the sample and the previously identified list. That process is performed on a zone-by-zone level with equal probability of the three actions (i.e., add, subtract, or swap) being considered. The developed algorithm was applied to California and Oregon to synthesize populations for their models. The California application served the California Statewide Travel Demand Model including short- and long-distance travel considering personal and commercial vehicles. The Oregon application served the Oregon Statewide Integrated Model, which included employment synthesis for 34 industries. Both model applications resulted in a near-perfect fit for synthesized populations. Generally, the population synthesis procedure using combinatorial optimization has proven to be fast and flexible with the possibility for application to both households and employment scenarios. However, this algorithm can be further improved by using multicore and parallel computing techniques.

5. Markov process-based approaches

As demonstrated, hereinabove, IPF, IPU, and combinatorial optimization approaches rely on cloning attributes that were captured in microdata. In addition, they all share key drawbacks including (a) fitting of a contingency table while ignoring other solutions matching the available data; (b) loss of heterogeneity that has been captured in the microdata due to cloning rather than true population synthesis; (c) dependency on the accuracy of captured data to determine the cloning weights which may replicate inherent inaccuracies; and (d) limited scalability, in terms of the number of attributes of synthesized agents. Hence, Markov process-based approaches were developed to overcome such drawbacks and to offer an approach that truly synthesizes populations instead of cloning them.

The earliest notable effort in this direction was pioneered in 2013 by Farooq et al. [21] who developed a Markov chain Monte Carlo (MCMC) simulation-based approach for synthesizing populations. The proposed approach is a computer-based simulation technique that can be used to simulate a dependent sequence of random draws from complicated stochastic models. To synthesize populations that approach uses three sources of data, namely, (a) zoning systems such as census blocks, census tracts, counties, and states; (b) sample of individuals such as the North American PUMS and the European Sample of Anonymized Records (SARs); and (c) cross-classification tables for socioeconomics and demographics like income by age at a certain zoning level. Assuming that in a given spatial region at any point in time there exists a true population, the MCMC simulation-based approach synthesizes that population by drawing the individual attributes from their uniquely joint distribution using the available partial views while ensuring that the empirical distribution in the synthetic population is as close as possible to the unique actual distribution of that population. The proposed approach was applied to the Swiss census data, and results were compared against those developed by a conventional IPF approach. Eq. (3) illustrates the standardized root mean square error (SRMSE)-based goodness-of-fit tests that were performed on each case, and results indicated that MCMC simulation-based synthesis outperformed IPF synthesis while featuring a higher level of heterogeneity:

$$SRMSE = \frac{\left(\sum_{i=1}^{m} \dots \sum_{j=1}^{n} \left(R_{i \dots j} - T_{i \dots j}\right)^2 / N\right)^{1/2}}{\sum_{i=1}^{m} \dots \sum_{j=1}^{n} T_{i \dots j} / N} \tag{3}$$

where N, is the total number of agents; $R_{i \dots j}$, is the number of agents with attribute values $i \dots j$ in the population synthesized; $T_{i \dots j}$, is the number of agents with attribute values $i \dots j$ in the actual population.

Two years later, in 2015, Casati et al. [22] proposed an extension of the MCMC simulation-based approach to simultaneously combine both individual- and household-level attributes in a process that was named hierarchical MCMC. Furthermore, generalized raking was introduced as a technique to fit the simulated synthetic population to actual observed control totals. The hierarchical MCMC is a combination of two methods: (a) an extension of the original MCMC method that allows producing hierarchies of persons grouped into households and (b) a post-processing method to satisfy known control totals on both the individual- and household-level. That extension aimed to synthesize populations with a hierarchical structure that is based upon ordering the agents living in the same household according to their household roles. The general formulation of the extension is based upon the definition of three groups of agent types (viz., owners, intermediate, and others) running Gibbs sampling on the three groups and merging subpopulations. The proposed approach was applied to the 2008 household interview travel survey of Singapore. The application resulted in realistic synthetic populations, and SRMSE-based test confirmed the goodness-of-fit of synthesized populations and their generated hierarchical structures.

