Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method

Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method

Cite this article as: Nicholas Fournier, Eleni Christofa, Arun Prakash Akkinepally and Carlos Lima Azevedo, Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method, Transportation https://doi.org/10.1007/s11116-020-10090-3

Noname manuscript No.
(will be inserted by the editor)

Integrated population synthesis and workplace assignment using an efficient optimization-based person-household matching method

Nicholas Fournier, Ph.D. · Eleni Christofa, Ph.D. · Arun Prakash Akkinepally, Ph.D. · Carlos Lima Azevedo, Ph.D.

the date of receipt and acceptance should be inserted later

Abstract Large scale activity-based simulation models inform a variety of transportation and planning policies using models that often rely on fixed or flexible workplace location in a synthetic population as input to work related activity, participation, and subsequent destination dependent travel decisions. Although discrete choice models can estimate workplace location with greater flexibility, disaggregate data available (e.g., travel surveys) are often too sparse to estimate workplace location at sufficient spatial detail. Alternatively, aggregated employment data are often readily available at higher spatial resolutions, but are typically only used in separately estimated *ad hoc* models, which introduces error if the estimations have divergent solutions. This paper's primary contribution is to reduce error by integrating population synthesis and workplace assignment, yielding a synthetic population with home and work locations included as attributes. The two are integrated using additional variables shared between population and workplace assignment (i.e., industry sector), but this increased matrix size can render conventional multilevel person-household re-weighting methods computational intractable. A secondary contribution is to mitigate this scalability challenge using more efficient optimization-based re-weighting approaches, substantially reducing computation time. The proposed process is applied to the Greater Boston Area, generating a population of 4.6-million persons within 1.7-million households across 965 census tract zones. The integrated process is compared against conventional *ad hoc* location assignment process, using both classical and contemporary synthesis techniques of Iterative Proportional Fitting, Markov chain Monte Carlo simulation, and Bayesian Network simulation. The integrated approach yielded an improvement in workplace location assignment, with only modest impact on population accuracy.

Keywords population synthesis · workplace assignment · robust regression · joint re-weighting · iterative proportional fitting

1 Introduction

Agent-based microsimulation is a mainstay for transportation and land-use planning, using an ever growing array of large-scale modeling platforms such as MATSim (Balmer et al., 2009), UrbanSim (Waddell, 2002), SimMobility (Adnan et al., 2016), ILUTE (Salvini and Miller, 2005; Wagner and Wegener, 2007), MUSSA (Martinez and Donoso, 2010), and DaySim (Bowman et al., 2014) to inform a variety of decisions, such as policy, investment, and operation. With the field of transportation simulation shifting away from classical trip-based approaches towards purely activity-based models, a great deal of research has focused on improving the synthesis methods for flexible and accurate dissagregate populations of agents with high

N. Fournier
University of California, Berkeley, CA 94720, USA
E-mail: nick.fournier@berkeley.edu

E. Christofa
University of Massachusetts, Amherst, MA 01003, USA
E-mail: christofa@ecs.umass.edu

A. Akkinepally
Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA
E-mail: arunprak@mit.edu

C. Azevedo
Technical University of Denmark, Anker Engelunds Vej 1, 2800 Kgs. Lyngby, Denmark
E-mail: climaz@dtu.dk

spatial resolution of home location. However, workplace location is still an important input to activity-based models for work and related travel activity, yet substantially less attention has been given to workplace assignment and overall synthesis frameworks. Conventionally, workplace location in a synthetic population is assigned using a separately estimated *ad hoc* model, potentially introducing error by not fitting for both targets (i.e., population and workplace location assignment) simultaneously. The motivation of this paper is to address the error introduced from divergent population synthesis and workplace assignment estimations by presenting a framework for integrating these two processes to reduce workplace assignment error. In addition, this paper introduces and evaluates a computationally more efficient re-weighting method for generating multilevel joint person and household populations. The improved efficiency is necessary for computational tractability in handling the increased matrix sizes introduced with integrated synthesis. However, the optimization-based re-weighting can be used in any multilevel population synthesis, making larger scale multilevel population synthesis more scalable in general.

1.1 Background

Population synthesis and workplace destination assignment utilize similar joint distribution fitting methods, such as Iterative Proportional Fitting (IPF), Markov chain Monte-Carlo simulations (MCMC), or Bayesian Networks (BN); yet to date the two processes have not been integrated. The benefits of such an integration not only provides a more seamless generation and assignment process, but can greatly reduce the potential for error. This paper describes such an integration applied to a population of 4.6-million persons and 1.7-million households allocated across 965 zones in the Greater Boston Area (GBA). This is achieved through a multi-step synthesis process where a joint distribution for a workplace assignment model of home (origin), workplace (destination), and industry sector is estimated and then subsequently used as a constraint in the joint distribution of persons. The industry sector acts as a shared variable between the workplace and person distributions, enabling the joint distribution estimation for population synthesis (e.g., IPF, MCMC, or BN) to minimize error in the population with respect to work place assignment, reducing overall workplace assignment error in otherwise potentially divergent solutions.

To ensure a high degree of accuracy is achieved when integrating persons and workplace assignment, the linking variable(s) (i.e., industry sector in this case) should be as detailed as possible. However, matrix dimensionality increases with detail and the assignment quickly becomes computationally intractable. This is particularly true during joint multilevel person-household re-weighting, a re-weighting step for allocating persons into household groups with household attributes. Conventionally, this process is achieved using an algorithm called Iterative Proportional Updating (IPU); however, this is highly computationally intensive and can fail to find a global optimum. This paper aims to fill this gap by proposing an optimization-based approach to re-weighting, achieving substantially faster computation times, which allows for a much more scalable population synthesis process.

1.2 Contributions

This proposed unified process makes two contributions; first by integrating population synthesis and workplace assignment, and second by developing a more efficient multilevel person-household re-weighting approach to handle the additional population attributes added. The integrated synthesis process is compared against conventional *ad hoc* workplace assignment using both classical synthesis methods of Iterative Proportional Fitting (IPF), as well as contemporary probabilistic methods of Markov chain Monte Carlo Gibbs (MCMC) sampler and Bayesian Networks (BN). Results yield an improvement in workplace assignment in the integrated process with only minor loss of person-household accuracy. The different synthesis methods also yielded trade-offs, with IPF achieving greater aggregated marginal fit and workplace assignment accuracy, but less accurate at the microdata joint distribution level compared to MCMC and BN methods. The proposed optimization-based multilevel person-household re-weighting method is compared against conventional Iterative Proportional Updating (IPU) using a classical quadratic non-negative least squares (NNLS) algorithm, a linear optimization of non-negative least deviation (NLAD), and cyclical coordinate decent (CCD). The results show the CCD method capable of achieving comparable re-weighting accuracy at nearly $^1/_{15}$ of the time required by IPU. Overall, these two contributions improve the accuracy and scalability of synthetic population generation, ultimately benefiting agent-based simulation models and their applications.

2 Literature review

Despite being able to share common fitting methods in population synthesis and workplace assignment, the two are typically performed as completely independent processes due to computational tractability or proprietary program scope (Briem et al., 2019). To clearly discuss the two, the following background discussion is divided into two main sections of population synthesis and workplace assignment.

2.1 Population synthesis

In general, population synthesis methods can be categorized into three broad groups: (1) Iterative Proportional Fitting, (2) Combinatorial Optimization (CO), and (3) Statistical Learning and Probabilistic Simulation-based approaches. The following literature review of population syntheses is structured around these three groups.

2.1.1 Iterative Proportional Fitting (IPF)

Population synthesis data can be cleaved into two distinct types, aggregated and disaggregated data. Aggregated data are the totals of a particular subject or variable (e.g., total number of men or women), referred to as *marginal* data. Aggregated population data in the U.S. generally is available from the U.S. Census Bureau (U.S. Census Bureau, 2010, 2015), which provides tabulated totals for variables, such as totals by age, sex, occupation, etc. Disaggregated data in contrast, are comprised of individual persons in the population and their characteristics, referred to as *microdata*. For decades the backbone of most population synthesizers has been IPF, a method for expanding a small microdata sample (called a seed) to match marginal totals through an iterative fitting process (Deming et al., 1940; Stephan, 1942; Choupani and Mamdoohi, 2016; Pritchard and Miller, 2012).

Introduced by Deming et al. (1940), IPF is an iterative process used to fit joint distribution cells in an n-dimensional contingency table when the marginal totals are known. Mosteller (1968) advanced IPF by showing that cross-product ratios could be used to adjust the table while preserving its structure at each iteration. Then Ireland and Kullback (1968) further showed that cell probabilities can be estimated for multi-way contingency tables, the importance of this is that IPF can be extended to high dimensional contingency tables. Wong (1992) tested the utility of IPF for generating populations for geographers, while Beckman et al. (1996) was the first to utilize IPF for population synthesis with disaggregated travel demand modeling.

