Available online at www.sciencedirect.com

ScienceDirect

Transportation Research Procedia 00 (2018) 000–000

Transportation Research Procedia

www.elsevier.com/locate/procedia

World Conference on Transport Research - WCTR 2019 Mumbai 26-31 May 2019

Practical population synthesis enhancements to support advanced and activity-based travel demand models

Ramachandran Balakrishna[a,*], Srinivasan Sundaram[a]

[a]*Caliper Corporation, 1172 Beacon Street, Suite 300, Newton, MA 02461, USA*

Abstract

An enumerated population is a key input to advanced travel demand models of the tour-based, activity-based and hybrid genres. A synthetic population allows the model to be sensitive to person-level behavioral heterogeneity and facilitates the use of demographic and other variables (such as gender) that are otherwise not intuitive in aggregate models. Synthesis assumes the availability of a seed sample of households and the persons living therein, estimates the number of each household type per geographic unit, and samples from the seed according to certain weights. Such a population must however reflect the distributions of key variables in the study region, as encapsulated by marginals collected at the household and person levels. Popular synthesis techniques have generally focused on matching household marginals but do not explicitly control for person-level distributions. While attempts have been made to extend these methods to simultaneously fit person marginals, the results appear to be experimental, create more data-related problems, and take a long time to run on large-scale, practical models. In this paper, we review the state of the art and state of the practice of population synthesis methods, identify the key limitations, and propose simple techniques to overcome the same. We also demonstrate the novel use of third-party data sources to correct errors in the marginals. The enhanced approach is applied on two large, real-world examples in the USA: Las Vegas, Nevada and the Central Coast, California. Empirical evidence supports the significant improvement in the quality of the synthesized population.

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY.

Keywords: Population synthesis; Activity-Based Models; Iterative Proportional Updating; Person marginals

1. Introduction

A crucial requirement for advanced travel demand models is a robust population synthesis. This step creates a full but synthetic enumeration of the study region's population in such a way that they sum up to measured aggregate totals at various levels of geography. Examples of such variables include the number of households by size, vehicle ownership and income; and the number of persons by age and gender. These totals, also called marginals, may be drawn from various standard data sources such as the Census or the American Community Survey (ACS). Advanced models can then predict the activity participation and/or the travel demand generated by each household or person in

Peer-review under responsibility of WORLD CONFERENCE ON TRANSPORT RESEARCH SOCIETY

this population, which can allow for models that are more sensitive to the heterogeneity between seemingly similar households. It can also facilitate the inclusion of person-specific variables such as gender and driver license ownership. All of these effects are difficult if not impossible to capture in the aggregate, trip-based approach deployed to date.

1.1. Traditional population synthesis approach

Most popular population synthesis methods adopt a framework such as that shown in Fig. 1. They start with a sample of household (HH) and person records, typically obtained from a survey of the study region, and tagged to a high-level geography such as zones. An Iterative Proportional Fitting (IPF) step then estimates the number of each type of household (identified in the survey) to be generated for each of the sub-zones (e.g. blocks) contained within a zone. This is achieved by identifying the number of households of each type that will add up to various household marginals at the sub-zone level. The term "Nested" indicates that the household marginals may themselves be specified at different (but nested) geographies: for example, the auto ownership marginals may be at the zone level while the income marginals may be at the blocks. Nested synthesis will drill all the marginals down to a common geography (usually the smallest/finest among all input geographies).

Finally, the predicted number of households in each sub-zone are sampled from the surveyed households using externally-supplied (initial) weights. All persons living within a chosen household are automatically copied into an output file, which eventually becomes the synthetic population.