Saadi et al. [23] proposed an integrated MCMC approach and profiling-based methods to capture the behavioral complexity and heterogeneity of synthesized agents. This approach used two types of datasets, namely, (a) aggregated sociodemographic and transportation-related variables derived from household travel surveys and (b) individual activity-travel diaries collected from travel diary surveys. The integrated approach consists of six steps that run on those two data types. The first step involves performing a MCMC simulation on the sociodemographic dataset. The second step concerns synthesizing population by a Gibbs sampling procedure. The third step selects sociodemographics to compare behaviors in the activity-travel patterns. The fourth step uses results from the previous two steps to cluster synthesized populations according to sociodemographics and related activity sequences. The fifth step utilizes multiple sequence alignments to estimate hidden Markov model (HMM). The final step characterizes clusters including mixed socioeconomic effects. The integrated approach was applied to the 2010 Belgian household daily travel survey. Results indicated that the integrated approach effectively captured the behavioral heterogeneity of travelers. In addition, comparisons against IPF and IPU approaches demonstrated that the proposed integrated approach is adequately adapted to meeting the demand for large-scale microsimulation scenarios of urban transportation systems.

Realizing the advantages of Markov process-based approaches, Saadi et al. [24] developed an extended HMM-based approach which promised better alternatives than the existing ones. More specifically, the proposed HMM-based approach promised great flexibility and efficiency in terms of data preparation and model training while being able to reproduce the structural configuration of a given population from an unlimited number of micro samples and a marginal distribution. The HMM-based approach considers population synthesis as a variant of the standard decoding problem, at which the state sequences are supposed to be unknown. Accordingly, the maximum likelihood estimators related to the transition states were determined through the Viterbi algorithm. An important advantage of the HMM-based approach is its ability to handle both continuous and discrete variables, which addresses the inherent issue of loss of information due to aggregation of continuous variables like age. Also, the proposed HMM-based approach satisfies the need to discretize continuous variables to meet the fundamental limitation of Markov process to discrete states. The statistical and machine Learning Toolbox of MATLAB was used to generate sequences from an estimated HMM that were applied to the 2013 Belgian National household travel survey. Three simulations were run to illustrate the HMM-based approach. The first simulation tested the combined effects of scalability and dimensionality. The second simulation compared the HMM-based approach against IPF, and the third demonstrated the advantage of the HMM-based approach over IPF using various samples. Simulation results indicated that the proposed HMM-based approach provided accurate results due to its ability to reproduce the marginal distributions and their corresponding multivariate joint distributions with an acceptable error. Furthermore, the HMM-based approach outperformed IPF for small sample sizes while using smaller amount of input data than IPF. In addition, simulation results demonstrated that the HMM-based approach can integrate information provided by several data sources to allow good estimates of synthesized population.

6. Fitness-based synthesis approach

To address the inability of the IPF approach to deal with multilevel controls, Ma and Srinivasan [25] developed the fitness-based synthesis approach that directly generates a list of households to match several multilevel controls without the need for determining a joint multiway distribution. The FBS approach generally involves selecting a set of households from the seed data, like PUMS, such that tract-level controls are satisfied. The FBS approach starts with an initial set of households that can either be a null set or a random sample from the seed data. Then, the population of each census tract is synthesized in an iterative fashion, with one household being either added or removed from the current list in each iteration. Count tables, defined in terms of control attributes, are used to track the number of households of each type that have already been included. The FBS approach implements an adding or removing procedure, while swapping is not considered. The main criteria in the FBS approach is the reduced sum of squared error for addition F_I^{in} and corresponding error for removal F_{II}^{in} as illustrated by Eqs. (4) and (5):

$$F_I^{in} = \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[\left(R_{jk}^{n-1}\right)^2 - \left(R_{jk}^{n-1} - H\,T_{jk}^{i}\right)^2\right] \tag{4}$$

$$F_{II}^{in} = \sum_{j=1}^{J}\sum_{k=1}^{K_j}\left[\left(R_{jk}^{n-1}\right)^2 - \left(R_{jk}^{n-1} + H\,T_{jk}^{i}\right)^2\right] \tag{5}$$

Subject to $F_I^{in} + F_{II}^{in} = -2\sum_{j=1}^{J}\sum_{k=1}^{k_j}\left(H\,T_{jk}^{i}\right)^2$

where $R_{jk}^{n-1} = T_{jk} - CT_{jk}^{n-1}$; j, is an index representing the control (and the corresponding count) tables; J, is the total number of control (or count) tables; jk, is an index representing the different cells in a table; T_{jk}, represents the value of cell k in control table j; CT_{jk}^{n-1}, represents the value of cell k in count table j after iteration $n-1$; R_{jk}^{n-1}, is the number of households/persons required to satisfy the target for cell k in control table j after iteration $n-1$; HT_{jk}^{i}, is the contribution of the i^{th} household in the seed data to the k^{th} cell in control table j.