IPF requires initial seed values to begin proportional fitting. Any zero cells in the seed will remain as a zero during IPF and not be fitted. There are two types of zero cells, “sampling” zeros that occur when there are no representatives captured in the sample (e.g., rare combinations), and “structural” zeros that represent impossible combinations in the data (e.g., a head of household that is under aged). The difficulty in handling zero cells is the need to preserve structural zeros while adding heterogeneity by filling sampling zeros. One solution to the zero cell problem is to simply set a very small arbitrary value (e.g., 0.001) for zero cells (Beckman et al., 1996). This allows the cell to be fitted and helps IPF to converge. However, this also removes any structural zeros in the seed, introducing the potential for impossible combinations to occur. Another solution is to substitute missing cells using values from a larger sample (e.g., the entire study area rather than a sub region). In order to ensure proportional unity, the borrowed values are adjusted proportionally by the ratio of the sub-sample size to the total sample size (Ye et al., 2009; Guo and Bhat, 2007).

2.1.2 Combinatorial Optimization

Though popular, IPF is not the only technique used in population synthesis. Another classical deterministic approach is CO (Openshaw and Rao, 1995; Voas and Williamson, 2000; Abraham et al., 2012). CO treats population synthesis as an optimization problem, where the number of representatives in the joint sample (i.e., sample weight) is optimized to match the marginal totals. CO also offers the possibility of integer optimization, eliminating the need for probabilistic sampling or decimal “integerization” (Lovelace and Ballas, 2013). However, a major weakness of using CO is the inherent disregard for attribute association and weight (i.e., the frequency of an attribute combination) (Pritchard and Miller, 2012). While IPF will preserve patterns in a microdata sample based on frequency, CO will minimize error even if it means setting unrealistic weights (e.g., zero). This potentially leads to over-fitting or loss of heterogeneity. In general, CO is less common and has several shortcomings, but can provide precise and computationally efficient results (Hermes and Poulsen, 2012).

2.1.3 Probabilistic Simulation

IPF and CO rely on classical fitting and re-weighting methods for populations, but more recently a pure simulation based probabilistic approach has proven superior in many regards. Rather than determining household weights using IPF and then drawing, simulation-based approaches effectively fit and draw samples simultaneously by sampling directly with a conditional MCMC. Farooq et al. (2013) used a Gibbs sampler to draw from a person level population sample, checking the fit against marginals to achieve a near perfect fit.

A potential weakness in MCMC simulation-based methods is a lack of heterogeneity in the sample, meaning that persons or households cannot be synthesized in the population if they are not represented in the sample (Farooq et al., 2013). Sun and Erath (2015) proposed a new approach using Bayesian Networks (BN) to map and reconstruct the joint conditional probabilities one pair of variables at a time from their partials in the population; in effect, reintroducing heterogeneity into the population that may have been lost by solely relying on full joint conditionals. This ability to reconstruct populations also means that the method requires smaller sample sizes than IPF to achieve a satisfactory level of accuracy. Furthermore, unlike IPF or CO which are limited to discrete categorical frequencies, a major benefit of probabilistic approaches is the ability to handle continuous variables as well as discrete variables. This not only increases flexibility, but can improve scalability by using a single parametric function (e.g., Gaussian) rather many small discrete segments.

Branching from the "expert knowledge" driven approaches of Bayesian Networks, fully unsupervised machine learning techniques are becoming increasingly utilized in population synthesis. Saadi et al. (2016) employed Hidden Markov Models (HMM) to capture hidden correlations between the diversity of variables in subgroups of the population. Machine learning techniques are gaining further attention as agent-based models demand increasing detailing synthetic populations, easily exceeding computational limits of IPF, MCMC, and BN approaches. Borysov et al. (2019) utilized a variational auto-encoder, which "decodes" a machine learned model to overcome scalability issues for very complex populations.

2.1.4 Synthesizing multilevel populations

Activity-based models often rely on decisions made at the household level (Guo and Bhat, 2007). For this reason it is often necessary to synthesize a multilevel population (i.e., persons and households). Generating multilevel populations tends to be one of the most challenging problems in population synthesis. In general, multilevel populations are synthesized by drawing households from a joint microdata sample of persons and households. The sampled households along with their associated persons are replicated into a pool of joint persons and households (Beckman et al., 1996; Auld and Mohammadian, 2010).

Beckman et al. (1996) estimated joint populations by fitting households using IPF, then used the IPF weights to draw from a joint sample. However, using only households leaves person characteristics uncontrolled, therefore introducing error. Error was partially mitigated by incorporating broad person level variables into households (e.g., number of workers, children, or adults). This also improved through sampling algorithms, relation matrices, multiple IPF steps, or improved classification and regression trees (Le et al., 2016; Zhu and Ferreira, 2014; Guo and Bhat, 2007; Arentze et al., 2007; Arentze and Timmermans, 2004). Ye et al. (2009) provided a breakthrough by proposing a novel fitting algorithm called Iterative Proportional Updating (IPU). IPU re-weights households in a microdata sample using separate population weights (e.g., from IPF) for persons and households as marginal constraints in the subsequent IPU step. This yields a single joint weight that accounts for both persons and households simultaneously. The algorithm is performed by structuring the joint person-household sample data into a joint list. The household and person types are combinatorial, meaning that there is a unique cell in a matrix for each possible combination of household or person variables. Depending on the sample size and possible combinations, the resulting table can become an extremely large and sparse matrix, quickly becoming computationally cumbersome.

Alternatively, sample-less populations may be generated using structured marginals (Barthelemy and Toint, 2013) with IPF. The weights are then integerized and replicated to form a near perfect disaggregate population (Lovelace et al., 2014; Ballas et al., 2005b,a). However, this destroys the intricate household-person relationships that can be extracted organically from a joint sample. Multi-level populations must then be reconstructed using an algorithm, but this often comes with a loss of accuracy (Lovelace and Dumont, 2016). Sample-based approaches tend to be preferred, largely because public use microdata are typically available in most countries where population synthesis is performed. Examples of such data include Public Use Microdata Sample (PUMS) in the United States (U.S. Census Bureau American Community Survey, 2015), Public Use Microdata Files (PUMFs) in Canada, and Samples of Anonymised Records (SARs) in the United Kingdom.

While probabilistic simulation based approaches (e.g., BN and MCMC) have yielded superiority in synthesizing individual populations, the techniques on their own do not possess the ability to synthesize multilevel populations (e.g., joint person-household). Casati et al. (2015) improved upon MCMC approaches by proposing a two-step method using a Gibbs sampler followed by a re-weighting step to satisfy both individual and household margins. Sun et al. (2018) further expanded their seminal BN approach to use latent class models with rejection sampling to synthesize multilevel populations.

2.2 Workplace assignment

Traditional trip-based models allocate aggregated travelers from origins to destinations using an origin-destination (OD) assignment matrix fitted with aggregated trip generation data. For example, the number of workers that live in each origin and the total number of workers that work in each destination. To fit the matrix, the cells in the matrix (i.e., OD pairs) are given an initial weight based on some weighting scheme, such as the common "gravity model" (Voorhees, 1956). Most aggregated trip-based models fall into this classical model of iterative fitting, but vary by their weighting procedures (Abdel-Aal, 2014), such as the maximum entropy (Wilson, 2011), intervening opportunities (Stouffer, 1940), or radiation laws (Simini et al., 2012). These models make alternative assumptions or add complexity in order to account for a variety of socioeconomic factors. However, these aggregated approaches all rely on IPF and are confined to a single trip purpose at a time (e.g., work trips).

With the development of discrete choice models and the ability to break free from single purpose OD matrices, transport modeling has largely shifted away from rigid deterministically fit assignment models (McFadden, 1978; Train, 1986; Ben-Akiva and Lerman, 1985). An ever growing family of increasingly complex models are being developed to model individual decisions (e.g., for mode, purpose, time of day, and destination) (Bowman and Ben-Akiva, 2001; Bowman et al., 1998; Dong et al., 2006; Recker, 2001). However, with increased spatial resolution the combinatorial problem quickly becomes intractable. While methodologies to deal with the limitations of discrete spatial choice modeling have been proposed (Guevara, 2010), this still poses a problem to fine grain destination choice models as sample data can become too sparse for accurate estimation.

Major advancements in population synthesis has been achieved through research in recent years, but much of the attention has been focused on improving statistical fit in a single region, and not a spatially distributed population. Probabilistic methods do not forbid integration of workplace assignment and population generation per se, but no examples were found in the literature yet. There is however, a burgeoning body of literature focused on extracting origin-destination activity behavior of individuals (Nakanishi et al., 2018; Anda et al., 2018; Li et al., 2019; Bassolas et al., 2019; Bachir et al., 2019). Such large-scale mobility data holds great data-fusion potential (Huang et al., 2018) and practical applications. One particularly relevant attempt by Zhang et al. (2019) used passively collected call records to generate a synthetic population with more detailed home locations. A major step towards breaking free of discrete traffic analysis zones.

3 Methodology

The proposed methodology makes two contributions, first to integrate population synthesis with workplace assignment for improved accuracy, and second to make joint multilevel person-household synthesis more scalable through a more efficient optimization based re-weighting approach. The proposed integrated population synthesis and workplace assignment process is displayed visually through a schematic flowchart in Figure 1. In general, the process is divided into four steps: (1) origin-destination-industry synthesis, (2) separate person and household synthesis, (3) joint re-weighting, and (4) joint sampling. For comparison, the conventional *ad hoc* workplace assignment is displayed as the dashed line.

