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

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.

to enrich the application of IPF, developing various extensions and mutations based on its principle to deal with emerging problems. For example, Pritchard and Miller (2012) proposed to use sparse matrix manipulation techniques to solve the memory consumption problem when the dimension of the contingency table (number of attributes of interest) is high. Guo and Bhat (2007) presented an improved IPF procedure to deal with the zero-cell value problem and the consistency between individual and household level attributes. In order to match household and person attributes as close as possible in a universal generator, Ye et al. (2009) proposed an Iterative Proportional Updating algorithm to control both levels simultaneously. To better control the fitting at both household and individual levels, hierarchical and multi-stage IPF procedures are proposed to preserve the inter-relationships at these two levels (Casati et al., 2015; Zhu and Ferreira, 2014). The CO approach tries to assign a weight parameter, which is computed by matching zonal marginals, to each sample (Voas and Williamson, 2001). This approach is less prevalent than IPF but it is suggested to produce less variance (Ryan et al., 2009). A comprehensive review about these fitting-based approaches and their extensions could be found in Müller and Axhausen (2011) and Farooq et al. (2013).

Another crucial shortcoming of these conventional fitting methods, which is less discussed in the literature, is that the synthesized population is created by cloning/replication rather than a true synthesis (Farooq et al., 2013). As a result, the quality of synthetic population is highly determined by the accuracy and amount of available microsamples. To build a model that is more flexible in terms of data requirement, Barthelemy and Toint (2013) developed a sample-free synthesis procedure based on a three-step optimization approach. However, this approach ignores the dependence between household and individual levels and does not maintain the consistency in these two layers. There also exists a body of literature in survey sampling, exploring weighting methods to control marginals at both individual and group levels (Deville et al., 1993; Casati et al., 2015). To deal with the lack of heterogeneity and the limitation of microsamples, researchers have been developed other statistical learning-based approaches. The principle of these methods is to update variables sequentially based on plausible conditional models. For example, Reiter (2005) suggested to use a model to update each variable in sequence, and in doing so the author employed classification and regression trees (CART) model to characterize conditional distributions and draw samples. This model has been further developed in Caiola and Reiter (2010), which replaced CART with the more advantageous random forest (RF). Farooq et al. (2013) proposed to used discrete choice models to estimate those conditionals and apply Markov chain Monte Carlo (MCMC) algorithm as the data generation model. In the numerical examples, the authors used directing counting and discrete choice models to construct conditional distributions. In terms of the framework of population synthesis, the probabilistic approach actually integrates the fitting and generation stages together. Moreover, instead of cloning individual agents, this approach is able to generate an infinite pool of potential agents as long as a list of full conditionals are specified in advance. Nevertheless, in practice it is a key challenge to prepare the full conditionals to initialize the process, since it requires us to have a comprehensive understanding of the underlying system. On the other hand, as mentioned, the difficulties in preparing full conditionals also increase with number of attributes due to the complex interactions among them, and thus the probabilistic approach is confined to the curse of dimensionality. With regard to this problem, one possible solution is to use partial conditionals, which are much simpler, to replace those full conditionals. However, professional knowledge is often required to identify the crucial relationships.

As an alternative modeling paradigm to identify causality and dependence among a set of random variables, the Bayesian network is a promising data-driven framework to abstract the complex relationships into a simple graphical model, transferring complex interdependence patterns into a concise and compact structure (Pearl, 2000; Koller and Friedman, 2009). Bayesian network models have been extensively used in probabilistic inference and reasoning problems. As a particular case, it has also been applied to analyze and interpret knowledge from survey data. For example, Sebastiani and Ramoni (2001) used a Bayesian network model to analyze and efficiently represent the General Household Survey data in UK. In this study, Bayesian network is used as an efficient tool to encode a complex probability distribution in a compact structure, providing people with a simple way to retrieve information from the data. Therefore, privacy and confidentiality are well preserved by using this approach. This also indicates that a Bayesian network model could be used as a possible approximation of the underlying distribution to produce artificial observations. As its first attempt in travel behavior research, Xie and Waller (2010) applied a Bayesian network model to quantify model choice behavior using household travel survey in San Francisco. A tabu search procedure was also presented for efficient structural learning. The successful application of these models also inspire us to Bayesian network models for the purpose of population synthesis. Further applications of Bayesian network models in transportation research include, but are not limited to: agent-based activity simulation and prediction (Janssens et al., 2006), accident/incident modeling (Zhang and Taylor, 2006), and traffic flow prediction (Castillo et al., 2008).

3. Methodology

The general population synthesis problem can be considered to be the inference of a multivariate probability distribution $\mathbb{P}(\mathbf{X})$, where $\mathbf{X} = \{X_1, \ldots, X_n\}$ is a set of attributes characterizing the demographic and socioeconomic information of individuals and households. In almost every case, we have little knowledge of $\mathbb{P}(\mathbf{X})$, except a set of sampled observations $\mathbf{D}$ that come from PUMS and travel surveys and aggregated marginal distributions from census. The difficulties in estimating $\mathbb{P}(\mathbf{X})$ arise from the fact that the number of attributes of interest is often very large, while observations from PUMS and travel surveys are usually too limited to describe the complex dependence and relationships of the underlying joint distribution $\mathbb{P}(\mathbf{X})$. As a result, it is difficult to draw samples from the unknown joint distribution $\mathbb{P}(\mathbf{X})$ directly. Essentially, the Bayesian

network approach that we are to introduce in this paper is also grounded on the inference of the joint distribution $\mathbb{P}(\mathbf{X})$. Before introducing the details of the Bayesian network approach, we first briefly review the strength and weaknesses of the sequential modeling strategy by taking MCMC as an example.

The principle of the MCMC approach is to use the Gibbs sampler to reproduce a complex joint distribution by exploiting all of its full conditionals. In doing so, the MCMC approach first prepares full conditional distributions $P(X_i|X_1,\ldots,X_{i-1},X_{i+1},\ldots,X_n) = P(X_i|X_{-i})$ for each variable X_i (we denote X_{-i} as all the other variables except X_i), and then applies the Gibbs sampler algorithm to draw a sequence of samples by updating each variable in turn. By doing so we also create an infinite pool of samples. Based on the ergodic theorems, the stationary distribution of those samples generated from the Gibbs sampler is the target joint distribution (Robert and Casella, 2004). In terms of population synthesis, a Gibbs sampler could work very well when all full conditionals are well defined and estimated in advance. However, in practice the required full conditionals are hardly available without a population-scale data set. And even with large amount of data, specifying parametric models for high-dimensional problem is still resource-intensive Caiola and Reiter (2010). Therefore, preparing all the full conditional distributions is considered a key challenge in its application. In a case study, Farooq et al. (2013) constructed $P(X_i|X_{-i})$ by directly counting frequencies of each outcome from a population-scale data set. Although the counting method is simple, in reality it is not always a good way to construct conditionals, even based on the full census. On the one hand, we may over fit $P(X_i|X_{-i})$ by counting the occurrences when we only have a small set of subpopulation with $X_{-i} = x_{-i}$. For instance, if we only have one observation with $X_{-i} = x_{-i}$, the estimated conditionals of $P(X_i|X_{-i})$ will be either 0 or 1, which may not be the correct values we want. This is particularly important if the number of attributes of interest is large while the relative size of available PUMS is small (in other words, we have a large sparse contingency table).