Three applications of the FBS approach were performed to demonstrate the feasibility of incorporating many controls at multiple levels in the synthesis and increased accuracy of synthesized population. The three applications were performed using the 2000 Census data for 12 census tracts in Florida. The first application involved population synthesis using the IPF approach with only household-level controls. The second application involved population synthesis using the proposed FBS approach with few household- and individual-level controls. The third application also involved population synthesis using the FBS approach but with significantly larger number of controls. Validation for the three applications was performed by comparing the mean absolute error against 22 artificial census tracts that were created by randomly selecting subsets of households from the 2000 PUMS. Validation results demonstrated that FBS outperformed IPF and demonstrated efficiency and scalability. In addition, FBS did not require many iterations as it required only one to three times the number of households to be synthesized. In addition, the proposed FBS approach addresses the notorious IPF issues of zero-cell problems, computational resources (memory), and non-integers cell value in the joint-distribution tables.

Hafezi and Habib [26] refined the FBS approach, and the refined FBS population synthesizer was examined by three models. The first model used household-level control tables. The second model used individual- and household-level control tables, and the third model used weighting individual-and household-level control tables. The models were applied to the province of Nova Scotia in Atlantic, Canada, using the 2006 Canadian Census and Public Use Microdata File (PUMF). The refined approach was implemented using the sparse matrix technique package in MATLAB that is based on high-level matrix programming for numerical computation. The three models were validated by error percentages and goodness-of-fit evaluation. Validation results indicated that the refined FBS approach can efficiently obtain a satisfactory result using both individual- and household-level control tables. However, higher homogeneity was achieved within the third model.

7. Emerging approaches

Other emerging approaches have been developed in an attempt to replace the IPF approach or to overcome one or more of its drawbacks. Emerging approaches include Bayesian network, annealing algorithm, linear programming, heuristic-based, copula-based, and entropy maximization approaches. The following paragraphs introduce each of the emerging approaches.

The Bayesian network approach was developed by Sun and Erath [27] in 2015. The proposed Bayesian network approach is a probabilistic population synthesizer that is intended as an alternative to approximate the inherent joint distribution in a more efficient manner. Using a graphical model, the proposed Bayesian network approach encodes probabilistic relationships, like causality or dependence, among a set of variables. The advantages of Bayesian network models lie in their ability to learn the structure of population systems, particularly when the number of attributes of interest is large using limited amounts of microdata. The Bayesian network approach was founded on the inference of the joint distribution—that is, perceiving the population synthesis problem as an inference of a multivariate probability distribution of demographic and socioeconomic household- and individual-level attributes. Like the Markov process-based approaches, the Bayesian network approach does not require marginals as input. In addition, it does not require any conditionals since structure learning and parameter estimation are inherently integrated in the learning model. The performance of the proposed Bayesian network approach was demonstrated through an application to the 2010 household interview travel survey of Singapore. The Bayesian network approach demonstrated good performance as illustrated by low SRMSE values. It also demonstrated good heterogeneity in synthetic population when the size of PUMS is less than 70% of the full population.

The simulated annealing (SA) algorithm was developed by Kim and Lee [28] to synthesize populations for activity-based models. The proposed SA algorithm is built upon the concepts of thermodynamics and metallurgy and was first introduced as a generic heuristic method for discrete optimization. The Metropolis-Hastings Algorithm was employed to solve the inherent problems of hill climbing and cooling schedule when applying SA to population synthesis. The proposed algorithm consists of seven steps. The first step concerns setting the maximum number of iterations. The second step sets up the total amount of columns and rows in the population and enters observed values of sample distribution. The third step sets up the before-distribution, which is composed by random numbers, while satisfying the total amount of restrictive conditions. The fourth step sets up the after-distribution, which is also composed by random numbers that satisfy total amount restrictive conditions. The fifth step involves calculation of absolute error on the before–/after-distributions as well as observed data. The sixth step involves calculation of selection probability. The seventh and final step iterates steps 4 through 6 and ends the calculations when the absolute error (calculated in the fifth step) has the smallest value or satisfies ending conditions. The SA algorithm was implemented using the household travel diary survey from the Korean National Statistics Office. Results from the implementation indicated the need for further verification of the accuracy of this algorithm.