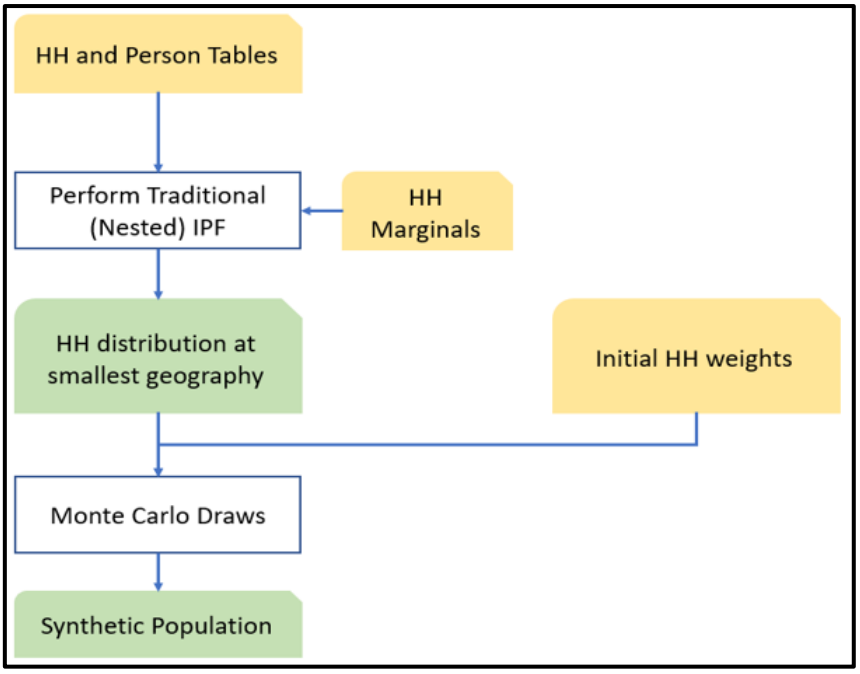

Fig. 1. General population synthesis process flow

IPF provides a very simple mathematical formulation to the above problem and can be easily implemented in a wide array of software tools and scripts. However, the above framework has serious limitations. The most important drawback is the lack of control on person marginals. Since the algorithm only controls for household variables, the obtained distribution of person attributes in any given sub-zone is largely a lottery. There is no guarantee that the chosen households will yield, say, the correct distribution of age, gender, etc. in the synthetic population.

An additional critique of the above approach is the common set of sampling weights used for all zones (and sub-zones) in the region. Since all sub-zones must pick from the same pool of candidate survey households falling within their parent zone, the Monte Carlo step assumes identical distributions of those household types in all sub-

zones. This is likely to be highly erroneous in practice, and a scheme that allows the weights to vary across sub-zones must be preferred over the default approach relying on initial weights alone.

Since advanced demand models are based on numerous disaggregate choice models that make heavy use of demographic variables, errors in person attributes can have a significant impact on the accuracy of the model's predictions of activity patterns and travel levels by mode and purpose. For instance, incorrect forecasts of the number of households with working-age adults or school-age children will result in biased estimates of mandatory travel related to work, school and child-care. It is therefore desirable to include person marginals in the population synthesis process.

2. Literature Review

The need for person marginals in the synthesis has been acknowledged in the literature, and several ideas have bee presented to fill this gap. Proposed solutions typically fall within two broad categories: those based on mathematical optimization frameworks, and those extending the IPF philosophy to include person variables in addition to household variables. While detailed reviews of population synthesis methods may be found elsewhere (Muller and Axhausen, 2011; Barthelemy and Johan, 2012), we present brief ideas of some of the relevant papers below.

An entropy-maximizing objective function was proposed by Vovsha et al. (2015), in which the household weights are adjusted in a "list-balancing" framework. First, the household weights are adjusted to improve the fit to both household and person marginals while remaining as close as possible to the initial weights. Subsequently, utilities are executed to systematically round any fractional weights. A slightly modified version of the above algorithm is deployed by Paul et al. (2018).

Both of the above methods tend to localize the estimated household weights around the initial input values, which can be sub-optimal given that the initial weights are often of unknown vintage.

Iterative Proportional Updating (IPU) has been proposed as a way of handling person marginals while essentially retaining the simplicity of the IPF framework (Xin et al., 2009). Here, each household's initial weight is adjusted while explicitly considering its contribution to household and person marginals. For example, households with children below age 5 years will have their weights adjusted to better match the total number of such children predicted to live in a given sub-zone. Other households (which do not feature such children) will not be impacted during this adjustment. Since there can be many marginals to be matched, the weights are adjusted sequentially (i.e. one marginal at a time), working with only the relevant subset of households each time. After all marginals have been treated in this manner, the procedure returns to the first marginal and continues to iterate until convergence (as defined by the modeler) is reached.

3. Proposed methodology

The adopted technical approach is illustrated by the flowchart in Fig. 2, the difference from Fig. 1 being the introduction of the IPU step which modified the initial weights for each sub-zone independently to provide more flexibility in matching both household and person marginals across the study region.