To strengthen the applicability of the MCMC approach when observations are limited, Farooq et al. (2013) introduced two methods to better characterize those full conditionals. The first method is to use parametric models (e.g., discrete choice models/multinomial linear logistic models) to construct full conditionals. This is similar to applying loglinear models to capture higher-order interactions in contingency tables (Agresti, 2002, chap. 8). This method enables us to efficiently capture the complex conditionals, as long as we have a fair understanding about the modeling of the higher-order interactions between different variables (Casati et al., 2015). However, implementing such a model computationally is challenging, in particular when the contingency table has enormous number of cells while the number of observations is limited. In order to get a satisfactory goodness of fit, the applied parametrical model has to impose a large number of coefficients to be estimated. The second method is to use partial/incomplete conditionals $P(X_i|X'_{-i})$ ($X'_{-i} \subseteq X_{-i}$ is a subset all other variables) to replace full conditionals $P(X_i|X'_{-i})$. Indeed, it seems to be reasonable to reduce the size of X_{-i} since not all attributes determines X_i directly, and by removing the non-relevant variables we can obtain X'_{-i} in a smaller size. The use of partial/incomplete conditionals not only prevents overfitting, but also mitigates the risk of encountering incomplete conditional distributions. However, this reduction essentially ignores some of the higher-order interactions and it should be determined by domain knowledge and proper assumptions, and thus it is still an open question that which conditional distribution should be simplified and to what format. In other words, identifying the best subset X'_{-i} for each variable X_i emerges as a new problem. In summary, specifying conditional distribution on each variable is still very time-consuming to account for potential joint interactions.

As a model that integrates causal relationships and probabilistic semantics, we consider Bayesian network an alternative tool to simplify the estimation of the joint distribution $\mathbb{P}(\mathbf{X})$. In the following of this section, we propose to apply Bayesian network models to efficiently approximate/reproduce $\mathbb{P}(\mathbf{X})$ through identifying the critical relationships among different attributes in population systems.

3.1. Bayesian networks

The Bayesian network is a graphical model that efficiently encodes probability distributions for a set of variables of interest (Heckerman, 1998; Pearl, 2000; Koller and Friedman, 2009). Essentially, a Bayesian network for a set of variables $\mathbf{X} = \{X_1,\ldots,X_n\}$ consists of two parts: (1) the qualitative part is a network structure G in the form of a directed acyclic graph (DAG), in which nodes are in one-to-one mapping with the random variables $\mathbf{X}$ and links characterize the dependence among connected variables, and (2) the quantitative part is a set of local probability distributions/tables $\Theta = \{P(X_1|\Pi_1),\ldots,P(X_n|\Pi_n)\}$ for each node/variable X_i, conditional on its parents Π_i (see Fig. 2 for examples of conditional probability tables). These conditional probability tables demonstrate the probability of X_i with respect to each combination of its parent variables. In a Bayesian network we refer to X_j as a parent of X_i if there exists a direct link from X_j to X_i. We use Π_i to denote the set of parent variables of X_i. If a variable has no parents, the local probability distribution collapses to its marginal $P(X)$. In a Bayesian network model, we refer to G as model structure and Θ as model parameter. The DAG topology of a Bayesian network only asserts conditional dependence of children given parents. Therefore, by integrating structure G and parameter Θ, the joint distribution for $\mathbf{X}$ in a Bayesian network can be decomposed, by using the chain rule, into a factorized form with smaller and local probability distributions, each of which involves one node and its parents only:

$$P(\mathbf{X}) = \prod_{i=1}^{n} P(X_i|\Pi_i). \tag{1}$$

In other words, the joint probability distribution $P(\mathbf{X})$ can be exclusively encoded by the pair (G, Θ). The Bayesian network representation allows us to approximate and represent an unknown distribution $\mathbb{P}(\mathbf{X})$ into a concise graphical form ($P(\mathbf{X}) \approxeq \mathbb{P}(\mathbf{X})$). Therefore, in terms of population synthesis, Bayesian network model offers us an intuitive framework to reproduce the $\mathbb{P}(\mathbf{X})$ of the studied population system. For example, considering a simple network with two nodes and one link (age)→(income), the root node tells the probability $P(\text{age})$ of an individual being Y yrs old and the conditional probability table $P(\text{income}|\text{age})$ contains the probability of a person having income M when knowing he/she is Y yrs old (for all Y). With these two tables we can easily calculate the joint probability of observing an individual with age Y and income M as $P(\text{age} = Y, \text{income} = M) = P(\text{age} = Y) \times P(\text{income} = M|\text{age} = Y)$.

We next discuss the concept of Markov blanket and the way to construct full conditionals in a Bayesian network model. The Markov blanket $\mathbf{Mb}(X_i)$ for a particular node X_i is the union of three sets: (1) its parents, (2) its children, and (3) the co-parents—a set consists of other parents of its children (excluding X_i). Note that we can derive $P(X_i|X_{-i}) = P(X_i, X_{-i})/P(X_{-i})$ and by canceling out all terms that do not involve X_i from both numerator and denominator, we have

$$P(X_i|X_{-i}) \propto P(X_i|\Pi_i) \prod_{k \in ch(j)} P(X_k|\Pi_k), \tag{2}$$

where $ch(j)$ denotes the children nodes of X_i.

Thus, given the expression in Eq. (2), the full conditional distribution $P(X_i|X_{-i})$ depends only on its Markov blanket $\mathbf{Mb}(X_i)$. This suggests that the inference and sampling of X_i can be achieved by looking at only its Markov blanket, instead of the full conditionals. On the other hand, if we assume that $\mathbb{P}(\mathbf{X})$ is indeed characterized by a Bayesian network, then an arbitrary simplification of full conditionals—even with professional domain knowledge—could be problematic. This also gives out a warning to the use of partial/incomplete conditionals to replace full conditionals in the MCMC approach. Particularly, one is required to identify the full Markov blanket instead of the parent set only.

3.2. Model selection and optimization

Having seen the definition of Bayesian networks, we next briefly introduce the learning problem in Bayesian network analysis. Details about learning Bayesian network models could be found in most textbooks and tutorials (Heckerman et al., 1995; Heckerman, 1998; Pearl, 2000). There are two types of learning problem given a set of observations $\mathbf{D}$: (1) learning only model parameter Θ when network structure G is known, and (2) learning both model structure G and model parameter Θ. For the first type, one needs to pre-define the network structure. An intuitive way to do so is to build G based on expert knowledge. In this case, the variables are identified and causal relationships are asserted by using professional domain knowledge. After knowing network structure G, the estimation of local probability Θ becomes straightforward by applying maximum likelihood (ML) or Bayesian estimation if prior knowledge is available.