The linear programming (LP) approach was developed by Vovsha et al. [29] to synthesize populations as part of an activity-based model developed for the Maricopa Association of Governments. The LP approach is an analytical method that balances a list or sample of household weights to meet the controls imposed at some spatial level, typically, for each traffic analysis zone (TAZ). Features of the LP approach include (a) the general formulation of convergence of the balancing procedure with imperfect controls, (b) optimized discretization of weights while preserving the best possible match to the controls, and (c) ability to set controls at multiple spatial levels. In addition, the proposed LP approach featured an innovative discretizing method applied for the household weights and integrated with the balancing procedure. While validation of the proposed LP approach is questionable, it still demonstrates reasonable accommodation to various fine-resolution spatial levels that are much needed by newer-generation activity- and agent-based models.

The heuristic-based approach was developed by Zhuge et al. [30] to address two IPF limitations that received less attention from earlier studies. The first limitation stems from the existence of various solutions for one target marginal distribution. The second limitation stems from the optimization nature of population synthesis with the objective function being minimizing the mean absolute percentage error (MAPE) of control variables. The proposed heuristic-based approach consists of 11 steps arranged in three parts. The first part, including steps 1 and 2, is used to generate the initial household weights. The second part, including steps 3 through 11, adjusts the household weights until a stop criterion is met. The third part, including steps 10 and 11, calculates the adjustment steps and adjustment range, which are two fundamental parameters of the approach. The 2007 household travel survey data from Baoding, China, were used as a case study. Results indicated that heuristic-based approach cannot perform as well as IPF-based on comparing MAPE values for both approaches.

Most recently, the copula-based approach was proposed by Kao et al. [31] to address previously identified limitations of IPF approach. Copulas are joint probability distributions with uniform marginal, which are a relatively new statistical tool. Hence, the copula-based approach was designed to preserve marginal distributions and dependence structure between variables. The proposed method was tested for the state of Iowa, and the results were compared with the IPF approach using mean, median, and correlation matrices. The synthesized households resulted in the same local statistics at each block group, but having similar intervariable correlations as described in the PUMS suggests the applicability of the copula-based approach.

Another recent effort to develop an alternative to IPF approaches resulted in the development of entropy maximization-based population synthesizer by Paul et al. [32] which handles multiple geographies and avoids algorithmic errors. The entropy maximization approach was developed as part of the Oregon Department of Transportation (ODOT) effort to utilize an open-source population synthesis platform. The approach consists mainly of two algorithms. The first algorithm, namely, list balancing, finds weights that match the given marginal control distributions. The second algorithm, namely, integerizing, implements a LP-based procedure to covert fractional weights to integers. The proposed entropy maximization-based approach was implemented in Python and made heavy use of the Pandas and NumPy libraries, which allow for vectorization of operations to reduce overall runtime. Validation results against those of IPF approach were promising and demonstrated reasonable match to controls.

8. Conclusion

This study presented a critical, comprehensive literature review of population synthesizers starting from the early efforts through the most recent approaches. The review and synthesis indicated that, despite its identified limitations and drawbacks, IPF approach is the most feasible and widely used population synthesizer. All other studies and efforts used it as a reference for comparison and produced similar or slightly improved results. Evidently, IPF has its drawbacks and limitations. Yet reviewed literature indicates that there is no single approach that can result in an efficient and accurate population synthesizer. However, an integration of robust methods appears as the most promising approach, like the effort of Fournier et al. [33] where the limitations of IPF are resolved by combining five methods into an integral framework for population synthesis. **Table 1**, in the *Supplemental Information* section, summarizes the advantages and disadvantages of the presented approaches.

Almost three-decade old, yet the IPF approach is still being used in state-of-the-art simulation platforms like MATSim. Given that IPF is the most studied approach and the fact that none of the alternatives provided an out-of-the-box solution, IPF is preferred approach by modelers and practitioners. This conclusion is confirmed by the findings of Saadi et al. [34], who investigated the influence of scalability on the accuracy of different population synthesizers using both fitting- and generation-based approaches. Their results revealed that simulation-based approaches are more stable than IPF approaches when the number of attributes increases; however, IPF approaches are less sensitive to changes in sample size.

