Transportation Research Part C 61 (2015) 49–62

Contents lists available at ScienceDirect

Transportation Research Part C

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

A Bayesian network approach for population synthesis

Lijun Sun [a,b,*], Alexander Erath [a]

[a] Future Cities Laboratory, Singapore-ETH Centre, Singapore 138602, Singapore
[b] The Media Laboratory, Massachusetts Institute of Technology, Cambridge, MA 02142, USA

ARTICLE INFO

Article history:
Received 12 June 2015
Received in revised form 5 October 2015
Accepted 12 October 2015
Available online 2 November 2015

Keywords:
Population synthesis
Agent-based model
Bayesian networks
Data-driven

ABSTRACT

Agent-based micro-simulation models require a complete list of agents with detailed demographic/socioeconomic information for the purpose of behavior modeling and simulation. This paper introduces a new alternative for population synthesis based on Bayesian networks. A Bayesian network is a graphical representation of a joint probability distribution, encoding probabilistic relationships among a set of variables in an efficient way. Similar to the previously developed probabilistic approach, in this paper, we consider the population synthesis problem to be the inference of a joint probability distribution. In this sense, the Bayesian network model becomes an efficient tool that allows us to compactly represent/reproduce the structure of the population system and preserve privacy and confidentiality in the meanwhile. We demonstrate and assess the performance of this approach in generating synthetic population for Singapore, by using the Household Interview Travel Survey (HITS) data as the known test population. Our results show that the introduced Bayesian network approach is powerful in characterizing the underlying joint distribution, and meanwhile the overfitting of data can be avoided as much as possible.

1. Introduction

The development of agent-based urban transportation and land use micro-simulation models, such as MATSim[1] (Balmer et al., 2006), UrbanSim[2] (Waddell, 2002) and ILUTE[3] (Salvini and Miller, 2005), has greatly benefited the process of urban policy making. In principle, these models simulate the behavior/activity patterns of each agent over time, helping researchers and decision makers to evaluate the impact of various policy scenarios related to transportation, land use and other urban environmental issues in a simulation-based setting. As an essential component, these micro-simulation models require a complete list of agents with detailed demographic and socioeconomic information at both individual and household levels.

Population synthesis is the process to generate an appropriate realization of the entire population, for each region/zone of interest, as the initial input to the aforementioned micro-simulation models. In doing so, we need to have a comprehensive understanding about the underlying structure of the studied population. Ideally, such information could be collected from census data at an individual/household level, and then we can draw a certain amount of samples as synthetic population. However, the use of such a detailed and disaggregated data set is highly sensitive, since one can easily identify a person

* Corresponding author at: 75 Amherst Street, E14-574A, Cambridge, MA 02142, USA. Tel.: +1 6173243782.
E-mail addresses: sunlijun@media.mit.edu (L. Sun), erath@ivt.baug.ethz.ch (A. Erath).

[1] http://www.matsim.org/ Accessed March 9, 2015.
[2] http://www.urbansim.org/Main/WebHome Accessed March 9, 2015.
[3] http://www.civ.utoronto.ca/sect/traeng/ilute/ilute_the_model.htm Accessed March 9, 2015.

http://dx.doi.org/10.1016/j.trc.2015.10.010
0968-090X/

by filtering people using those presented demographic and socioeconomic criteria. As a result, the use of disaggregated census data is highly restricted in most countries and such data is almost never accessible to researchers for the purpose of urban modeling. Instead of releasing the complete data, most governments and agencies do provide a subset sampled from the whole population at a rate—ranging from 1% to 10%—for the purpose of urban modeling. This subset of microsamples is usually referred to as public use micro sample (PUMS). For instance, the Integrated Public Use Microdata Series (IPUMS)[4] project collects and distributes PUMS from USA (IPUMS-USA)[5] and around the world (IPUMS-International).[6] These microdata sets are made available to researchers for free upon protecting statistical confidentiality. When PUMS is not available or accessible, travel surveys that capture complete demographic and socioeconomic attributes in a comparable sampling rate can act as a replacement. In addition to these microsamples, aggregated marginal information on regional/zonal level is usually available from the bureau of statistics. The goal of population synthesis is to effectively and efficiently utilize the available microsamples—together with the complementary aggregated/marginal information on each attribute of interest—to create a realization of population that could satisfy the underlying population structure as much as possible.