3.1 IPU enhancements

Our proposed IPU, however, deviates from that published in the literature. Xin et al. (2009) starts with a seed table that treats the incidence of each record against every possible **combination** of variable levels across both households and persons. If household auto ownership is divided into four levels (0, 1, 2, 3+), household income is split into three levels (low, medium, high), person age is divided into five categories (0-5, 5-18, 18-25, 25-60, 60+), and person gender is divided into two levels (male, female), this gives rise to 4*3*5*2 = 120 columns. A typical survey would thus have its records spread extremely thinly across these many columns. It also bears remembering that synthesis applications in practice may employ more variables, which only exacerbates the problem.

The published work on IPU mentions the above issue along with its consequence, the zero-cell problem, in which many columns are largely filled with zeroes and hence causes the IPU to either struggle or fail. In our adaptation, this

problem is largely alleviated by only looking to match the marginals of individual variable levels and not their exhaustively enumerated combinations. The example above would thus involve only 4+3+5+2 = 14 columns, which gives the IPU a significantly better chance at converging. An added attraction is the greatly reduced running time, which is orders of magnitude lower than that expected from the previously published IPU approach.

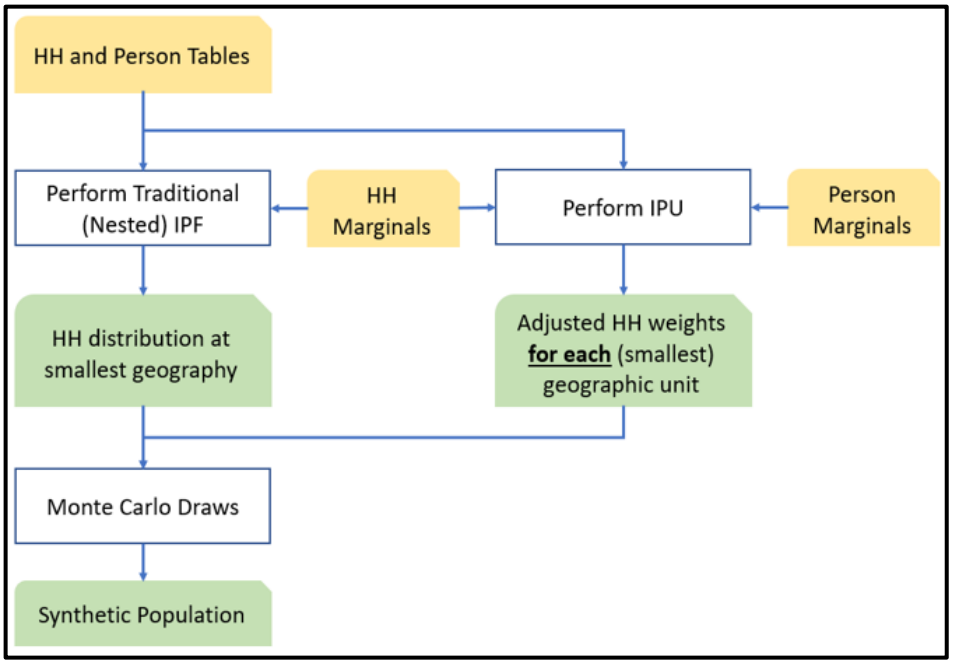

Fig. 2. Enhanced population synthesis with integrated IPU

4. Case studies

We tested our methodology on two real-world datasets corresponding to locations that are actively exploring advanced travel demand models with disaggregate components. The first is the Regional Transportation Commission (RTC) in charge of modeling for the Las Vegas, Nevada region. The other is the Central Coast region of California, comprised of three planning agencies for Monterey Bay, Santa Barbara and San Luis Obispo. While the former is investigating a hybrid travel demand model with the selective and judicious introduction of disaggregate destination choice models, the latter is building a data-driven activity-based model (ABM) from state-wide survey data.

4.1 The Las Vegas application

The numerical results confirm that the no-IPU case can indeed match household marginals with a high degree of accuracy. This is to be expected, as the baseline synthesis process explicitly controls for these marginals. Fig. 3, for example, shows the fit to household marginals for size 5+, with the horizontal and vertical axes representing the target and synthesized totals.