Approach	Advantage(s)	Disadvantage(s)
Iterative proportional fitting (IPF)	Synthesized estimates maintain the same odds ratios as those in the sample table Most studied and improved approach with more than 20 years of continuous refinements Widely available with ready-to-use implementations in several computer programming languages	Does not provide an answer to the zero-cell problem in the public use microdata sample (PUMS) Unable to control for statistical distributions of both household- and individual-level attributes Limited number of attributes that can be synthesized per agent
Iterative proportional updating (IPU)	Addresses the issue of control for individual-level attributes and joint distributions of personal characteristics Computationally efficient Described in 23 computational steps that can be easily coded in most programming languages	Cannot overcome the zero-marginal problem that may result due to nonexistence of a certain attribute in the households of a certain geographic area
Combinatorial optimization	Generally simpler and more direct than IPF Fast and flexible with the possibility for application to both households and employment scenarios	Implementation is limited to a proprietary computer program Resource-demanding and needs multicore, parallel computers
Markov process-based approach	Truly synthesizes populations instead of cloning them Meets the demand for large-scale microsimulation scenarios Can handle both continuous and discrete variables	Requires extensive knowledge of computer programming Difficult to trace errors Refinement for specific scenarios or locations requires substantial redevelopment of the computer algorithm
Fitness-based synthesis (FBS)	No need for determining a joint multiway distribution Addresses the notorious IPF issues of zero-cell problems	Requires extensive knowledge of the sparse matrix technique package in MATLAB that is based on high-level matrix programming for numerical computation
Emerging approaches	Scalable and adaptive Addresses all disadvantages of IPF approach	Requires advanced expertise in Python and makes heavy use of the Pandas and NumPy libraries Limited successful applications compared to IPF

Table 1.
Key advantages and disadvantages of population synthesis approaches.

Overall, this study provides a critical review and comprehensive synthesis of population synthesis approaches that can serve as a valuable reference to future efforts focusing on population synthesis for activity- and agent-based transportation models.

Acknowledgements

This study was performed in support of the project "Technology Influence on Travel Demand and Behaviors" that was sponsored by the US Department of Transportation Office of the Assistant Secretary for Research and Technology (OST-R) through the Southeastern Transportation Research, Innovation, Development, and Education (STRIDE) Center under Contract No. 69A3551747104 with matching funds from the Alabama Department of Transportation (ALDOT).

Author details

Ossama E. Ramadan and Virginia P. Sisiopiku*
University of Alabama at Birmingham, Birmingham, AL, USA

*Address all correspondence to: vsisiopi@uab.edu

References

[1] Axhausen KW, Gärling T. Activity-based approaches to travel analysis: Conceptual frameworks, models, and research problems. Transport Reviews. 1992;**12**(4):323-341

[2] Bowman JL, Rousseau G, editors. Validation of Atlanta, Georgia, regional commission population synthesizer. In: Innovations in Travel Demand Modeling Conference; 2006; Austin, TX. Washington, DC: Transportation Research Board; 2008

[3] Müller K, Axhausen KW, editors. Population synthesis for microsimulation: State of the art. In: Transportation Research Board 90th Annual Meeting. Washington, DC: Transportation Research Board; 2011

[4] Deming WE, Stephan FF. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. The Annals of Mathematical Statistics. 1940;**11**(4):427-444

[5] Papacostas CS, Prevedouros PD. Transportation Engineering and Planning. 3rd ed. Englewood Cliffs, NJ: Prentice Hall; 2001

[6] Beckman RJ, Baggerly KA, McKay MD. Creating synthetic baseline populations. Transportation Research Part A: Policy and Practice. 1996;**30**(6):415-429

[7] Arentze T, Timmermans H, Hofman F. Creating synthetic household populations: Problems and approach. Transportation Research Record. 2014;**2007**:85-91

[8] Guo JY, Bhat CR. Population synthesis for microsimulating travel behavior. Transportation Research Record. 2014;**2007**:92-101

[9] Srinivasan S, Ma L, Yathindra K. Procedure for Forecasting Household Characteristics for Input to Travel-Demand Models. Report No. TRC-FDOT-64011-2008. Tallahassee, FL: Florida Department of Transportation; 2008

[10] Pritchard DR, Miller EJ, editors. Advances in agent population synthesis and application in an integrated land use and transportation model. In: Transportation Research Board 88th Annual Meeting. Washington, DC: Transportation Research Board; 2009

[11] Auld J, Mohammadian A. Efficient methodology for generating synthetic populations with multiple control levels. Transportation Research Record. 2010;**2175**:138-147

[12] Lee D-H, Fu Y. Cross-entropy optimization model for population synthesis in activity-based microsimulation models. Transportation Research Record. 2011;**2255**:20-27

