Transportation Research Part A 85 (2016) 247–264

Contents lists available at ScienceDirect

Transportation Research Part A

journal homepage: www.elsevier.com/locate/tra

An empirical assessment of factors affecting the accuracy of target-year synthetic populations

Lu Ma [a,*], Sivaramakrishnan Srinivasan [b,1]

[a] MOE Key Laboratory for Urban Transportation Complex Systems Theory and Technology, School of Traffic and Transportation, Beijing Jiaotong University, Beijing 100044, PR China

[b] Dept. of Civil and Coastal Engineering, University of Florida, 513-A Weil Hall, PO Box 116580, Gainesville, FL 32611, United States

ARTICLE INFO

Article history:
Received 16 August 2014
Received in revised form 26 January 2016
Accepted 28 January 2016
Available online 13 February 2016

Keywords:
Synthesized population
Data-fusion method
Target-year
Fitness-based synthesis

ABSTRACT

This study contributes by presenting an empirical assessment of the accuracy of the target-year populations synthesized with different base-year populations, data-fusion methods, and control tables. Forty-five synthetic populations were generated for 12 census tracts in Florida for this purpose. The empirical results indicate the value of synthesizing base-year populations more accurately by accommodating multi-level controls. Although fewer controls are typically available for target years, the use of multi-level controls in the target year with appropriate synthesis methods does benefit the accuracy of the synthetic population. This study also establishes that the magnitude of the overall error in the synthesized population appears to be linearly related to the magnitude of the input errors introduced via the control tables. The improvements in accuracy are statistically significant and hold after controlling for differences in population sizes and growth rates for the different census tracts. Overall, efforts to accurately synthesize base-year populations and to good forecasts of target-year controls can help synthesize accurate target-year populations.

1. Introduction

The need for detailed socio-economic population characteristics as an important pre-requisite to the application of disaggregate, microsimulation-based travel-demand models such as the activity-based models has increased significantly in the recent years. The true characteristics of the population of an area is usually very time consuming, expensive and sometimes impossible (e.g., for future population) to obtain (Barthelemy and Toint, 2013). Therefore, a synthetic population, described in terms of several household and/or person attributes (socio-economic characteristics), is developed to serve as a proxy for the true population.

Once a synthetic population for a target year (generally a future year for which planning is being undertaken) is available, disaggregate – travel-demand – models can be applied to the artificial households and persons in this population to determine the travel patterns. The travel patterns of the individual persons and households can be suitably aggregated to determine the overall demand and system-performance. This approach improves forecast by addressing the issue of aggregation bias that current aggregate-models suffer from (i.e., the behavior of an average person in the population is not truly representative of the overall behavior of the population; see for example, Koppelman, 1974; Landau, 1978). This is because dis-

* Corresponding author. Tel.: +86 1051684599; fax: +86 1051840080.
E-mail addresses: lma@bjtu.edu.cn (L. Ma), siva@ce.ufl.edu (S. Srinivasan).
[1] Tel.: +1 352 392 9537x1456.

http://dx.doi.org/10.1016/j.tra.2016.01.016
0965-8564/

aggregate models can accurately represent the underlying activity-travel generation process by making the individual decision maker, their choices, and their decision-making processes the center of the modeling paradigm. Further, the approach allows for extensive scenario testing; i.e., assessments of travel patterns under a variety of socio-economic futures. Such scenario-based planning methods are becoming increasing important and relevant (Bartholomew, 2007; Swartz and Zegras, 2013). Finally, the overall disaggregate approach also enables the assessments of benefits and costs of any transportation project/policy on several specific population sub-groups (such as the elderly, the low-income, and the transit-captive). There is increasing emphasis (Alsnih and Hensher, 2003; Chakraborty, 2006) being placed on such detailed environmental justice analyses. Existing aggregate modeling methods are limited in the number of market segments that can be studied; in contrast, disaggregate models with synthetic populations can deal with any number of market segments as each individual/household is explicitly represented in the model.