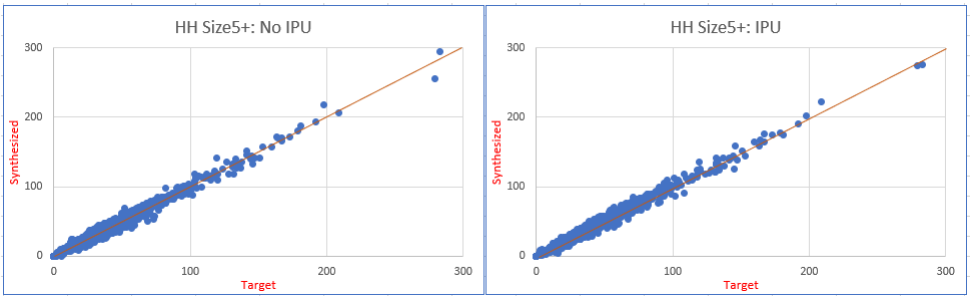

Fig. 3. Household size 5+ marginals (a) Without IPU and (b) With IPU

Similar near-perfect fit is obtained both with and without IPU for all other levels of household size, as well as each level of household auto ownership and income. However, the benefits of the IPU corrections are realized when comparing person marginals. Fig. 4, for instance, compares the fit to male residents across zones, indicating a significant tightening of the solution when IPU is employed.

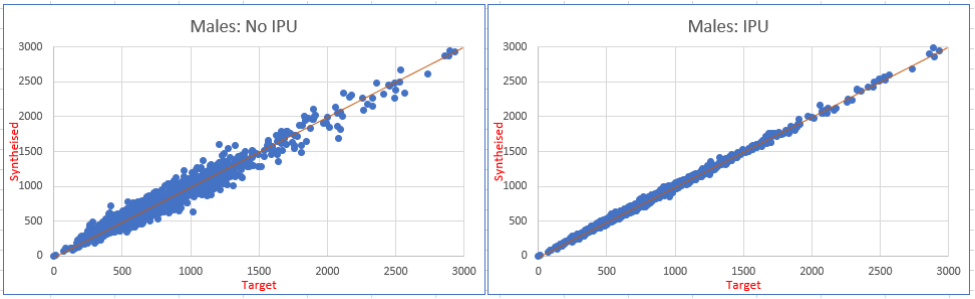

Fig. 4. Gender (male) marginals (a) Without IPU and (b) With IPU

An even more distinct comparison is obtained with the age marginals. The power of IPU is made obvious by the comparison of an age marginal in Fig. 5:

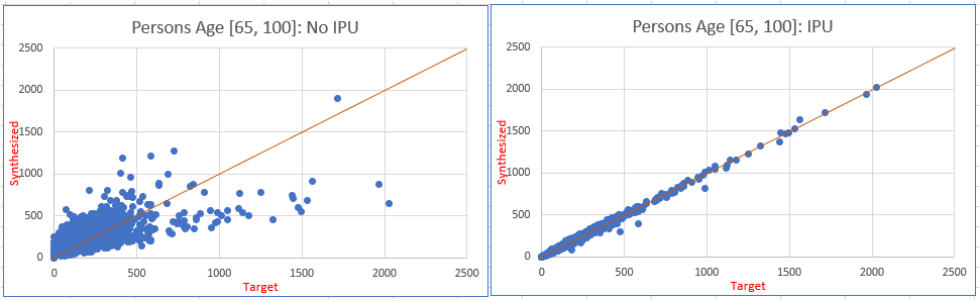

Fig. 5. Age (65+ years) (a) Without IPU and (b) With IPU

4.2 The Central Coast application

The numerical evidence for the Central Coast dataset reinforces the findings from the Las Vegas example. Fig. 6 summarizes the fit to person age marginals for the 20-24 year category, indicating that the IPU is successful in generating more representative populations than the initial weights alone.

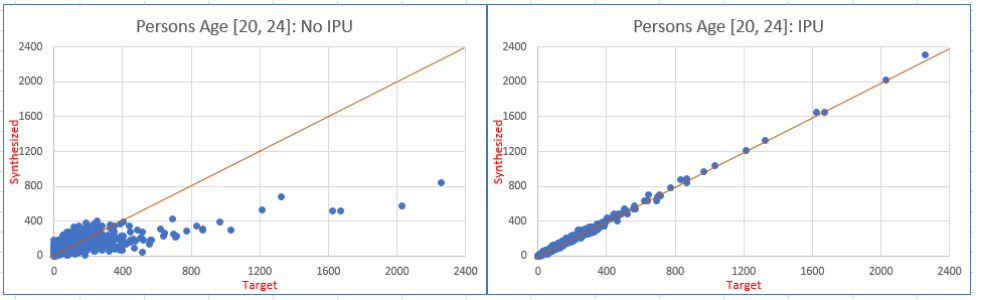

Fig. 6. Age 20-24 years (a) Without IPU and (b) With IPU

Further proof is provided in the 65+ age category, in Fig. 7:

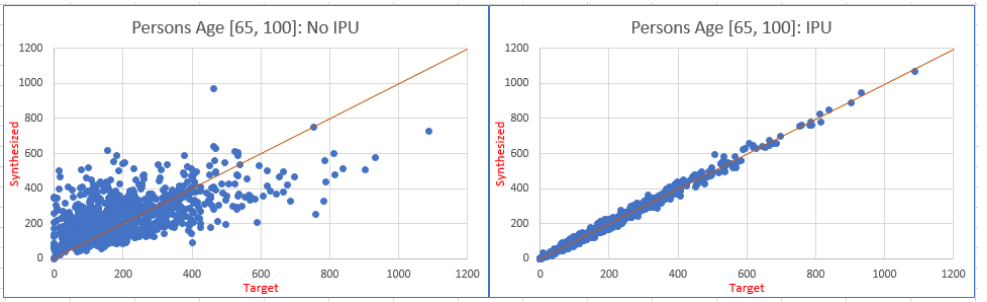

Fig. 7. Age (65+ years) (a) Without IPU and (b) With IPU

4.3 A note on the consistency of marginals

The success of the IPU procedure relies greatly on the consistency of the various marginals involved. In practice, it is highly likely that the household and person marginals do not always add up. This situation was noted in both test cases described above. When such inconsistencies exist in the data, IPU is able to match either the household marginals or the person marginals, but not both.

The primary cause for such inconsistencies was identified as the highly-skewed average household size estimates for the last household size category (in our case, these were the households with seven or more residents). The Census estimates for the average household size for 7+ households were consistently in the range of 17 to 42, which is extremely high for reasonable households. Our hypothesis is that multi-family and senior living situations were combined into single households, which tends to skew the average occupancy upwards. For the purpose of travel demand modeling though, these units should be treated separately.

We adopted a novel corrective process in which a third-party data source, one sourced from retail and other transactional records, was used to arrive at more reasonable and realistic estimates of household size for the 7+ case. This was used as a control variable to re-distribute “excess” Census households into lower-size bins proportional to their incidence in the Census. Such corrections were found to be critical in ensuring data consistency for IPU. In addition, realistic household sizes are also essential in downstream components of ABMs, since they can result in more plausible interactions between the members of smaller households. Models of joint activity participation among household members can also become tractable and obviate the need for some of the simplifying assumptions seen in practice.

5. Conclusion

In this paper, we identify key limitations in existing approaches to handling person marginals in population synthesis techniques. We propose enhancements that address these gaps, while simplifying the problem and resolving critical issues (such as the zero-cell problem) raised in the existing body of work. Our proposed methodology was implemented and tested on two large-scale and real-world case studies, and the empirical analysis clearly shows the efficacy of the enhanced methodology. We also provide a qualitative discussion on algorithmic efficiency and show that our simplifications also lead to more robust convergence behavior. The enhanced population synthesis is expected to form the backbone of future hybrid and activity-based model development for the two regions.

Acknowledgements

The authors thank the Regional Transportation Commission of Southern Nevada (RTC SNV), Association of Monterey Bay Area Governments (AMBAG), Santa Barbara County Association of Governments (SBCAG) and San Luis Obispo Council of Governments (SLOCOG) for providing the data and funding for this practice-ready research study.

References

Barthelemy, J., Cornelis, E., 2012. Synthetic Population: Review of the existing approaches, FUNDP – University of Namur, Belgium, Working paper no. 2012-8.

Muller, K., Axhausen, K.W., 2011. Population synthesis for microsimulation: State of the art, Presented at the TRB annual meeting.

Paul, B.M., Doyle, J., Stabler, B., Freedman, J., Bettinardi, A., 2018. Multi-level population synthesis using entropy maximization based simultaneous list balancing. Presented at the TRB annual meeting.

Vovsha, P., Hicks, J.E., Paul, B.M., Livshits, V., Maneva, P., Jeon, K., 2015. New features of population synthesis. Presented at the TRB annual meeting.

Xin., Y., Konduri, K., Pendyala, R.M., Sana, B., Waddell, P., 2009. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. Presented at the TRB annual meeting.