[13] Williamson P, Birkin M, Rees PH. The estimation of population microdata by using data from small area statistics and samples of anonymised records. Environment and Planning A: Economy and Space. 1998;**30**(5):785-816

[14] Ryan J, Maoh H, Kanaroglou P. Population synthesis: Comparing the major techniques using a small, complete population of firms. Geographical Analysis. 2009;**41**(2):181-203

[15] Zhu Y, Ferreira J. Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. Transportation Research Record. 2014;**2429**:168-177

[16] Choupani A-A, Mamdoohi AR. Population synthesis in activity-based models. Transportation Research Record. 2015;**2493**:1-10

[17] Choupani A-A, Mamdoohi AR. Population synthesis using iterative proportional fitting (IPF): A review and future research. Transportation Research Procedia. 2016;**17**:223-233

[18] Otani N, Sugiki N, Vichiensan V, Miyamoto K. Modifiable attribute cell problem and solution method for population synthesis in land use microsimulation. Transportation Research Record. 2018;**2302**:157-163

[19] Ye X, Konduri KC, Pendyala RM, Sana B, Waddell P, editors. Methodology to match distributions of both household and person attributes in generation of synthetic populations. In: Transportation Research Board 88th Annual Meeting. Washington, DC: Transportation Research Board; 2009

[20] Abraham JE, Stefan KJ, Hunt JD, editors. Population synthesis using combinatorial optimization at multiple levels. In: Transportation Research Board 91st Annual Meeting. Washington, DC: Transportation Research Board; 2012

[21] Farooq B, Bierlaire M, Hurtubia R, Flötteröd G. Simulation based population synthesis. Transportation Research Part B: Methodological. 2013;**58**:243-263

[22] Casati D, Müller K, Fourie PJ, Erath A, Axhausen KW. Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. Transportation Research Record. 2015;**2493**:107-116

[23] Saadi I, Mustafa A, Teller J, Cools M. Forecasting travel behavior using Markov chains-based approaches. Transportation Research Part C: Emerging Technologies. 2016;**69**:402-417

[24] Saadi I, Mustafa A, Teller J, Farooq B, Cools M. Hidden Markov model-based population synthesis. Transportation Research Part B: Methodological. 2016;**90**:1-21

[25] Ma L, Srinivasan S. Synthetic population generation with multilevel controls: A fitness-based synthesis approach and validations. Computer-Aided Civil and Infrastructure Engineering. 2015;**30**(2):135-150

[26] Hafezi MH, Habib MA, editors. Synthesizing population for agent-based microsimulation modeling in Atlantic Canada. In: Transportation Research Board 94th Annual Meeting. Washington, DC: Transportation Research Board; 2015

[27] Sun L, Erath A. A Bayesian network approach for population synthesis. Transportation Research Part C: Emerging Technologies. 2015;**61**:49-62

[28] Kim J, Lee S. A simulated annealing algorithm for the creation of synthetic population in activity-based travel demand model. KSCE Journal of Civil Engineering. 2015;**20**(6):2513-2523

[29] Vovsha P, Hicks JE, Paul BM, Livshits V, Maneva P, Jeon K, editors. New features of population synthesis. In: Transportation Research Board 94th Annual Meeting. Washington, DC; 2015

[30] Zhuge C, Li X, Ku C-A, Gao J, Zhang H. A heuristic-based population synthesis method for micro-simulation in transportation. KSCE Journal of Civil Engineering. 2017;**21**(6):2373-2383

[31] Kao S-C, Kim HK, Liu C, Cui X, Bhaduri BL. Dependence-preserving approach to synthesizing household characteristics. Transportation Research Record. 2018;**2302**:192-200

[32] Paul BM, Doyle J, Stabler B, Freedman J, Bettinardi A, editors. Multi-level population synthesis using entropy maximization-based simultaneous list balancing. In:

Transportation Research Board 97th Annual Meeting. Washington, DC: Transportation Research Board; 2018

[33] Fournier N, Christofa E, Akkinepally AP, Azevedo CL, editors. An integration of population synthesis methods for agent-based microsimulation. In: Transportation Research Board 97th Annual Meeting. Washington, DC; 2018

[34] Saadi I, Eftekhar H, Teller J, Cools M, editors. Investigating the scalability in population synthesis: A comparative approach. In: Transportation Research Board 96th Annual Meeting. Washington, DC: Transportation Research Board; 2017