However, most practical problems belong to the second category, in which expert knowledge is not available or not sufficient enough for us to build the network structure from scratch. Therefore, we should make full use of available observations to learn G and Θ simultaneously. This process is often referred to as structural learning. In general, structural learning can be divided into two stages: model selection and model optimization. In the selection stage, we try to use a universal criteria to evaluate the quality of different hypothetical structure G^h. In the optimization stage, we focus on identifying the best structure.

To proceed with selection, we usually apply a score-based approach, computing a score function that quantifies how well a hypothetical structure G^h fits the data. In doing so, the estimation of local probability $\hat{\Theta}$ is used as an inner loop to quantify score of G^h. Intuitively, a natural candidate score function to quantify the quality of a Bayesian network model is the maximum likelihood:

$$l\left(G^h|D\right) = \max_G \sup_\Theta l(G, \Theta|D) = \max_G l(G, \hat{\Theta}|D), \tag{3}$$

where $l(G, \Theta|D) = \log P(D|G, \Theta)$ is the log-likelihood of a provided pair (G, Θ) given observation $\mathbf{D}$. However, log-likelihood is actually not an appropriate score function. On one hand, maximizing likelihood will always lead to a fully connected (complete) DAG, in which every pair of nodes are connected, regardless of what the underlying structure should be. This is because adding a link in an incomplete network will always increase or at least maintain $l\left(G^h|D\right)$. Introducing a new link will also increase the complexity of a model, since more parameters are introduced to estimate the local conditionals. This makes us over weight the characteristics of the sample, which may not representative in the underlying population structure. As a result, the large size of parameters and the high complexity of a complete DAG will lead to the problem of overfitting.

In practice the most used score function is the Bayesian information criterion (BIC), which is defined by (Schwarz, 1978):

$$\text{BIC}\left(G^h|D\right) = \log P\left(D|G^h, \hat{\Theta}\right) - \frac{d}{2} \log m, \tag{4}$$

where $\hat{\Theta}$ is the maximum likelihood estimates of parameter given a hypothetical structure G^h, d is the number of free parameters (degrees of freedom) in Θ, and m is the size of observation $\mathbf{D}$. As can be seen, the first term on the right hand side is the optimal likelihood, which quantifies how well the hypothetical structure G^h fits the data; the second term is a penalty function on the complexity of the model, preventing the overall structural learning process from overfitting. The structure of BIC make it a very prevalent choice practically when sample size is large.

Another popular candidate score function is the so-called Akaike information criterion (AIC), which is given by (Akaike, 1974):

$$\mathrm{AIC}\left(G^h|D\right) = \log P\left(D|G^h, \hat{\Theta}\right) - d. \tag{5}$$

Similar to the definition of BIC, AIC also consists of two parts. The first term is still the optimal likelihood, while penalty term is just the number of free parameters in Θ, being independent from the size of observations. Therefore, both BIC and AIC are constructed by adding penalty terms to the optimal likelihood, which balances model fit and model complexity. The only difference between them is that BIC penalize free parameters more strongly than AIC. Other score functions, such as Cooper–Herskovits (CH) score, minimum description length (MDL), holdout validation likelihood (HVL) and cross validation likelihood (CVL), could also employed under different scenarios (e.g., when data is divided into training set and test set).

After selecting a score function, the goal of the optimization stage is to identify the hypothetical structure with the highest score. Ideally, a straightforward way is to enumerate all potential candidates and evaluate score of each of them. However, in practice this is infeasible as the number of candidates increase super-exponentially with the number of nodes/variables (Robinson, 1973). For example, a network with six nodes has about 3 million possible DAGs, and this number becomes 1.1 billion when it has seven nodes. To cope with this problem, the common approach is to apply heuristic search techniques, including hill-climbing, hill-climbing with restarts, tabu, best-first search, K2, and MCMC methods (Heckerman, 1998).

Tabu search method is an iterative searching procedure to move from one solution to its neighboring solution until some stopping criterion is satisfied (Glover and Laguna, 1997). Although tabu search is still a local searching technique, its performance is enhanced by using a memory structure (tabu list) while exploring the neighborhood of each solution during the searching processes. Tabu search is also capable of escaping from local optima, in which normal local search techniques often get stuck. We refer the readers to Glover and Laguna (1997) for a complete description of this technique. We will also skip the details about other heuristic search technique. Interested readers may be referred to the tutorial of Heckerman (1998) and references inside for more on these heuristic algorithms. Similar to the MCMC approach, the Bayesian network method does not require marginals as input. Moreover, it does not require any conditionals as well, since structure learning and parameter estimation are integrated in the learning of a Bayesian network model. Therefore, despite the set of observations $\mathbf{D}$, the only input that is required in learning a Bayesian network model is a specified score function (e.g., BIC and AIC). For structure learning, we used the `R` package `bnlearn`, which implements tabu search as one of the score-based structure learning algorithms (R Core Team, 2015; Scutari, 2010).

3.3. Realization of synthetic population

As mentioned, an interpretation of the population synthesis problem is to produce a pool of samples from $\mathbb{P}(\mathbf{X})$. After learning both model structure and model parameter, we can generate/sample values of $\mathbf{X}$ given the factorized joint probability distribution $P(\mathbf{X})$ defined by the Bayesian network. Unlike the MCMC approach, samples generated from the Bayesian network are independent, and thus the procedure can be paralleled. On the other hand, there is no need to thin the results to reduce correlation between sequential samples. In fact, using the factored decomposition introduced in Eq. (1), one can also compute the probability of observing a particular sample realization $(X_1 = x_1, \ldots, X_k = x_k)$ easily.

The Bayesian network model also provides us with an efficient approach to sample based on evidence. For example, after learning a Bayesian network for the whole population, we may want to use this model to generate synthetic population in zonal level, in which each zone has unique marginal distributions on one or more variables. In this case, we may use this information (e.g., the marginal distributions or cross validation tables) as evidence to control the global sampling of $\mathbf{X}$. A more appropriate sampling technique, such as Gibbs sampling and forward sampling, might be used to match the known evidence. In fact, in the early development of MCMC, one of the obvious applications of the Gibbs sampler is on graphical models.

As the learning of Bayesian networks relies on observation $\mathbf{D}$, the quality and quantity of $\mathbf{D}$ may determine the functionality of estimated Bayesian network substantially. As mentioned, in applying the MCMC approach, we are likely to encounter the problem of unidentified distributions and overfitting when the amount of observations is not enough. Although Bayesian network models have made the structure of local conditional as concise as possible (reducing full conditionals to parent-based conditionals), in practice a local conditional distribution may still be not fully estimated when a combination of parent set $\Pi_i = \pi$ for variable X_i is not observed in data $\mathbf{D}$. In this case, part of the conditional table $P(X_i|\Pi_i = \pi)$ will be unidentified and sampling X_i given $\Pi_i = \pi$ is impossible. This occurrence of unidentified local conditional distributions increases with the size (in terms of number of potential combinations) of parent set. In other words, the larger the size of a parent set is, the higher the chance we will encounter unidentified local conditional distributions. This is particularly worth noting when the estimated Bayesian network is used for prediction and sampling purpose, since sampling from an unidentified conditional

probability will give us a missing entry. To reduce such inconsistence, we would like to keep the structure of a Bayesian network as simple as possible. This can be achieved by adopting an appropriate score function that penalizes model complexity (e.g., BIC and AIC). On the other hand, we can also reduce the number of categories in each variable, because the total number of parent combinations gets reduced as well. The problem can also be avoided by adopting a Bayesian framework, specifying prior distributions (e.g., Dirichelet) of potential parameters.

