Transportation Research Part B 114 (2018) 199–212

Contents lists available at ScienceDirect

Transportation Research Part B

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

A hierarchical mixture modeling framework for population synthesis

Lijun Sun[a,*], Alexander Erath[b], Ming Cai[c,*]

[a] Department of Civil Engineering and Applied Mechanics, McGill University, Montreal, Quebec H3A 0C3, Canada
[b] Future Cities Laboratory, Singapore-ETH Centre, Singapore 138602, Singapore
[c] School of Intelligent Systems Engineering, Sun Yat-sen University, Shenzhen, Guangdong 518107, China

ARTICLE INFO

Article history:
Received 22 September 2017
Revised 3 June 2018
Accepted 4 June 2018

Keywords:
Population synthesis
Multilevel latent class
Mixture model
Probabilistic tensor factorization

ABSTRACT

Synthetic population is a key input to agent-based urban/transportation microsimulation models. The objective of population synthesis is to reproduce the underlying statistical properties of real population based on available microsamples and marginal distributions. However, characterizing the joint associations among a large set of attributes is challenging because of the curse of dimensionality, in particular when attributes are organized in a hierarchical household-individual structure. In this paper, we use a hierarchical mixture model to characterize the joint distribution of both household and individual attributes. Based on this model, we propose a framework of generating representative household structures in population synthesis. The framework integrates three models: (1) probabilistic tensor factorization, (2) multilevel latent class model, and (3) rejection sampling. With this framework, one can generalize not only the associations of within- and cross-level attributes, but also reproduce structural relationships among household members (e.g., husband-wife). As a case study, we implement this framework based on the household interview travel survey (HITS) data of Singapore, and then use the inferred model to generate a synthetic population pool. This model demonstrates great potential in reproducing the underlying statistical distribution of real population. The generated synthetic population can serve as a replacement for census in developing agent-based models, with privacy and confidentiality being protected and preserved.

1. Introduction

Agent-based microsimulation models have become increasingly important in urban/transportation planning practices (e.g., MATSim Balmer et al., 2006). Compared with traditional aggregated planning models, agent-based models simulate the decisions and activities of each individual person over time, providing more detailed and accurate information for planning and policy evaluation. A first and critical step in developing such models is to prepare a list of population (agents) with comprehensive demographic and socioeconomic attributes that may affect agents' decision-making and activity patterns. An ideal data source for this purpose is the census data of a city, since it registers full information of the whole population. However, due to privacy concerns, the full census data is strictly confidential and even the use of samples and marginals is highly sensitive. To reduce the risk of disclosure, a typical practice of statistical bureaus is to release two sets of reproduced

* Corresponding authors.
E-mail addresses: lijun.sun@mcgill.ca (L. Sun), erath@ivt.baug.ethz.ch (A. Erath), caiming@mail.sysu.edu.cn (M. Cai).

https://doi.org/10.1016/j.trb.2018.06.002
0191-2615/

data separately: (D1) a small fraction (e.g., 1–5%) of disaggregated microsamples, and (D2) marginal distributions for different attributes. Data set (D1) is often referred to as the public use micro samples (PUMS). Data set (D2) is usually provided as one-way, two-way, and sometimes multi-way cross-tabulations aggregated from the full census.

Because of the confidentiality and privacy issues in using census, developing methodologies to generate a synthetic population has received considerable attention in the literature. A common objective of these models is to take full use of the available microsamples and marginals to create a representative list of agents, which can reproduce the underlying structure and statistical properties of the real population as much as possible. In practice, it is not easy to achieve this goal and there are three challenging problems to be addressed in designing population synthesis models. The first challenge is to preserve the high-dimensional dependency structure and match the aggregated data without introducing potential biases. For example, at the individual level, *age* and *income* are clearly associated, while intuitively we may consider *age* and *sex* to be independent across the population. At the household level, a good example is the dependency between *dwelling type* and *number of household members*—a household with more members needs a large house. These association structures become more and more complex when the number of attributes gets larger, and an ideal population synthesis model should be able to fully capture these structures. Most existing works on population synthesis have focused on solving this problem. The second challenge is to associate both household-level attributes and individual-level attributes in a unified manner (Anderson et al., 2014). For example, *car availability* (as a household attribute) should be strongly related to whether household members have *driving licenses* (as an individual attribute). The third challenge is to reproduce the interdependencies among agents in the same household (e.g., the relationship of *husband* and *wife*), even this type of structural relationship is not reported in the census data. While the latter two issues are as critical as the first one, in the past little attention is paid to reproduce the cross-level and within-household associations.