3.1 Integrated synthesis methods

For comparison, this paper runs the entire process in Figure 1 using three different synthesis methods in steps (1) and (2): Iterative Proportional Fitting (IPF), a Markov chain Monte-Carlo Gibbs sampler (MCMC), and Bayesian Network based simulation (BN). Within each full generation process, a synthesis method (i.e., IPF, MCMC, or BN) is used at three separate instances: persons, households, and origin-destination-industry (ODI). The ODI synthesis is performed in step labeled (1) as a pre-processing step. The purpose of the pre-processing step is to obtain a multidimensional joint distribution for origin, destination, and industry from separate "flat" two-dimensional marginal tables. The resulting joint distribution is then subsequently used as a marginal in step (2) for the person level synthesis.

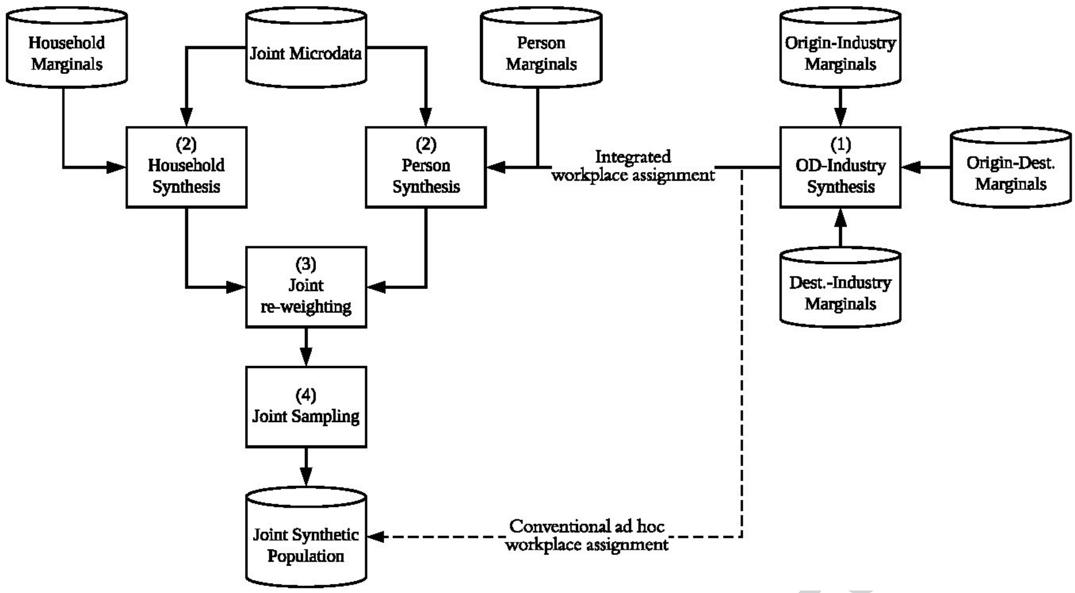


Fig. 1: Modeling framework

For further comparison of the proposed integrated assignment, the entire process with each synthesis method is run a second time using a conventional workplace assignment process. In conventional assignment, steps (1) and (2) are performed independently of each other, where the joint ODI distribution is not used as a marginal and skips the second step. Workplace location is then probabilistically assigned directly from the ODI distribution after the full population has been synthesized (see the dashed line in Figure 1).

To perform the integrated workplace assignment when generating a population of persons, a three-dimensional matrix was generated for origin, destination, and industry (see Figure 2); however, the process is flexible in that it can accommodate higher dimensional matrices by incorporating additional socio-demographic stratification. In this case, the three-dimensional matrix is formed by three two-dimensional tables of origin by industry (OI), destination by industry (DI), and origin by destination (OD) available from the US census. The joint ODI distribution can be obtained either through IPF by treating the tables as marginals, or alternatively by calculating the conditional probabilities from the tables and using an MCMC sampler to yield the joint probability distribution. Unlike parametric OD assignment models, such as the gravity model, the proposed joint distribution for origin-destination-industry (ODI) is created using observed ODI totals from census data, meaning that the resulting matrix is already fit to observed empirical data, not an assumed model such as the gravity model.

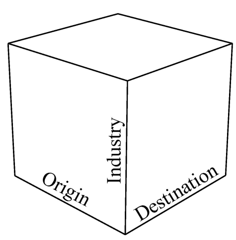


Fig. 2: Origin-destination-industry conceptualization

Once the origin-destination-industry (ODI) matrix is fitted in step (1), the joint distribution of destinations can be treated as the destination marginal or partial conditional for persons in step (2). This can be imagined on a zone-by-zone basis as taking a slice of the three-dimensional cube for each origin, yielding a joint table of workers in each destination by industry. The joint table is then used as a marginal constraint (i.e, with IPF) or partial conditional (i.e., with MCMC or BN) along with the person demographic variables (e.g., age, gender, industry, destination).

In conventional workplace assignment, the person-household population is synthesized completely separate from the home-workplace location model. In this case, workplace location is not population attribute in person synthesis, and is probabilistically assigned after the population has been synthesized.

The following three subsections describe the synthesis methods used in steps (1) and (2) with IPF, MCMC, and BN in further detail.

3.1.1 Iterative Proportional Fitting

Before any IPF step can proceed, the marginals must be checked for consistency between the origin, destinations, and person marginals (i.e., the marginal totals are equal). It is possible that the census tables will not perfectly match the home-workplace origin and destination data due to sampling error, shifts in the population over time, changes in employment, or persons that enter/leave the study region. Although the differences may be minor, IPF requires perfect consistency between marginals. The minor differences between the OD, OI, and DI marginals can be corrected by proportionally adjusting each marginal to match each other. In this case the census tables are assumed to be correct and the origin and destination tables are adjusted to match the census tables.

The adjustment process begins by treating the population marginals as the OI marginal. The OD and DI tables will be adjusted to match the census based OI table. The OD table is adjusted first to match the OI table, then the DI table is adjusted to match the OD; effectively using the OD table as a bridge between origins and destinations. At this point, non-working persons are excluded because the aggregated employment data only accounts for employed persons. Once the tables are adjusted, the missing portion of non-working persons are added back to the OD, DI, and OI tables using the original total of unemployed persons in the population marginals. Since the aggregated data reflects workplace only, the non-working person's origin (i.e., home zone) is counted as also their destination to ensure the totals are consistent. To account for trips that leave the study area, the region outside is treated as a single zone with trips being counted as going to that zone. Trips entering the study area from outside can be ignored because only trips for persons in the study area needed to be synthesized.

Once the marginals are consistent, the IPF process begins with generating the ODI joint distribution. Then once the ODI joint distribution table has been synthesized, it is then used in the second IPF process for persons as a marginal for workplace destination.

3.1.2 Markov chain Monte-Carlo Gibbs sampler

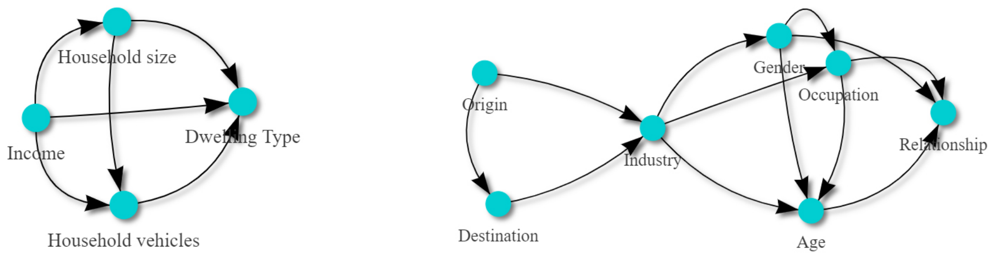
The MCMC technique used in this paper is a direct Gibbs sampler, which generates a simulated population by sequentially drawing each variable from the local conditional probabilities in a Markov chain. Eventually with a sufficiently large number of draws, the joint probability distribution will converge as the posterior joint distribution. A sufficiently large pool of 1-million random draws were generated for persons and households, respectively. However, given that there are 965 census tract zones and 14 industry sectors, the ODI distribution (965x965x14 cells) is substantially larger than the person distribution (8x2x5x6x14 cells), thus requiring a much larger number of draws (10-million) to ensure that the very small joint probabilities are captured. This is further true for the integrated person-destination distribution, which was given 100-million draws.

Full conditional probability tables for the person and household populations can be easily calculated directly from the microdata sample. However, microdata for the ODI matrix does not exist, instead partial conditionals are formed using the OD, OI, and DI tables. The resulting posterior ODI joint distribution and the calculated person conditional probabilities are then treated as partial conditional tables in a second MCMC simulation to generate the integrated posterior person-destination distribution.

Within step (2) (see Figure 1), the posterior joint distribution can be tailored to fit a desired marginal for individual census zones using Generalized Raking (Casati et al., 2015; Deville et al., 1993). Generalized Raking is functionally similar to IPF in that it adjusts the joint distribution values to satisfy marginal totals, but uses regression-like error minimization methods rather than proportional fitting. This provides fast fitting, tuning capabilities, and flexible variable handling (e.g., continuous variables), but is generally suited for subsequent calibration of a sample rather than baseline synthesis. However, this step can be avoided with the integrated process since the population is already allocated to census zones via the ODI conditional.