By sampling from the obtained Bayesian network we are allowed to generate a large list of individuals (or households if one applied a hierarchical approach as introduced in Section 4.3) as population pool. To incorporate marginal information (e.g., age, sex distribution at zonal level), one may apply survey sampling method to get the weight/probability of each individual/household belonging to a particular zone (Deville et al., 1993). Using this weight one may get a designed subset from the pool for the region of interest.

4. Population synthesis for Singapore

This section is devoted to demonstrating the performance of the proposed Bayesian network approach by conducting numerical experiments on generating synthetic population in Singapore. In doing so, we also assess the performance of other existing techniques, including IPF and MCMC. The population data used in the following experiments comes from Household Interview Travel Survey in 2012 (HITS2012), which is conducted by the Land Transport Authority of Singapore. Two numerical examples are provided in this section, following the proposed hierarchical approach in Casati et al. (2015). The first toy model focuses on the synthesis of household owners, while the second is a more complex example on both owners and spouses.

4.1. Data sources

HITS2012 data is an essential input for various urban and transportation planning for Singapore. Planning agencies and researchers have been using this data set (or similar data sets from other years, e.g., HITS2008) to conduct population synthesis for agent-based urban modeling (Zhu and Ferreira, 2014). The survey collected comprehensive demographic/socioeconomic attributes at both individual and household levels, covering 35,714 individuals from 9635 households (about 1% of the total population). Despite demographic/socioeconomic information, travel information such as trips/journeys in one day was also registered to study travel behavior and activity patterns. For the purpose of population synthesis, we are interested in only demographic/socioeconomic data of individuals and households.

To better evaluate the performance of different models, we assume HITS2012 to be a known full population and then use micro samples (PUMS) that are randomly generated under different sampling rates (ranging from 1% to 100%) as test data sets.

4.2. A toy model for individual synthesis

We first conduct an experiment on synthesizing individuals (household owners) with seven attributes taken into consideration. Table 1 lists the seven variables of interest and their descriptions. The total number of cells in the contingency table for IPF is $7 \times 4 \times 2 \times 12 \times 2 \times 12 \times 2 = 32,256$.

The owner of a household is determined as the individual with highest income. If there exist multiple candidates, the one with the highest age is chosen. When a conflict still exists, a randomly selected candidate will be assigned (Casati et al., 2015). In total, there are 9635 observations in this owner data set (the same as number of household). As mentioned, for this toy model we consider these observations as known population, and consider samples with different size (1% to 20%, in a 1% step, of the total population) to be available PUMS. Therefore, only 96 observations are used when sampling rate is 1%.

We next compare the synthesis results of the introduced Bayesian network approach with IPF and MCMC. To make a fair assessment of MCMC and Bayesian network approaches, we do not provide full population information to either of them. In this case, we construct full conditionals of MCMC by counting PUMS exclusively. To avoid unidentified conditionals, the initial seed of the Gibbs chain is randomly selected from each PUMS dataset. In addition to the PUMS data, we also provide IPF with a set of marginal distributions of all the seven attributes from the full population. In this sense, we provide IPF with more information than MCMC and Bayesian network approaches. In the learning process of the Bayesian network, we assume that 'age' is a natural attribute that does not depend on any other variables and it is constrained as the root node in model structure G.

As presented in Section 3, the structure of a Bayesian network model is also determined by the choice of score function. In order to avoid the change of network structure with varying sample size, we choose AIC as the score function in searching the best structure. The reason is that AIC is independent from the size of observations, while BIC prefers an overly simplified model with the increase of sample size (see the penalty terms of Eqs. (4) and (5)). We apply tabu search algorithm, which is implemented in `bnlearn` package, to learn the structure of Bayesian networks. To avoid local optima and try to exploit the space as much as possible, we use a large tabu list with length of 100 and start 24 runs with different initial network structure in parallel. Other tuning parameters for the tabu method are chosen by their default settings. This parallel multistart model is applied on each estimation process and the final result is selected as the best one among all the 24 runs given their

Table 1
Attributes of household and owner.

Level	Variable	Definition [number of categories]	Values
Household	dwell	Dwelling type [7]	HDB 1–2 rooms; HDB 3 rooms; HDB 4 rooms; HDB 5 rooms and larger; Condo; Landed property; Other
	eth	Ethnicity [4]	Chinese; Indian; Malay; Other
	car	Car availability [2]	Yes; No
Individual	age	Age of owner [12]	15–19; 20–24; 25–29; 30–34; 35–39; 40–44; 45–49; 50–54; 55–59; 60–64; 65–69; 70/ +
	sex	Gender of owner [2]	Male; Female
	Inc.	Income of owner [12]	SGD: No income; 1–999; 1000–1499; 1500–1999; 2000–2499; 2500–2999; 3000–3999; 4000–4999; 5000–5999; 6000–7999; 8000/+; Refused
	licen	Driving license of owner [2]	Yes; No

scores. The computational time of a tabu search run with a 20% sample on a PC with an Intel Core i7 3.40 GHz and 16 GB RAM is about 0.17 s.

To quantify the accuracy/fitness of different approaches, we adopt a popular measure in the literature of population synthesis, which is *Standard Root Mean Square Errors* (SRMSE) defined by Müller and Axhausen (2011):

$$\mathrm{SRMSE} = \sqrt{\sum_{m_1=1}^{M_1} \cdots \sum_{m_n=1}^{M_n} \left(f_{m_1,\ldots,m_n} - \hat{f}_{m_1,\ldots,m_n}\right)^2 \times (M_1 \times \cdots \times M_n)}, \tag{6}$$