There is a vast literature on population synthesis modeling. In general, previous work can be split into three categories: (1) synthetic reconstruction (SR) (e.g. Deming and Stephan, 1940; Beckman et al., 1996), (2) combinatorial optimization (CO) (e.g. Williamson et al., 1998; Voas and Williamson, 2001), and (3) statistical learning (SL) (e.g. Farooq et al., 2013; Sun and Erath, 2015; Saadi et al., 2016; Hu et al., 2017). In terms of development, SR and CO-based models have been studied for decades and applied in various projects. However, these models often have some problems in implementation and a critical one is that SR and CO only replicate existing agents in the PUMS. Thanks to the advances in statistical learning theory and application, probabilistic and SL-based models have become emerging in the development of population synthesis models recently. In comparison with SR and CO, SL-based approaches try to encode the structure of population as a probabilistic model, and thus it is able to generate "real" synthetic data by sampling from the distribution instead of cloning (Farooq et al., 2013). Among those SL-based approaches, notably, Hu et al. (2017) proposed to model household-individual association using a nested latent class structure. This seems to be the first SL-based work addressing the association issues among household-individual attributes. The Dirichilet process is used to capture the number of latent classes in a non-parametric Bayesian setting. This model shows great potential in capturing the interdependencies among individuals within the same household by using a household class-specific conditional distribution. Although this model is both flexible and effective, the underlying assumptions still create some problems in real-world implementations. First of all, given the conditional independence assumption for individuals in the same household, it cannot fully characterize structural relationships in households. Second, since individual classes are defined separately for each group-level class, the model needs a large number of parameters and the computational cost in Bayesian inference could be high when the size of input data and the number of attributes of interest become large.

In this paper, we use a hierarchical probabilistic model to capture and reproduce the structure of population at both household and individual levels. To better characterize the underlying joint distribution for both households/individuals and the within-/cross-level association structures, the proposed framework integrates three models: (1) probabilistic tensor factorization (Sun and Axhausen, 2016), which is applied to model the joint distribution for nominal categorical variables at each level using a mixture structure, (2) multilevel latent class model (Vermunt, 2003; 2008), which captures the interaction between household-level and individual-level latent classes, and (3) rejection sampling, which further filters the synthetic population to preserve the structural relationships among individuals in the same household (e.g., *husband-wife-child*). The first two models provides an integrated probabilistic model for full household observations. Based on the estimated model, we can generate a large pool of synthetic population from the inferred model. The first two models ensures this hierarchial mixture framework to capture the association among household- and individual-level attributes; however, it still cannot reproduce meaningful individual associations within a household due to the assumption that individuals are independent given household class label. To correct this, rejection sampling (the third model) is used as a postprocessing step, in which we transform the structural relationships into a target distribution to filter those created households/individuals. The remaining samples after rejection sampling are used as the final synthetic population. This integrated framework allows us to learn the underlying structure distribution of population from limited PUMS data. In applying this model, one only need the PUMS data as input and define two hyperparameters (G and M for numbers of latent classes at the household level and the individual level, respectively). The MATLAB codes for this project are available at https://github.com/lijunsun/population_synthesis_hierarchical.

The remainder of this paper is organized as follows. In Section 2, we review previous literature on population synthesis modeling. Section 3 first provides an overview of the hierarchical population synthesis problem, and then presents a model that integrates probabilistic tensor factorization and multilevel latent class model. In addition, an efficient expectation maximization (EM) algorithm is derived for model inference. Using the household interview travel survey (HITS) data in