3.1.3 Bayesian Networks

Simulating the population with a Bayesian Network is performed similarly to a MCMC Gibbs sampler, but instead follows a Bayesian Network of partial conditionals along a directed acyclic graph (DAG). Generally there are three methods to construct a Bayesian Network: a *data-driven* approach where the structure and parameters are learned from a data set, an *expert-driven* approach where the network structure and parameters are user defined, or a combination of the two. In this paper the data-driven learning is performed using a "Tabu" search method (Glover, 1989, 1990) as part of the "bnlearn" package (Scutari, 2014). The household population network is created using an entirely data-driven learning while the person-destination population is created with a hybrid approach. The person population is first created using data-driven learning, which is then augmented using the known ODI conditionals. The Bayesian Network used for households and person-destination are shown in Figure 3. Since the Bayesian Network must be acyclic, care must be taken when constructing a custom network to avoid introducing cyclical loops in the network when adding the conditionals.



(a) Households (b) Person-destination

Fig. 3: Bayesian Networks

3.2 Joint re-weighting

To generate a multilevel joint population of households and persons in step (3) (see Figure 1), the multilevel person and household microdata sample must be re-weighted to fit the separate joint household and person-destination distributions previously created. A common re-weighting method is Iterative Proportional Updating (IPU) (Ye et al., 2009). However, IPU is computationally intensive and can require a very long time to converge when given many variables. To improve the computational efficiency of multilevel re-weighting, the re-weighting problem is recast as an optimization problem with the objective to minimize error.

The performance using several different optimization algorithms are compared against IPU, specifically a classical non-negative least squares (NNLS) algorithm, a simplex based solution to non-negative least absolute deviation (NLAD), and a fast gradient descent method.

3.2.1 Formulation

The problem can be formulated into an optimization problem by first restructuring the joint multilevel person-household microdata into a frequency table, such as the example in Table 1.

Table 1: Example joint multilevel frequency table

Joint type	Microdata person-household sample						Joint target
	x_1	x_2	x_3	x_4	x_5	x_6	b
Household Type 1	1	0	0	1	0	1	35
Household Type 2	0	1	1	0	1	0	45
Person Type 1	1	0	1	3	0	1	124
Person Type 2	1	2	3	0	1	0	137

Each column is an individual record from the person-household microdata sample and each row contains the frequency of each joint person or household type in the record. There can be multiple person types in each household record, but only one household type (i.e., only one household per household). The joint target column is the joint distribution values estimated from the separate person and household synthesis step (i.e., IPF, BN, or MCMC) that the microdata is to be re-weighted to match. From this table, the problem may be easily formulated into the familiar $Ax = b$ format, as shown in

$$\begin{array}{c} \text{Household Type 1} \\ \text{Household Type 2} \\ \text{Person Type 1} \\ \text{Person Type 2} \end{array} \begin{bmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 3 & 0 & 1 \\ 1 & 2 & 3 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} 35 \\ 45 \\ 124 \\ 137 \end{bmatrix} \tag{1}$$

where each joint sample is a decision variable a vector of x, the sample household/person type values are constraints in an A matrix, and the right hand side target values b are the separately synthesized joint person and household distributions (i.e., from IPF, MCMC, or BN). Two fundamental objective functions can then be formulated, first to minimize the least square error as

$$\min \ ||b - Ax||^2 \tag{2a}$$
$$\text{s.t. } x \geq 0 \tag{2b}$$

or alternatively to minimize the least absolute deviation as

$$\min \ |b - Ax| \tag{3a}$$
$$\text{s.t. } x \geq 0 \tag{3b}$$

with the additional non-negative boundary constraint imposed in each to prevent negative weights (there cannot be negative persons or households). The NNLS objective in Equation (2) is often solved using a well established algorithm developed by Lawson and Hanson (1995). However, given the quadratic nature of the formulation, the algorithm quickly becomes computationally inefficient and intractable for large scale problems. In contrast, NLAD in Equation (3) remains linear and can be efficiently solved using linear programming methods, such as the simplex algorithm.

Other than computation, the difference between the two formulations is that least squares will find the mean value while least absolute deviation will find the median value. This property of least absolute deviation makes it resistant to outliers and is often called "robust" regression (Bloomfield and Steiger, 1984; Davis and Dunsmuir, 1997). The two objectives functions are analogous to the variable selection technique Least Absolute Shrinkage and Selection Operator (LASSO) and ridge regression. In this space exists methods to handle regularized regression very quickly, such as a hybrid ridge-LASSO called "elastic net" which uses cyclical coordinate descent (CCD) of the likelihood function to achieve optimization (Friedman et al., 2010; Simon et al., 2011; Friedman et al., 2007).

This paper compares conventional IPU against the above optimization problem, solved with three different methods with the following implementations:

- IPU was coded as a custom *R* package in C++ by the authors to provide a competitive performance comparison.
- NNLS utilized an open source software package called "nnls", which is based on the Lawson and Hanson (1995) algorithm and is coded in Fortran (Katharine M. Mullen and Ivo H. M. van Stokkum, 2015).
- NLAD with linear programming utilizes an open source commercial grade optimization package written in C++ called "Clp", developed and maintained by Computational Infrastructure for Operations Research (COIN-OR) Foundation (2017).
- CCD utilized an open source software package called "glmnet" (Friedman et al., 2019). The tuning parameters were set with a penalty of zero to achieve pure coordinate descent optimization of the maximum likelihood function without variable selection.

The project work flow and data handling is written in *R*, but the optimization algorithms are coded as dedicated functions using the more efficient programming languages and simply executed with *R*. The project and suite of tools used will be made available in a public repository via `https://github.com/nick-fournier/poptools`. In all cases, the joint sample is stored as a sparse matrix in *R* before being passed to the respective algorithms, greatly reducing the required memory and improving overall performance for all methods.

3.3 Joint sampling

The results for all re-weighting methods are decimal weights for each joint record in the microdata. The joint weights can then be used as weighted probabilities to generate the final population with Monte-Carlo sampling in step (4) of the process (see Figure 1). This sampling process is no different than with existing methods (e.g., IPU). However, microdata typically does not contain OD information and cannot be re-weighted with respect to OD. Instead a simple two step random sampling procedure is used to generate the final population. First, joint household-persons are generated by sampling from the microdata using the new joint weights. Then from this joint sample, the destination is drawn using the person-destination IPF weights as proportional probabilities for each person given their person type. This process is effectively the same as with conventional OD assignment, the difference being in how the OD distribution is generated. The integrated OD distribution contains all person variables which were jointly fitted while the conventional OD distribution only contains industry as a stratified variable.

4 Application

The proposed method is applied to obtain a population of 4.6-million people in the Greater Boston Area (GBA) with work location incorporated as an attribute of the population. The GBA is defined by the Boston Region Metropolitan Planning Organization's (MPO) Central Transportation Planning Staff (CTPS). The GBA consists of 965 census tracts, shown in Figure 4. The following section first describes the data used for population synthesis and workplace assignment, followed by the results of this synthesis.

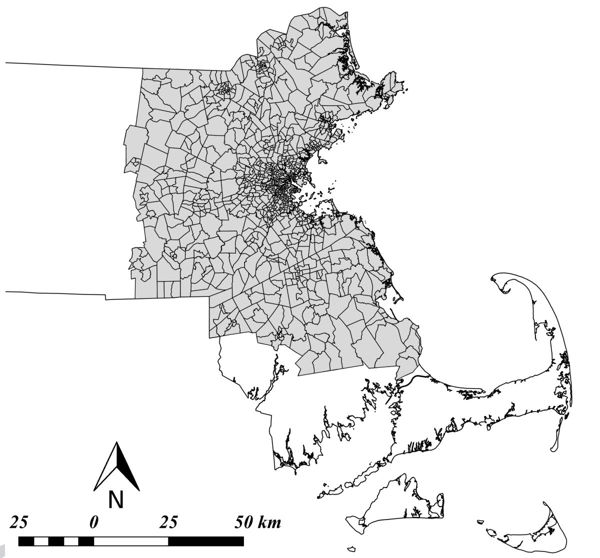


Fig. 4: Boston Metropolitan Area census tracts

4.1 Data

Data utilized in this paper consists of aggregated marginal totals, disaggregated microdata samples, and aggregated OD totals by industry. All data are publicly available from the United States Census Bureau, and are summarized in Table 2. The marginal tables are provided by the United States Census Bureau's American Community Survey (ACS) (U.S. Census Bureau, 2015). As opposed to the decennial census, which is a full census collected only every 10 years, the ACS is a program that performs ongoing data collection used to estimate adjusted tables for more recent years between decennial census years. The microdata are also managed by the ACS program of the United States Census Bureau, referred to as Public Use Microdata Samples (PUMS). The PUMS are provided as roughly a five percent sample of the households and persons in the population.

The OD totals are managed by the Center for Economic Studies of the United States Census Bureau under the Longitudinal Employer Household Dynamics (LEHD) program. This program also collects home and work locations of individuals, with origins and destinations aggregated by various stratification (e.g., industry sector), called the LEHD Origin-Destination Employment Statistics (LODES). The LODES data provides aggregated OD pair totals for census blocks in a set of data tables stratified by demographics. The demographic data stratification are provided for origins or destinations separately, not simultaneously. For

Table 2: Data used in population synthesis

Table/Dataset Name	Year	Program	Description
		Marginal data	
B19001	2015	ACS 5-year	Household income
B25124	2015	ACS 5-year	Household size and dwelling type
B08201	2015	ACS 5-year	Household size and vehicles
B09019	2015	ACS 5-year	Relationship to householder
C24050	2015	ACS 5-year	Industry & occupation
B01001	2015	ACS 5-year	Age & sex
		Microdata	
ss15pma	2011-2015	PUMS	Disaggregate persons sample
ss15hma	2011-2015	PUMS	Disaggregate households sample
		Origin-Destination data	
ma_wac_S000_JT00	2015	LODES	Workplace destination by industry
ma_rac_S000_JT00	2015	LODES	Workplace origin totals by industry
ma_od_main_JT00	2015	LODES	Workplace origin-destination totals

example, the total number of workers for each origin-destination pair are provided in one table with two separate tables stratified for origin by industry and destination by industry.