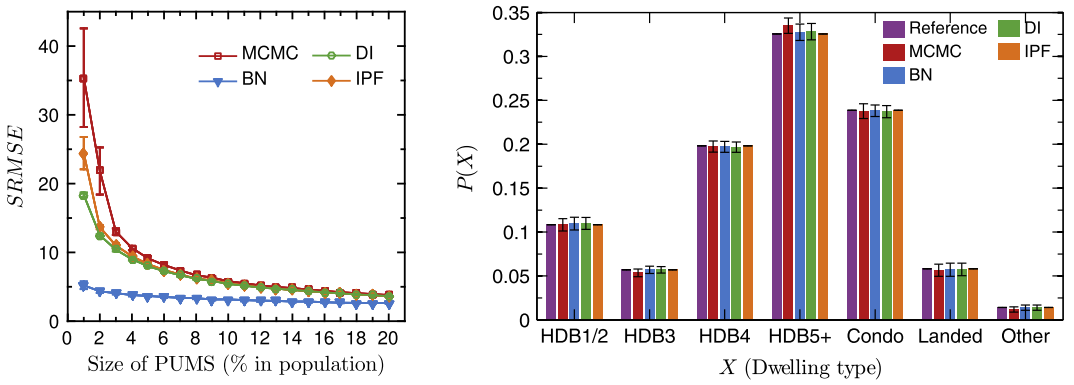
where $f_{m_1,\ldots,m_n}$ and $\hat{f}_{m_1,\ldots,m_n}$ are relative frequencies of a particular combination appears in the reference (the known population) and synthesized population, respectively; M_i is the total number of categories for attributes X_i; and thus $M_1 \times \cdots \times M_n = 32,256$ is the total number of cells in the corresponding contingency table. A value of zero means a perfect match between reference and synthetic population, while a high SRMSE value represents a bad fit. The relative frequency $\hat{f}_{m_1,\ldots,m_n}$ is measured from the a pool of synthesized population that is ten times of reference data (for each approach, we create 96,350 samples as synthetic population). To take into account the burn-in stage and avoid correlation of sequential samples in the MCMC approach, the population is created by discarding the first 10,000 samples and selecting every 10th sample in the following chain. In sampling the Bayesian network model, we only keep the samples without any unidentified entries.

To capture the variance of different approach, we create 16 groups of PUMS for each sampling rate, and thus in total 320 (16×20) sets of PUMS are generated. As a comparison, we also show the result of direct inflating (expanding/cloning by a factor—1/rate) the PUMS to match the size of total population. We plot the variation of SRMSE with the size of available PUMS in Fig. 1a. The markers represent mean values and error bars correspond to the standard deviations of 16 groups of samples. As can be seen, the performance of all these approaches increases with the size of PUMS.

When sample size is small, the performance of IPF is even worse than directly inflating (DI) the PUMS. This is because the heterogeneity of PUMS is not enough to calibrate a good contingency table (i.e., the zero-cell problem). Therefore, instead of improving the fitting, the imposed marginal constraints become a burden to the fitting of joint distribution. The MCMC approach also exhibits a low accuracy because of the large degrees of freedom in defining the full conditional distributions, and by the naïve counting we are definitely in the trouble of overfitting: the model may fit the training data (PUMS) very well but fail to capture the overall relationship at a population scale. In the case of population synthesis where we do not have enough observations, a complex model could over-capture the feature of those observations rather than the real underlying structure. This overfitting problem, together with the randomness in the simulation, undermines the performance of MCMC when sample size is smaller than 20%.

Notably, the Bayesian network (BN) approach always gives the lowest SRMSE values, demonstrating universally good performance, which is almost invariant to the size of PUMS. As a guide, we also depict the marginal distribution on dwelling type, when the size of PUMS is 20%, in Fig. 1b. The Bayesian network approach provides a similar degree of consistency as direct expanding. IPF, given its principle, gives a perfect match, while MCMC exhibits strong uncertainty. To explore why the Bayesian network approach has such good performance, we depict the structure of each estimated Bayesian network. Remarkably, we find that all these structures G are identical, as shown in Fig. 2 (household and owner attributes are marked in green and red respectively), no matter what the sampling rate is. We also show the conditional probability table of sex given age and a partial table of dwelling type given owner's income, car availability and sex. In this case, the difference and variation of the Bayesian network approach results exclusively from the estimation of model parameter $\hat{\Theta}$ from different PUMS observation $\mathbf{D}$. This invariant structure suggests that the Bayesian network is capable of identifying the core structure of the underlying population system efficiently. As a matter of fact, it also suggests that the investigated population (HITS2012) is indeed well characterized by a graphical structure, which can be well captured by using only a 1% sample set.

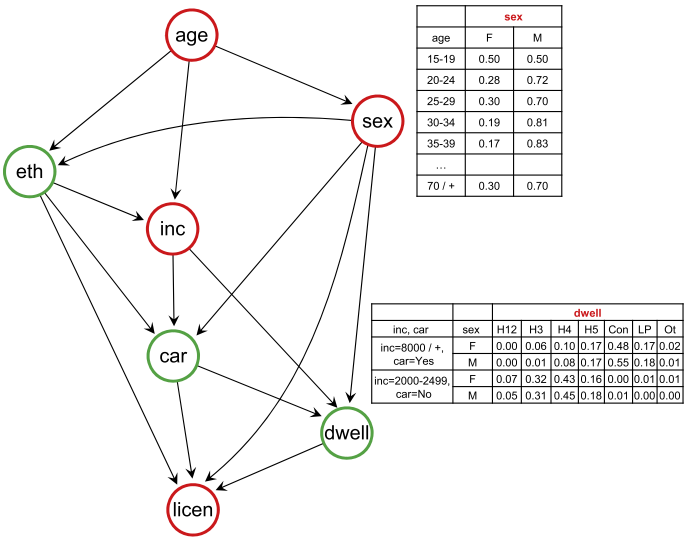
To have a better overview about the advantages and disadvantages of each approach, the same experiment is also run with larger sizes of PUMS (from 30% to 100%). We display the results of mean SRMSE in Fig. 3. As can be seen, IPF and MCMC begin to outperform the Bayesian network approach when we have 40% of total population as samples, although in practice we hardly get such a large PUMS. In fact, IPF and direct inflating/expanding will provide us a perfect match when the size of



(a) SRMSE given varying PUMS size

(b) Marginal fit of dwelling types with 20% PUMS

Fig. 1. Performance comparison of different approaches.



	sex	
age	F	M
15-19	0.50	0.50
20-24	0.28	0.72
25-29	0.30	0.70
30-34	0.19	0.81
35-39	0.17	0.83
...		
70 / +	0.30	0.70

		dwell						
inc, car	sex	H12	H3	H4	H5	Con	LP	Ot
inc=8000 / +, car=Yes	F	0.00	0.06	0.10	0.17	0.48	0.17	0.02
	M	0.00	0.01	0.08	0.17	0.55	0.18	0.01
inc=2000-2499, car=No	F	0.07	0.32	0.43	0.16	0.00	0.01	0.01
	M	0.05	0.31	0.45	0.18	0.01	0.00	0.00

Fig. 2. Model structure G on household[green]/owner[red] information. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

PUMS is 100% (as shown in Fig. 3), since the initial seed is the population itself. In this circumstance, the MCMC approach also attains the best estimates of those full conditional distributions. However, due to randomness in the simulation processes, there is still a small SRMSE in MCMC.