While the needs and theoretical benefits of the disaggregate approach is established, it is also evident that the benefits realized are subject to accuracy of the synthetic populations and the effectiveness of the demand models. The focus of this study is on the accuracy of the target-year synthetic populations. Toward this end, this study contributes to the literature by presenting an empirical assessment of the accuracy of the target-year populations synthesized with different seed data, data-fusion methods, and control tables.

The rest of this paper is organized as follows. Section 2 presents a synthesis of literature and positions the paper in the context of past literature. Section 3 presents the analysis framework. The data used in the study are discussed in Section 4 and the results are summarized and discussed in Section 5. The paper ends (Section 6) by presenting an overall summary of the work, the major conclusions, and the directions for future work.

2. Literature synthesis

A conceptual overview of the procedure for synthesizing target-year population characteristics is presented in Fig. 1. The population for a "base year" is generated first. The base year is usually the most recent census year in the past. The synthesis of the base-year population is performed by "fusing" aggregate data in the form of control totals of select attributes with detailed (disaggregate) population characteristics available for a sample of households in the area (called the seed data and are typically available from the census). The data fusion was first accomplished using the Iterative Proportional Fitting (IPF) methodology (Beckman et al., 1996) and using only household-level attributes as controls. Other than the commonly used IPF approach, earlier studies also developed combinational optimization approaches for the generation of base-year populations (Williamson et al., 1998; Voas and Williamson, 2000), and more recently, Farooq et al. (2013) developed a method using the Markov Chain Monte Carlo principle to generate base-year populations. Subsequently, other methods for data fusion have also been developed which simultaneously accommodate multi-level controls (such as household and person level) thereby relaxing the IPF procedures' requirement that all control tables be at the same "universe" (see for instance, Srinivasan et al., 2008). In order to combine information from different "universe" (for example, household and person information), approaches based on modified or multi-stage IPF procedures (Arentze et al., 2007; Guo and Bhat, 2007; Ye et al., 2009; Auld and Mohammadian, 2010; Müller and Axhausen, 2011; Pritchard and Miller, 2012; Zhu and Joseph, 2014); entropy optimization methods (Bar-Gera et al., 2009; Lee and Fu, 2011); heuristic search techniques (Ryan et al., 2010; Abraham et al., 2012; Ma and Srinivasan, 2015) or a bipartite graph approach (Anderson et al., 2014) have been proposed during the last several years. Methods to generate populations from only aggregate data also exist (see, for example, Gargiulo et al., 2010; Barthelemy and Toint, 2013). These approaches are particularly useful when the prototypical households (seed data) are not available.

Given the base-year synthetic population, there are two approaches for generating the target-year population (Fig. 1). In one approach (called the evolution approach), each base-year household is "grown" over time to determine its characteristics at the target year. This involves modeling phenomenon such as ageing, births, deaths, formation (marriage) and dissolution (divorce) of households, employment and education choices, children moving out of the household, automobile ownership decisions, and emigration from or immigration to the study region. Some of the currently available model systems that adopt such an approach include MIDAS (Goulias and Kitamura, 1996), MASTER (Mackett, 1990), CEMSELTS (Eluru et al., 2008; Pendyala et al., 2012), DEMOS (Sundararajan and Goulias, 2003), and the HA module of the Oregon2 model system (Hunt et al., 2004). Other studies have focused on evolving specific attributes of the population. For example, Paleti et al. (2011) formulate the automobile holding patterns of households by simulating the activities of disposing, replacing and adding for household vehicles. Zhu et al. (2013) evolve the ageing process to predict the marginal age distribution for a target year. Such methods are appealing as they try to mimic the real processes that households and persons go though and model behavioral decisions made at different stages of the life cycle. However, limited theoretical knowledge on the complex socio-economic evolution processes and the minimal availability of relevant data at the household level limit our ability to specify and estimate good models of household evolution (Eluru et al., 2008).

An alternate approach for generating the target-year population employs the data fusion technique which is similar to the one used in base-year population synthesis. The base-year synthetic population will serve as seed data in target-year population synthesis along with projected aggregate control totals of select attributes in the target year. Thus, unlike the evolution approach, the data fusion approach does not require evolution models and, therefore, is practical for target-year population synthesis. In this paper, we will focus on data fusion methods for target-year population synthesis.