Since both the PUMS and census tables are managed and provided by the U.S. Census Bureau, they largely share the same variables and data structure, requiring very little adjustment to make them compatible. In some cases however, continuous variables (e.g., income and age) in the disaggregated PUMS needs to be binned as discrete variables to match the grouping used in the aggregated census tables. Table 3 summarizes the overall variables used for person and household synthesis for the respective home and work locations. The industries and occupations are grouped in Table 4 based the PUMS data using the 2017 North American Industry Classification System (NAICS), as reported in the U.S. Census (U.S. Census Bureau, 2010).

Table 3: Control variables

Household				***Person***				
Vehicles	Income	Dwelling	Members	Sex	Age	Relation	Industry	Occupation
0	<\$15k	1 unit	1	Male	0-9	Head	10-560	10-3540
1	\$15k-\$25k	2-4 units	2	Female	10-14	Spouse	570-760, 6070-6460	3600-4650, 9800-9830
2	\$25k-\$35k	5-19 units	3		15-19	Child	770-1060	4700-5940
≥3	\$35k-\$50k	≥20 units	≥4		20-24	Relative	1070-4060	6000-7630
	\$50k-\$75k				45-54	Non-relative	4070-4660	7700-9750
	\$75k-\$100k				55-64		4670-6060	9800-9830
	\$100k-\$150k				>65		6470-6860	0
	>\$150k						6870-7260	
							7270-7790	
							7860-8490	
							8560-8690	
							8770-9890	
							8770-9290	
							0-9, 9920-9999	

Table 4: Code grouping of the North American Industry Classification System (NAICS)

Industry sector		***Occupation***	
Code range	Description	Code range	Description
10-560	Natural resources	10-3540	Management, business, scientific, and arts
570-760, 6070-6460	Transportation and utilities	3600-4650, 9800-9830	Service
770-1060	Construction	4700-5940	Sales, office, and administration
1070-4060	Manufacturing	6000-7630	Natural resources, construction, and maintenance
4070-4660	Wholesale trade	7700-9750	Production and transportation
4670-6060	Retail trade	9920-9999	None
6470-6860	Information		
6870-7260	Finance and real-estate		
7270-7790	Professional, scientific, and management		
7860-8490	Educational and social-work		
8560-8690	Arts and accommodation		
8770-9890	Public administration or other		
0-9, 9920-9999	None		

5 Results

The results are described in three subsections: joint re-weighting method comparison, multilevel person-household population generation results, and workplace assignment results. Joint re-weighting results present the accuracy and computational performance comparison between IPU, NNLS, NLAD, and CCD when performed for a single zone. The subsequent sections then demonstrate the final population and workplace assignment results using the CCD re-weighting method. A full population was not generated using all re-weighting methods due to the excessive computation time required to synthesize all 965 census tracts with the other methods. A comparison between the conventional and integrated workplace assignment is presented, but for clarity only the IPF based generation is presented graphically, with the other synthesis methods (i.e., MCMC and BN) being presented in a summary table in the final subsection.

The results are validated using Root Mean Square Error (RMSE) and Root Mean Square Normalized Error (RMSN). RMSE is calculated as

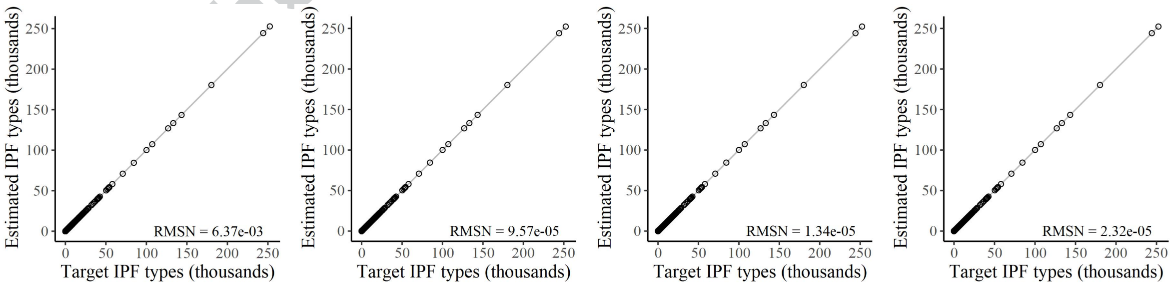
$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(\hat{b_i} - b_i)^2}{n}} \tag{4}$$

where n is the number of values being compared, $\hat{b}_i$ is the estimated value of variable i, and b_i is the actual values. For example, b_i is the frequency of person or household type i. A good fit will yield a small RMSE. However, since the following comparisons contain a wide range of values (e.g., between tracts, total region, and ODs) a normalized RMSE value is used in order to make the errors more comparable across tests. A commonly used alternative is to normalize the RMSE value by the mean $\bar{b}$ to account for relative error between differently sized values, further calculated as

$$RMSN = \frac{RMSE}{\bar{b}} \tag{5}$$

5.1 Joint re-weighting results

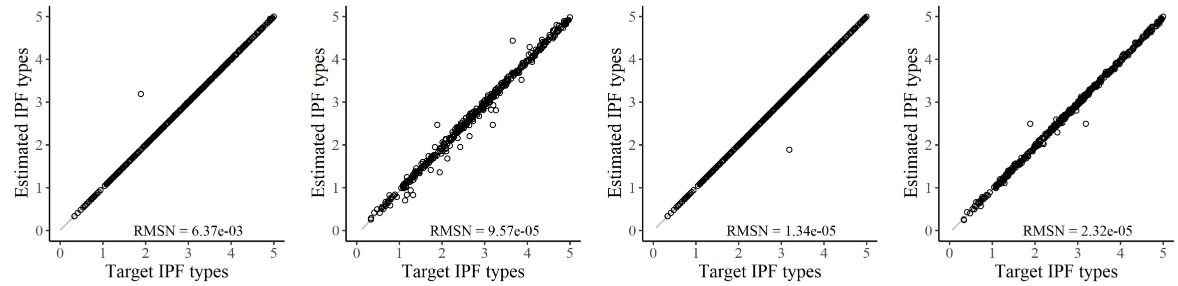
As a general comparison of fitting accuracy, persons and households are jointly re-weighted using the four re-weighting methods of (1) NNLS, (2) NLAD, (3) CCD, and (4) IPU for the entire Greater Boston Area treated as a single zone. Figure 5 is a comparison of the fit results for the methods. The target values for the separately synthesized persons and households (i.e., the b values) are shown on the horizontal axes and the vertical axes are the fit results when the joint weights are multiplied by the joint sample matrix (i.e., the Ax result). A good fit will be along the diagonal, meaning that the correct number of both persons and households are fitted when $Ax = b$ is evaluated.



(a) IPU (b) NNLS (c) NLAD (d) CCD

Fig. 5: Comparison of fit by method (1:1 scale)

Note that the weights at this point are decimal values, which is why the results are near perfect. Error will be introduced when weights are sampled as discrete persons and households, but as a measure of fitting performance that fact is irrelevant at this point. Overall the results appear near identical, with only a minor difference in the calculated RMSN. However, some interesting insights emerge upon closer inspection at 50,000:1 scaled zoom (see Figure 6). At this scale the underlying properties begin to emerge with NLAD tends to fit either perfectly or poorly, but NNLS and CCD tend to yield a small yet consistent variation. Meanwhile IPU lies somewhere in between, yielding very small consistent variation but also some outliers.



(a) IPU (b) NNLS (c) NLAD (d) CCD

Fig. 6: Comparison of fit by method (50,000:1 scale)

All methods achieved an excellent fit results between 3.17×10^{-2} to 1.34×10^{-5} RMSN. While the simplex algorithm for NLAD will take finite steps to reach a solution, NNLS, IPU, and CCD merely need to reach a specified error tolerance threshold. Thus it is possible to achieve better or worse results depending upon the threshold set by the users. However, the important distinction is the time it takes for each method to reach a similar level of accuracy, as summarized in Table 5.