In terms of the Bayesian network approach, however, the SRMSE almost remains at a constant level, with only a little improvement. As mentioned, the selected score function AIC penalize the complexity (number of free parameters, degrees of freedom) of Bayesian network models. In this case, the simple model structure G (or limited number of parameters) of the Bayesian network model can no longer improve the fitness of the complex structure of the underlying joint distribution $\mathbb{P}(\mathbf{X})$. With regard to this fact, we should select a more appropriate score function that better utilizes this new information and penalizes model complexity even less. For example, an effective alternative is to integrate the prior knowledge as constraints, and search for a new structure that maximizes posterior likelihood based on the sample observations (Buntine, 1991; Heckerman et al., 1995; Friedman and Koller, 2003).

As the principle of conventional approach is to replicate/clone the agents with PUMS, in general we also miss the heterogeneity of the underlying population and the synthesized population is lack of representativeness. To further access the goodness of fit in terms of heterogeneity, we calculate the total share of reference data (full census) that are not present in the synthesized samples $L = \sum_{m_1,\ldots,m_n} f_{m_1,\ldots,m_n} \times \mathbb{I}\left(\hat{f}_{m_1,\ldots,m_n} = 0\right)$, where $\mathbb{I}(e)$ is an indicator function that equals to 1 when e is true and 0 otherwise. Therefore, the measure L quantifies how much heterogeneity we will lose by using a limited PUMS. Fig. 4 shows the mean and standard deviation of L from 16 groups of PUMS. Clearly, we see that the Bayesian network approach is superior in enriching model heterogeneity in synthetic population when size of PUMS is less than 70% of the full

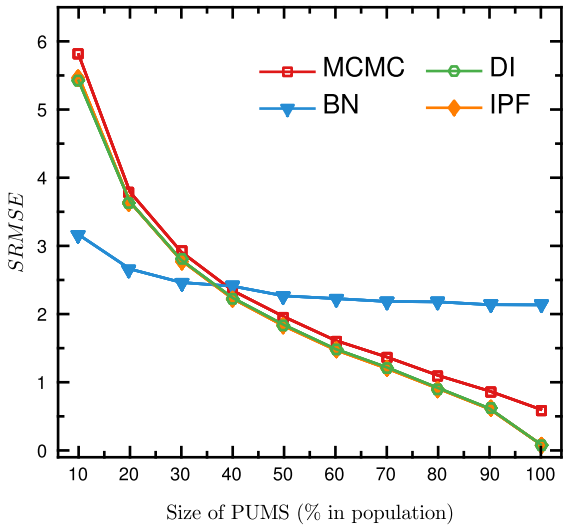


Fig. 3. SRMSE with larger sizes of PUMS (ranging from 10% to 100%).

population, generating samples that are not observed but belong to the underlying population. Given the constraints on model complexity, it cannot reproduce a 100% match when full census is provided.

In summary, the simplicity of model structure and the fitness to data are two contradictory goals. In practice, we need to find a compromise solution that takes both aspects into consideration. In the case where we have strong prior knowledge, a posterior likelihood Bayesian network approach—which integrates the graph models with Bayesian paradigm—seems to be a good alternative. As our focus is on population synthesis, in this paper we do not explore further the balance between model complexity and fitness. In the next subsection, we apply the Bayesian network approach on a more general problem—the synthesis of full household structure. In doing so, we impose a hierarchical structure to represent the configuration of households.

4.3. A hierarchical household configuration

In this part we apply the Bayesian network approach on a more complex case, which include demographic/socioeconomic information of both the owner and his/her spouse in the household. Same as the definition in Casati et al. (2015), the spouse of a household is selected as the agent with the minimum age difference from the owner among those with opposite sex.

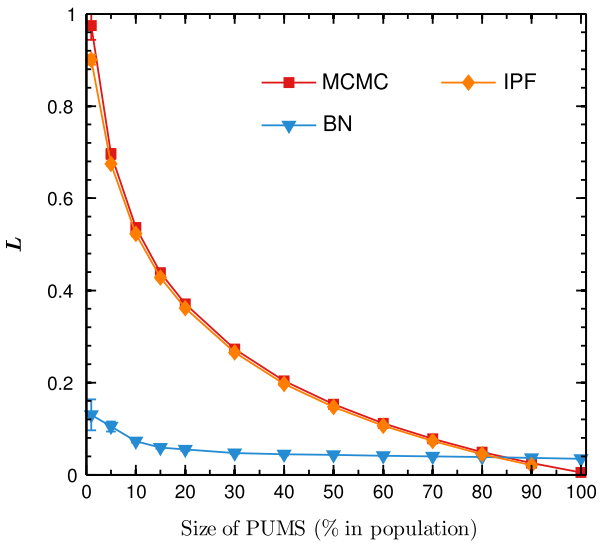


Fig. 4. Loss of heterogeneity by using limited PUMS.

Note that in the reference data a spouse may not be identified in some households (e.g., households with only one agent) and for this experiment we only have 5383 (<9635) pairs of owners/spouses.

Five more variables are introduced to this hierarchical model, including number of people in the household (*npax* in six categories: two; three; four; five; six; seven and above) and another set of demographic/socioeconomic variables for the spouse. We assume that the new Bayesian network follows a hierarchical structure. In so doing, we impose the attributes of a spouse to be the children of the aforementioned household/owner attributes in the Bayesian network.

We still apply the AIC as score function and each time a 10% sample of the 5,383 observations (i.e., 538 households) is used as PUMS in learning the Bayesian network model. Similar to the learning of individuals, the length of tabu list is set to be 100 and other tuning parameters are set as their default values. The tabu search is run for 24 time in parallel, and each estimation takes about 2 s. Given the proposed hierarchical household structure, in this new model link directions between the set of spouse attributes and owner attributes are restricted. These restrictions are considered in generating neighbor DAGs in the tabu search procedures. This numerical experiment is run for ten times with randomly initialized PUMS. Remarkably, the resulted optimal Bayesian network structures with the highest AIC scores are all identical (see Fig. 5). In particular, the partial network that only contains household and owner attributes is identical to the previously estimated model illustrated in Fig. 3, further confirming the modularity and expandability of Bayesian network models. In other words, based on its graphical structure, a Bayesian network model can be adapted in a flexible manner: on the one hand, it can be decomposed into smaller submodels to better focus on particular part of interest; on the other hand, it is also capable of being expanded to include more variables and the corresponding dependence.

To visualize the goodness of fit, we next map the synthesis results as two-dimensional distributions. As illustrations, here we only show two sets of joint distributions. Fig. 6 shows the joint distribution of owner income and spouse age in the Bayesian network. The results come from one randomly selected case from the ten runs. The left panel and the middle panel shows the joint distributions in the reference data and the synthesized data, respectively. The right panel displays the fit of these two dimensions between the two sets of data. As can be seen, these two attributes are definitely not independent, since each spouse age group has a unique owner income distribution. It is difficult to capture such higher-order interactions using conventional approach such as IPF. Fig. 7 shows a similar results for the joint distribution on spouse age and dwelling types. In this case, the two variables are approximately independent. Taken together, despite the variance from sampling, Bayesian network approach do provide satisfactory goodness of fit, even though only 10% (538) samples are used in the learning process.