One of the most popular existing techniques for generating synthetic population is Iterative Proportional Fitting (IPF), which focuses on fitting a contingency table constructed from the microsamples to marginal constraints from aggregated census data (Beckman et al., 1996; Agresti, 2002). Although IPF was proposed as a general numerical method to analyze contingency tables (Deming and Stephan, 1940), it fits the description of population synthesis problem very well and has long been considered a milestone in the field of population synthesis research. Given its widespread application, various extensions and mutations have been developed based on the general IPF procedure to generate population with more complex structures. The classical IPF model can be considered a loglinear model without interactions terms (Agresti, 2002, chap. 8.).

Another branch of models follow a probabilistic framework, which assumes, essentially, all agents come from a population that is characterized by an underlying multivariate distribution. Such a joint distribution is capable of capturing not only the marginal information, but also the complex dependence and higher-order interactions between different variables. By sampling from this distribution, we are able to create an infinite pool of attribute-stamped population. However, in most cases this joint distribution is not accessible or manageable directly and to reproduce this joint distribution becomes a primary task for most population synthesis techniques. As summarized in Caiola and Reiter (2010), current practice typically employs sequential modeling framework, which impute each variable based on the others (i.e., impute X_1 based on $(X_2, X_3, \ldots, X_n)$, impute X_2 based on $(X_1, X_3, \ldots, X_n)$, impute X_3 based on $(X_1, X_2, \ldots, X_n)$, and so on). However, considering the complex interactions among different variables, specifying conditional distributions/models is not a easy task, in particular when we have many variables of interest.

The purpose of this paper is to introduce a new alternative in the probabilistic framework. We propose to use a Bayesian network model as an alternative to approximate the inherent joint distribution in a more efficient manner. A Bayesian network encodes probabilistic relationships (causality or dependence) among a set of variables by using a graphical model. Given the high efficiency and advantages provided by its graphical representation, this data-driven approach is able to determine the core structure of a population system with a limited number of microsamples. In this sense, Bayesian network models are powerful tools for learning the structure of population systems, particularly in the case where the number of attributes of interest is large while the amount of available microsamples are limited. This paper is devoted to illustrating the application of this new alternative for population synthesis.

The remainder of this paper is structured as follows. In Section 2, we briefly review existing approaches on population synthesis and the use of Bayesian networks in transportation modeling. In Section 3, we introduce the main methodology for using Bayesian network to efficiently characterize the core structure of a population system. This structure is then used as a representation of the underlying joint distribution. With this graphical reorientation and estimated local conditionals, we can produce a realization of population by sampling the estimated Bayesian network. As an illustration, in Section 4 we apply the proposed Bayesian network approach to generate synthetic population of Singapore based on information collected from a large-scale travel survey. Concluding remarks are discussed in Section 5.

2. Literature review

Essentially, the development of any population synthesis techniques can be divided into two stages – fitting and generation (Müller and Axhausen, 2011). The fitting stage aims at characterizing the multiway distribution of all attributes of interest based on the microsamples and available marginal information. The second stage focuses on generating a list of individuals/households by sampling from the fitted distribution. The fitting stage has long been considered to be difficult, since it involves estimating a complex multivariate distribution from limited observations.

To cope with the fitting problem, various techniques have been developed, including the aforementioned IPF and other Combinatorial Optimization (CO) based approaches. Given its simplicity and good performance, IPF has become the primary choice in population synthesis since its development (Deming and Stephan, 1940; Beckman et al., 1996). In general, the IPF model is a particular type of loglinear model that only preserve the main effects. Researchers are making continuous efforts

[4] https://www.ipums.org/, Accessed August 8, 2015.
[5] https://usa.ipums.org/, Accessed August 8, 2015.
[6] https://international.ipums.org/, Accessed August 8, 2015.