Table 5: Computation time comparison of re-weighting methods

Method	RMSN	Computation time
NNLS [Lawson-Hanson algorithm]	$9.57\times^{-5}$	17.9 hours
IPU	$3.17\times^{-2}$	13.5 minutes
NLAD [simplex algorithm]	$1.34\times^{-5}$	1.6 minutes
CCD	$2.32\times^{-5}$	51.5 seconds

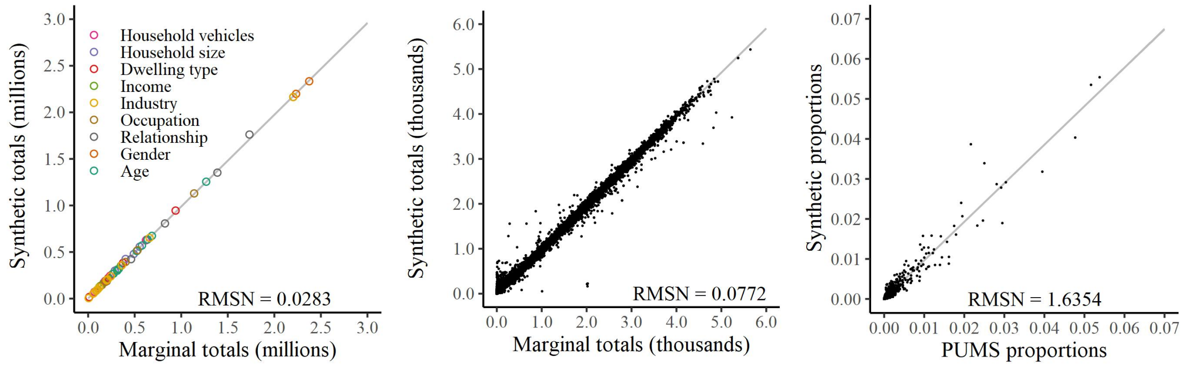
It is clear from the comparison in Table 5 that CCD achieved the best results in computation time. Although NLAD managed to achieve a slightly higher level of accuracy in this case, it required nearly twice as long. When extrapolated over the 965 census tracts, this additional computation time becomes very large. For example, approximately 13 hours for CCD, 25 hours for NLAD, 9 days for IPU, and 2 years for Lawson-Hanson NNLS. Although this time was cut down through parallel processing, it was still intractable to generate a full multilevel synthetic population for all 965 census tracts using all methods and would provide relatively little or no improvement. The following full generation results are done using only the CCD method, regardless of population synthesis methods (i.e., IPF, MCMC, and BN).

5.2 Multilevel person-household population generation results

The population validation is compared from three perspectives: marginal totals for the region, marginal totals for each census tract, and the cell level microdata proportions. The marginal comparisons measure how well the aggregated variable totals in the synthetic population fit the actual census totals. The cell level validation compares the individual combinatorial person and household type frequencies between the synthetic population and the PUMS microdata sample. A cell level validation helps ensure that the actual individual person and household types (i.e., joint distribution) are properly synthesized and not just matching the marginal totals. In general, when a microdata sample is adjusted to match the marginal totals it will no longer fit the original microdata sample. For example, a disaggregate population can be perfectly synthesized to match the microdata sample using Bayesian Networks, but will fall out of fit if it is then raked to fit marginal totals in individual zones. The challenge in population synthesis is expanding the sample to match the marginals without destroying too much of the original population's structure.

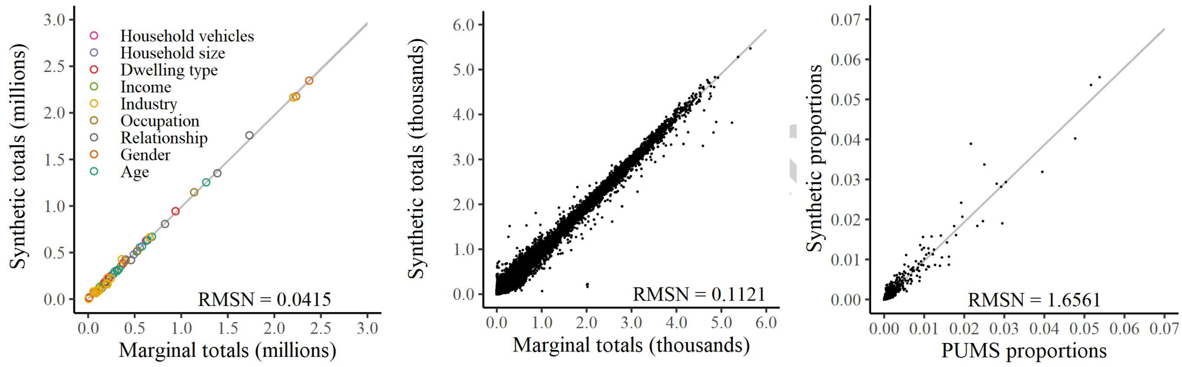
Validation of the final multilevel synthetic population is performed twice, first when using a conventional workplace assignment (shown in Figure 7) and then again with the integrated workplace assignment (shown in Figure 8). This is done in order to show any impact that the integrated assignment may have on synthesis. The marginals can be validated in absolute numbers, meaning whole integer frequencies, but the PUMS is only a sample, thus the comparison must be performed as proportions. These comparisons in Figures 7 and 8 show the final realized population results (i.e., not just weighted fit) on the vertical axes, against the expected census totals shown on the horizontal axes.

The marginal validation is shown at two scales. First for the entire aggregated GBA, achieving an RMSN of 0.0283 for conventional workplace assignment and 0.415 for integrated assignment (see Figures 7a and



(a) Marginal fit for entire GBA (b) Marginal fit for tracts (c) Cell level validation

Fig. 7: Population generation results with conventional workplace assignment



(a) Marginal fit for entire GBA (b) Marginal fit for tracts (c) Cell level validation

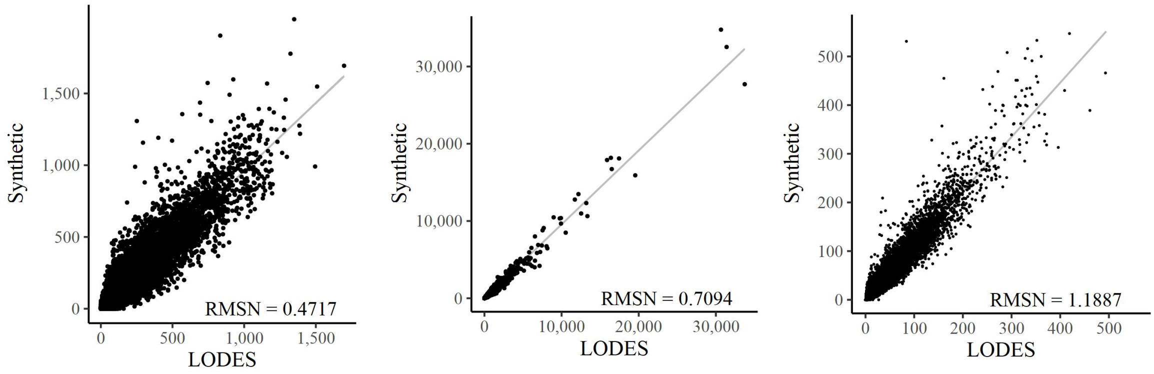
Fig. 8: Population generation results with integrated workplace assignment

8a). Then at the tract level where variables are accounted for each tract separately, achieving an RMSN of 0.0772 for the conventional assignment and 0.1121 for integrated assignment (see Figures 7b and 8b). The cell level comparison achieved an RMSN of 1.6354 for conventional assignment and 1.6561 for integrated assignment (see Figures 7c and 8c). It is clear that a possible gain in workplace assignment accuracy can come at the expense of person level accuracy with the integrated approach, particularly at the marginal census tract level.

5.3 Origin-destination results

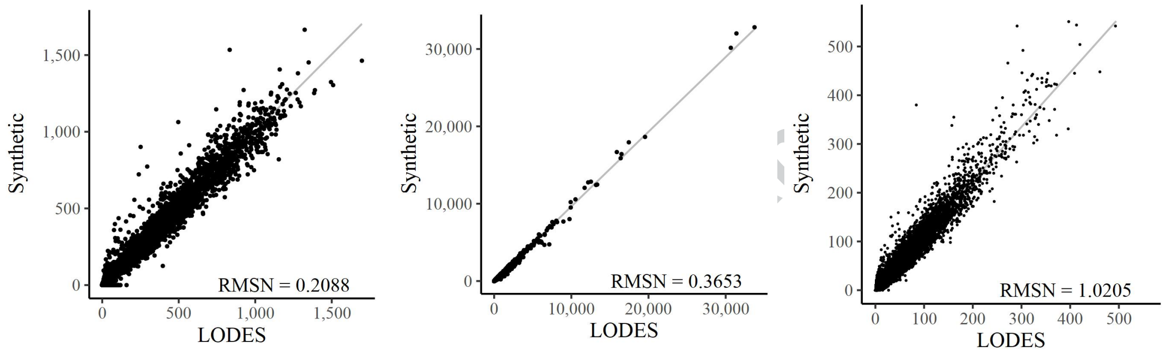
Results up to this point only considered demographic variables, not workplace assignment. A final check is to cross-validate the allocation of synthesized persons to origins and destinations using the LODES origin-industry, destination-industry, and origin-destination tables. This is performed at the aggregated level by comparing the aggregated totals in the synthetic population to the actual totals in the LODES marginals. This comparison is similar to the validation for the synthetic population, but can only be performed at the aggregated level because OD microdata at this fine grain resolution is not available.

Similarly with the demographic population, the workplace assignment validation is performed twice, once for conventional workplace assignment (see Figure 9) and again for integrated workplace assignment (see Figure 10). The reason that the figures are plotted on different scales is due to the variation between origin and destination totals. This is a byproduct of census tracts being delineated roughly by population size, but not by employment size; in other words, while residential location is dispersed fairly evenly, it is likely that certain census tracts (e.g., downtown) will attract a high concentration of workers and others very few. Although RMSN is normalized for comparison across different RMSN calculations, it does not normalize between values. This means that large outliers can dominate an RMSN result in an uneven distribution since the absolute difference between large values is greater than smaller values. For example, in Figures 9b and 10b three points in the upper right corner appear to be very dense employment destinations.