4.4. Generating whole synthesis population

Following this procedure we can further estimate a Bayesian network that characterize the rest people (e.g., children) in this hierarchical approach as in Casati et al. (2015) to match total number of people (*npax*). To generate a full household, one can first sample the information of household, owner and spouse, together with number of people (*npax*) using the hierarchical network in Fig. 5. Then, according to *npax* and a network that characterize the 'other' group, we can generate the rest people to fill the full household. In doing so, we assume that individuals belong to the 'other' group are independent and exchangeable, which may not be true in reality. For example, age difference of two or more children in a household should follow a certain distribution, while independent samples generated using in this way prevent us from reproducing the distribution. In other words, the joint interactions of people in the 'other' group may not be preserved. One should pay more attention if this joint relationships are of interest.

We follow the procedure shown in Fig. 8 for the implementation to generate whole synthesis population for MATSim Singapore.[7] Firstly, we divided all 9635 households into three categories: (1) with single member [535 observations], (2) multiple member with clear owner-spouse relation [5383 observations], and (3) multiple member without clear owner-spouse relation (e.g., single parent with children) [3717 observations]. For type (1), a Bayesian network G_1 is learnt based on the 535 samples of household and owner attributes (same as Fig. 2). For type (2), we first estimate a Bayesian network G_{21} based on 5383 samples of household, owner and spouse attribute (same as Fig. 5). Next, we create a temporary dataset focusing 'other' people in these household. Taking a household with three children $(C1, C2, C3)$ and two parents (owner and spouse) $(P1, P2)$ as an example, three 'other' observations $(P1, P2, H, C1), (P1, P2, H, C2)$ and $(P1, P2, H, C3)$ with household information H are created. The size of this temporary data set equals to the total number of 'other' people. Using this data set, we train another network G_{22} in which only those links initialized from household, owner, spouse attributes to attributes of the 'others' are allowed. From this network we can identify those factors determining the profiles of 'others' and extract a subnetwork with only attributes of 'other' and their parent nodes. Using this subnetwork we can sample 'other' person—conditional on the household, owner, spouse information generated from G_{21}—to fill vacancies [*npax*-2] of household. A similar procedure is applied to estimate two networks G_{31} and G_{32} for those households belonging to type (3). As mentioned, in sampling 'other' people, some of the intra-dependence between those 'other' people is not preserved. Applying this procedure we can create a large list of potential households (say one million) and this list could be used as the base/pool to create dedicated population at zonal levels. For example, when marginal distributions on individual and household level are available, one may apply the generalized raking method for survey sampling to get the weight(or probability) of each household appearing in a particular zone (Deville

[7] http://matsim.org/scenario/singapore Accessed August 8, 2015.

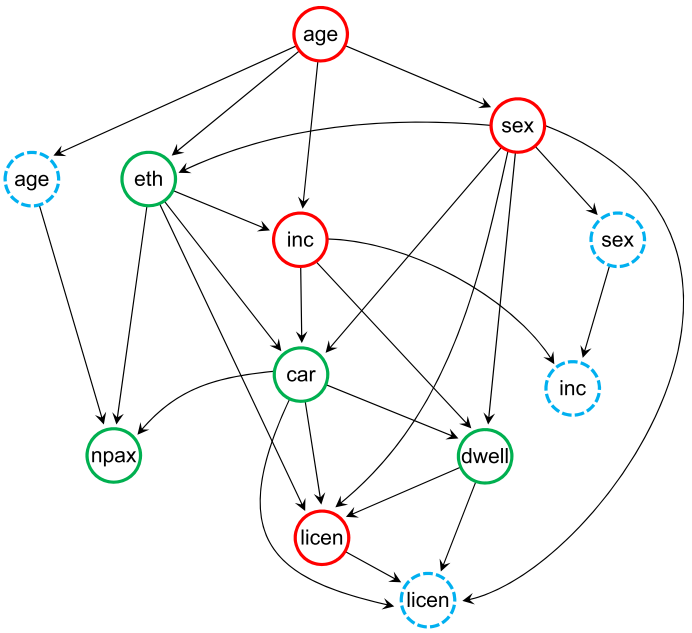


Fig. 5. Model structure G on household[green]/owner[red]/spouse[blue] information. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

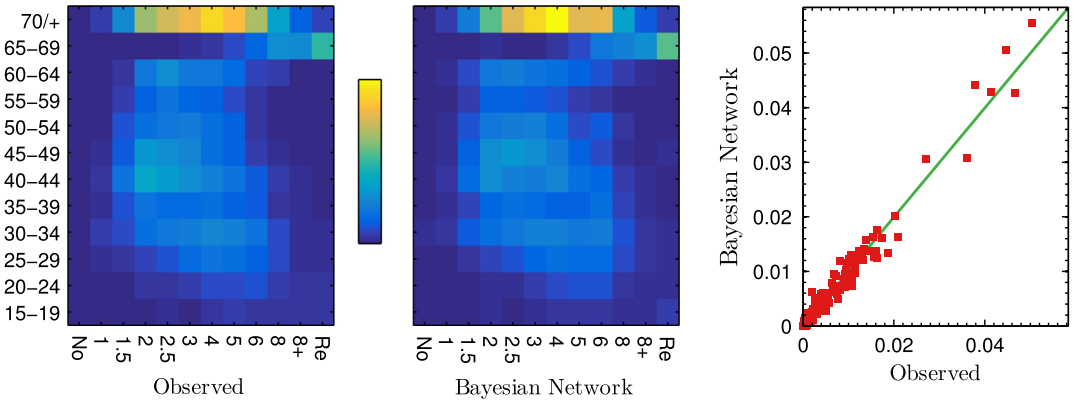


Fig. 6. Joint distribution of owner income and spouse age (10% PUMS).

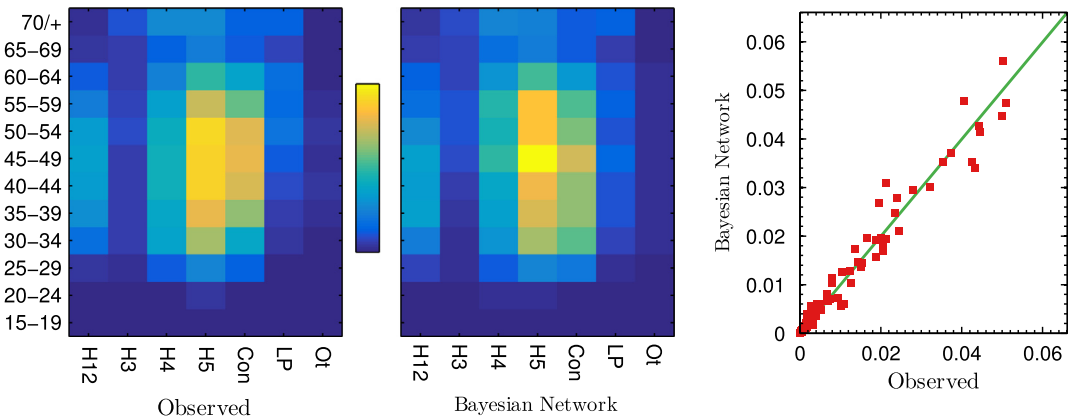


Fig. 7. Joint distribution of dwelling type and spouse age (10% PUMS).