(a) Origin-industry allocation (b) Destination-industry allocation (c) Origin-destination allocation

Fig. 9: Conventional workplace assignment results



(a) Origin-industry allocation (b) Destination-industry allocation (c) Origin-destination allocation

Fig. 10: Integrated workplace assignment results

In general, the workplace assignment results improved with the integrated approach over conventional assignment. The RMSN of the workplace assignment reduced from 0.472, 0.709, and 1.189 in the conventional assignment to 0.209, 0.365, and 1.021 in the integrated assignment. This trend is a reversal of what occurred for the demographic variables. There still appears to be a substantial amount of points dispersion in the origin-industry and origin-destination case compared to the destination-industry case. It is likely that this error is a result of discrepancies between the census population and LODES tables (i.e., that the industry totals in LODES do not perfect match the industry totals in the census). The overall results when compared using other population synthesis methods of MCMC and BN are presented in Table 6.

Table 6: Summary of RMSN results for all population generation and workplace assignment methods

Method		Marginal totals	Person-household microdata	Person microdata	Household microdata	Origin by industry	Destination by industry	Origin by destination
IPF	Conventional	0.077	1.635	1.493	0.934	0.475	0.656	1.318
	Integrated	0.112	1.656	1.569	0.935	0.209	0.336	1.127
MCMC	Conventional	0.134	1.468	1.504	0.804	0.461	0.692	1.112
	Integrated	0.313	1.239	0.605	0.781	0.372	0.974	1.101
BN	Conventional	0.137	1.459	1.465	0.807	0.460	0.690	1.105
	Integrated	0.381	1.781	2.492	0.781	0.365	1.198	1.178

In general, the results are relatively comparable to each other for all synthesis methods with trade-offs depending upon the target measure. For example, integrated IPF yielded a substantial improvement in workplace assignment while MCMC and BN achieved a modest or worse fit with an integrated approach. The reason for this is uncertain, but it is possible that the sparse discrete origin-destination tables contain many local optima, or difficult to reach optima, that cause a Markov chain in MCMC or BN to become stuck in a local optima or not fully converge. Further research in this area is necessary as BN and MCMC methods possess the ability to provide superior accuracy and greater flexibility than IPF.

6 Conclusions

As travel demand models shift towards pure activity-based models, workplace assignment is still an important input for activity-generation in state-of-the-art microscopic travel demand models. For example, many travel related activities take place in conjunction with work trips, such as shopping trips on the way home from work or picking up school-age children. Although discrete choice spatial models are possible to use, aggregated employment data is often readily available at a higher spatial resolution than in disaggregated samples, making the use of classically fit models attractive. This paper presents and applies an integrated population synthesis and workplace assignment method using aggregated employment data and an efficient person-housing matching method based on non-negative least deviation fitting. Such an integrated approach can be easily integrated in current common practice in existing models in the United States and elsewhere. The specific application described in this paper synthesized a population of 4.6-million people and 1.7-million households in the Greater Boston Area, which is ultimately utilized for an energy assessment simulation of an activity-based demand and multi-modal supply simulation (Fournier et al., 2018). The resulting population achieved an overall marginal level fit RMSN of 0.0415, 0.112 at the census tract level, and a microdata cell level fit RMSN of 1.656. While the integrated assignment approach resulted in a slight loss of population accuracy, it yielded an improved workplace assignment fit over conventional assignment with an RMSN of 0.209, 0.365, and 1.021 for origin by industry, destination by industry, and origin by destination, respectively.

The overall application for the population synthesis, workplace assignment, and person-household matching achieved good fit results. However, there are several areas of potential refinement. An immediate area of improvement is to investigate and resolve the noticeable error dispersion among the less frequent persons and household types incurred with the integrated assignment process (see Figure 7b). A second area worth further investigation is the impact of using an optimization based re-weighting approach (i.e., NLAD), as opposed to traditional proportional fitting (i.e., IPU). Where the outlier resistant property of NLAD is useful in variable selection (e.g., LASSO), it is uncertain whether this property is beneficial or harmful in population synthesis. It could mean that redundant or duplicate person-households records are ignored, or that person-household heterogeneity may be reduced in the population.

Another obvious area of future improvement is to incorporate additional stratification variables other than industry (e.g., age and gender). This is likely to improve the workplace assignment by providing additional constraints during the fitting process. Additional stratification variables are likely to improve results for the BN and MCMC approaches as well, which currently rely entirely upon a single variable to link workplace assignment and population variables, as shown in Figure 3. Any error in this linkage will propagate throughout the population when the sampler traverses the network during generation. Additional linking variables may help resolve the accuracy issues encountered with the BN and MCMC approaches.

A final proposed future research area, and possibly farther reaching, is to smooth the very fine grain discrete LODES data (i.e., small census blocks) into smooth continuous Cartesian coordinates (e.g., latitude and longitude or geographic projections) using kernal density estimation. Such a process when coupled with flexible probabilistic methods (e.g., BN or MCMC) would obviate the need for cumbersome zone-by-zone estimation, thus yielding a zoneless synthetic population allocated to home and work locations stored as continuous coordinates. This would be beneficial computationally in reducing generation to a single zone, but is also likely to improve accuracy as well because a single large zone is less susceptible to local survey error and heterogeneity loss than many small census zones fitted individually.

The proposed integrated process makes two contributions. First it integrates population synthesis and workplace assignment for improved workplace allocation. This minimizes errors that would be introduced through independently estimated models. Second, this paper introduces an efficient optimization based approach to multilevel joint person-household re-weighting, substantially reducing computation time compared to the conventional iterative proportional updating (IPU) method. This new re-weighting approach makes the integrated process more feasible by being able to efficiently handle additional shared attributes in the population and workplace data (e.g., employment).

Acknowledgments

This research was funded in part by the US DOE's Advanced Research Projects Agency-Energy (ARPA-E) under the Traveler Response Architecture using Novel Signaling for Network Efficiency in Transportation

(TRANSNET) program, with Award Number DE-AR0000611. On behalf of all authors, the corresponding author states that there is no conflict of interest.

Authors' contributions

N. Fournier: Literature review, manuscript writing, methodological development and analysis. E. Christofa: Methodological guidance, content planning, and manuscript editing. A. Akkinepally: Methodological guidance, interpretation of results, and manuscript editing. C. Azevedo: Methodological guidance, literature review, and manuscript editing.

References

Abdel-Aal MMM (2014) Calibrating a trip distribution gravity model stratified by the trip purposes for the city of Alexandria. Alexandria Engineering Journal 53(3):677–689

Abraham JE, Stefan KJ, Hunt JD (2012) Population synthesis using combinatorial optimization at multiple levels. Transportation Research Record 17

Adnan M, Pereira FC, Azevedo CML, Basak K, Lovric M, Raveau S, Zhu Y, Ferreira J, Zegras C, Ben-Akiva M (2016) SimMobility: A Multi-scale Integrated Agent-based Simulation Platform. In: Transportation Research Board 95th Annual Meeting, Transportation Research Board, p 18

Anda C, Ordonez Medina SA, Fourie P (2018) Multi-agent urban transport simulations using OD matrices from mobile phone data. Procedia Computer Science 130:803–809

Arentze T, Timmermans H, Hofman F (2007) Creating Synthetic Household Populations: Problems and Approach. Transportation Research Record: Journal of the Transportation Research Board 2014:85–91

Arentze TA, Timmermans HJ (2004) A learning-based transportation oriented simulation system. Transportation Research Part B: Methodological 38(7):613–633

Auld J, Mohammadian A (2010) Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels. Transportation Research Record: Journal of the Transportation Research Board 2175(1):138–147

Bachir D, Khodabandelou G, Gauthier V, El Yacoubi M, Puchinger J (2019) Inferring dynamic origin-destination flows by transport mode using mobile phone data. Transportation Research Part C: Emerging Technologies 101:254–275

Ballas D, Clarke G, Dorling D, Eyre H, Thomas B, Rossiter D (2005a) SimBritain: A spatial microsimulation approach to population dynamics. Population, Space and Place 11(1):13–34

Ballas D, Clarke GP, Wiemers E (2005b) Building a dynamic spatial microsimulation model for Ireland. Population, Space and Place 11(3):157–172

Balmer M, Rieser M, Meister K, Charypar D, Lefebvre N, Nagel K (2009) MATSim-T: Architecture and simulation times. In: Multi-agent systems for traffic and transportation engineering, IGI Global, pp 57–78

Barthelemy J, Toint PL (2013) Synthetic population generation without a sample. Transportation Science 47(2):266–279

Bassolas A, Ramasco JJ, Herranz R, Cantú-Ros OG (2019) Mobile phone records to feed activity-based travel demand models: MATSim for studying a cordon toll policy in Barcelona. Transportation Research Part A: Policy and Practice 121(January):56–74

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

Ben-Akiva ME, Lerman SR (1985) Discrete choice analysis: theory and application to travel demand, vol 9. MIT press

Bloomfield P, Steiger WL (1984) Least Absolute Deviations. Birkhäuser Boston, Boston, MA

Borysov SS, Rich J, Pereira FC (2019) How to generate micro-agents? A deep generative modeling approach to population synthesis. Transportation Research Part C: Emerging Technologies 106:73–97

Bowman J, Ben-Akiva M (2001) Activity-based disaggregate travel demand model system with activity schedules. Transportation Research Part A: Policy and Practice 35(1):1–28

Bowman JL, Bradley M, Shiftan Y, Lawton TK, Ben-Akiva ME (1998) Demonstration of an activity based model system for Portland. In: 8th World Conference on Transport Research, Antwerp, Belgium

Bowman JL, Bradley M, Castiglione J, Yoder SL (2014) Making advanced travel forecasting models affordable through model transferability. Tech. rep., Bowman Research and Consulting, URL http://jbowman.net

Briem L, Mallig N, Vortisch P (2019) Creating an integrated agent-based travel demand model by combining mobiTopp and MATSim. Procedia Computer Science 151:776–781

Casati D, Müller K, Fourie PJ, Erath A, Axhausen KW (2015) Synthetic Population Generation by Combining a Hierarchical, Simulation-Based Approach with Reweighting by Generalized Raking. Transportation Research Record: Journal of the Transportation Research Board 2493:107–116

Choupani AA, Mamdoohi AR (2016) Population Synthesis Using Iterative Proportional Fitting (IPF): A Review and Future Research. Transportation Research Procedia 17:223–233

Computational Infrastructure for Operations Research (COIN-OR) Foundation (2017) Clp. URL `https://www.coin-or.org/`

Davis RA, Dunsmuir WTM (1997) Least Absolute Deviation Estimation for Regression with ARMA Errors. Journal of Theoretical Probability 10(2):481–497

Deming WE, Stephan FF, Frederick F Stephan (1940) On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. The Annals of Mathematical Statistics 11(4):427–444

Deville JC, Särndal CE, Sautory O (1993) Generalized Raking Procedures in Survey Sampling. Journal of the American Statistical Association 88(423):1013–1020

Dong X, Ben-Akiva ME, Bowman JL, Walker JL (2006) Moving from trip-based to activity-based measures of accessibility. Transportation Research Part A: Policy and Practice 40(2):163–180

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

Fournier N, Chen S, Needell Z, Lima IVD, Deliali K, Araldo A, Prakash AA, Azevedo CML, Christofa E, Trancik J, Ben-Akiva M, Akkinepally A (2018) Integrated Simulation of Activity-Based Demand and Multi-Modal Dynamic Supply Simulation for Energy Assessment. In: 21st IEEE International Conference on Intelligent Transportation Systems

Friedman J, Hastie T, Höfling H, Tibshirani R (2007) Pathwise coordinate optimization. The Annals of Applied Statistics 1(2):302–332

Friedman J, Hastie T, Tibshirani R (2010) Regularization paths for generalized linear models via coordinate descent. Journal of statistical software 33(1):1

Friedman J, Hastie T, Tibshirani R, Simon N, Narasimhan B, Qian J (2019) glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. URL `https://cran.r-project.org/package=glmnet`

Glover F (1989) Tabu Search—Part I. ORSA Journal on Computing 1(3):190–206

Glover F (1990) Tabu Search—Part II. ORSA Journal on Computing 2(1):4–32

Guevara CA (2010) Endogeneity and sampling of alternatives in spatial choice models. PhD thesis, Massachusetts Institute of Technology

Guo J, Bhat C (2007) Population synthesis for microsimulating travel behavior. Transportation Research Record: Journal of the Transportation Research Board 2014(2014):92–101

Hermes K, Poulsen M (2012) A review of current methods to generate synthetic spatial microdata using reweighting and future directions. Computers, Environment and Urban Systems 36(4):281–290

Huang Z, Ling X, Wang P, Zhang F, Mao Y, Lin T, Wang FY (2018) Modeling real-time human mobility based on mobile phone and transportation data fusion. Transportation Research Part C: Emerging Technologies 96:251–269

Ireland CT, Kullback S (1968) Contingency tables with given marginals. Biometrika 55(1):179–188

Katharine M Mullen, Ivo H M van Stokkum (2015) The Lawson-Hanson algorithm for non-negative least squares. URL `https://cran.r-project.org/web/packages/nnls/nnls.pdf`

Lawson CL, Hanson RJ (1995) Solving Least Squares Problems. Society for Industrial and Applied Mathematics

Le Dt, Cernicchiaro G, Zegras C, Ferreira J (2016) Constructing a Synthetic Population of Establishments for the Simmobility Microsimulation Platform. Transportation Research Procedia 19:81–93

Li M, Gao S, Lu F, Zhang H (2019) Reconstruction of human movement trajectories from large-scale low-frequency mobile phone data. Computers, Environment and Urban Systems 77:101346

Lovelace R, Ballas D (2013) Truncate, replicate, sample: A method for creating integer weights for spatial microsimulation. Computers, Environment and Urban Systems 41:1–11

Lovelace R, Dumont M (2016) Spatial microsimulation with R, 1st edn. CRC Press

Lovelace R, Ballas D, Watson M (2014) A spatial microsimulation approach for the analysis of commuter patterns: from individual to regional levels. Journal of Transport Geography 34:282–296

Martinez F, Donoso P (2010) The MUSSA II land use auction equilibrium model. In: Residential Location Choice, Springer, pp 99–113

McFadden D (1978) Modelling the choice of residential location. Spatial Interaction Theory and Planning Models 673(477):75–96

Mosteller F (1968) Association and Estimation in Contingency Tables. Journal of the American Statistical Association 63(321):1

Nakanishi W, Yamaguchi H, Fukuda D (2018) Feature Extraction of Inter-Region Travel Pattern Using Random Matrix Theory and Mobile Phone Location Data. Transportation Research Procedia 34:115–

122
Openshaw S, Rao L (1995) Algorithms for reengineering 1991 Census geography. Environment and planning A 27(3):425–446
Pritchard DR, Miller EJ (2012) Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. Transportation 39(3):685–704
Recker WW (2001) A bridge between travel demand modeling and activity-based travel analysis. Transportation Research Part B: Methodological 35(5):481–506
Saadi I, Mustafa A, Teller J, Farooq B, Cools M (2016) Hidden Markov Model-based population synthesis. Transportation Research Part B: Methodological 90:1–21
Salvini P, Miller EJ (2005) ILUTE: An Operational Prototype of a Comprehensive Microsimulation Model of Urban Systems. Networks and Spatial Economics 5(2):217–234
Scutari M (2014) Bayesian Network Constraint-Based Structure Learning Algorithms: Parallel and Optimised Implementations in the bnlearn R Package. Journal of Statistical Software 77(1), 1406.7648
Simini F, González MC, Maritan A, Barabási AL (2012) A universal model for mobility and migration patterns. Nature 484(7392):96–100
Simon N, Friedman J, Hastie T, Tibshirani R (2011) Regularization paths for Cox's proportional hazards model via coordinate descent. Journal of statistical software 39(5):1
Stephan FF (1942) An Iterative Method of Adjusting Sample Frequency Tables When Expected Marginal Totals are Known. The Annals of Mathematical Statistics 13(2):166–178
Stouffer SA (1940) Intervening opportunities: a theory relating mobility and distance. American sociological review 5(6):845–867
Sun L, Erath A (2015) A Bayesian network approach for population synthesis. Transportation Research Part C: Emerging Technologies 61:49–62
Sun L, Erath A, Cai M (2018) A hierarchical mixture modeling framework for population synthesis. Transportation Research Part B: Methodological 114:199–212, URL https://www.sciencedirect.com/science/article/pii/S0191261517308615
Train K (1986) Qualitative choice analysis: Theory, econometrics, and an application to automobile demand, vol 10. MIT press
US Census Bureau (2010) 2010 Decennial Census Tables. URL https://www.census.gov/data.html
US Census Bureau (2015) 5-year American Community Survey Tables. URL https://www.census.gov/data.html
US Census Bureau American Community Survey (2015) 2011-2015 ACS 5-year PUMS. URL https://www.census.gov/data.html
Voas D, Williamson P (2000) An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. Population, Space and Place 6(5):349–366
Voorhees AM (1956) A general theory of traffic movement. Transportation 40(6):1105–1116
Waddell P (2002) UrbanSim: Modeling Urban Development for Land Use, Transportation, and Environmental Planning. Journal of the American Planning Association 68(3):297–314
Wagner P, Wegener M (2007) Urban land use, transport and environment models: Experiences with an integrated microscopic approach. disP-The Planning Review 43(170):45–56
Wilson AG (2011) Entropy in urban and regional modelling, vol 1. Routledge
Wong DWS (1992) The Reliability of Using the Iterative Proportional Fitting Procedure. The Professional Geographer 44(3):340–348
Ye X, Konduri K, Pendyala RM, Sana B, Waddell P (2009) A methodology to match distributions of both household and person attributes in the generation of synthetic populations. In: 88th Annual Meeting of the Transportation Research Board, Washington, DC
Zhang D, Cao J, Feygin S, Tang D, Shen ZJ, Pozdnoukhov A (2019) Connected population synthesis for transportation simulation. Transportation Research Part C: Emerging Technologies 103:1–16
Zhu Y, Ferreira J (2014) Synthetic Population Generation at Disaggregated Spatial Scales for Land Use and Transportation Microsimulation. Transportation Research Record: Journal of the Transportation Research Board 2429:168–177