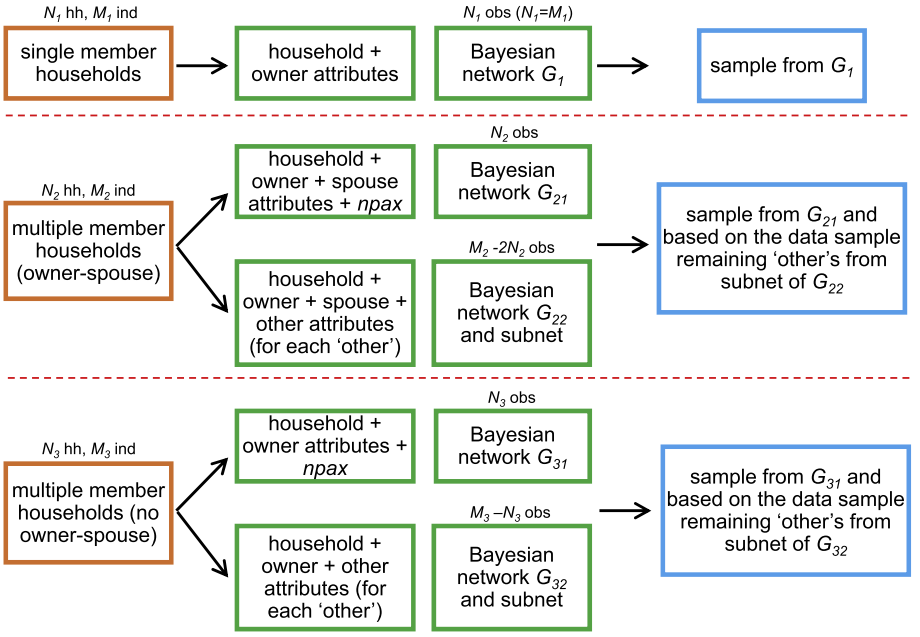


Fig. 8. Procedure to generate full synthesis population.

et al., 1993; Casati et al., 2015). Sampling from the one million potential household given this weight will create a realistic subset that meets the zonal marginal distributions/control totals on both individual and household levels.

5. Conclusions and discussion

In this paper, we introduce a new Bayesian network based approach for population synthesis. The Bayesian network integrates graphical thinking into probabilistic modeling, serving as an efficient tool to reproduce the underlying joint distribution when privacy and confidentiality is of primary concern.

As a popular model with the machine learning communities, the Bayesian network is devoted to identifying a concise structure that is able to capture and reproduce to complex dependence and higher-order interactions among a large set of variables. With the increasing number of variables of interest in emerging micro-simulation models, conventional approaches are essentially trapped by the curse of dimensionality, overfitting, resolution and scalability issues. For example, the curse of dimensionality can easily break the IPF and MCMC algorithm if one intends to sample the full structure at once. To this end, we show that Bayesian network models avoid these issues smartly, by abstracting the structure of population systems using a DAG and local conditional probabilities. In this sense, the proposed Bayesian network model can be considered as a MCMC with partial/incomplete conditionals. However, instead of choosing partial/incomplete conditionals arbitrarily, Bayesian network tries to identify the optimal structure, which in turn is easy to interpret and communicate.

Regarding the tradeoff between model complexity and robustness, the Bayesian network model avoids the overfitting by introducing penalty on size of parameters. The problem of overfitting appears when the training data is not fully representative to the underlying relationship, which is case for the population synthesis (i.e., a large sparse contingency table with only 1–5% observations). Thus, considering the large number of parameters in full conditional distributions (in discrete choice/multinomial linear logistic models), the full conditional MCMC approach is inevitable to the risk of overfitting. In other words, by using full conditional one may fit the PUMS very well given the excessive number of parameters, but fail to characterize the real underlying population structure.

Our results also suggest that a general population system in nature, or at least in the case of Singapore, is very well structured. On the other hand, such structural information of a population system can actually be revealed by a very limited number of observations. This is rather counterintuitive and suggests that the full structure can be well characterized by partial knowledge based on the network structure. In this way, the information from observations are extracted more efficiently. Taking the Bayesian network model demonstrated in Section 4 as an example, we can achieve a very good fit to the underlying population by only using a 1% sample (96 individuals). This is almost infeasible by using conventional approaches. Therefore, for the purpose of privacy protection, Bayesian network can indeed outperform other existing techniques. In terms of heterogeneity, the conciseness of a Bayesian network model allows us to enrich the sampling pool, yet avoids structure zeros in the meanwhile. By integrating other advanced methods such as model averaging, one can also capture uncertainty and enrich the heterogeneity that cannot be characterized by a single network (Dash and Cooper, 2004). The Bayesian

network approach also enables us to efficiently deal with incomplete/missing data. Despite the estimation of model parameter in a known network with missing data, which can be solved by using the Expectation–Maximization (EM) algorithm, the Bayesian network is also able to handle the case that observations are incomplete for an unknown network by integrating the structural EM (SEM) algorithm (Friedman, 1998).

Acknowledgements

The research is funded by the National Research Foundation of Singapore—the funding authority of the Future Cities Laboratory (FCL). We thank the Land Transport Authority of Singapore for providing the HITS2012 data. We are also grateful to the two reviewers for providing us with valuable comments and insightful feedback.

References

Agresti, A., 2002. Categorical Data Analysis, second ed. John Wiley & Sons, Hoboken, NJ.
Akaike, H., 1974. A new look at the statistical model identification. IEEE Trans. Autom. Control 19 (6), 716–723.
Balmer, M., Axhausen, K.W., Nagel, K., 2006. Agent-based demand-modeling framework for large-scale microsimulations. Transport. Res. Rec.: J. Transport. Res. Board 1985, 125–134.
Barthelemy, J., Toint, P.L., 2013. Synthetic population generation without a sample. Transport. Sci. 47 (2), 266–279.
Beckman, R.J., Baggerly, K.A., McKay, M.D., 1996. Creating synthetic baseline populations. Transport. Res. A: Policy Pract. 30 (6), 415–429.
Buntine, W., 1991. Theory refinement on bayesian networks. In: Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence. Morgan Kaufmann Publishers Inc., Los Angeles, CA, pp. 52–60.
Caiola, G., Reiter, J.P., 2010. Random forests for generating partially synthetic, categorical data. Trans. Data Privacy 3, 27–42.
Casati, D., Müller, K., Fourie, P.J., Erath, A., Axhausen, K.W., 2015. Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. In: Transportation Research Board 94th Annual Meeting. Washington, D.C.
Castillo, E., Menéndez, J.M., Sánchez-Cambronero, S., 2008. Predicting traffic flow using bayesian networks. Transport. Res. B: Meth. 42 (5), 482–509.
Dash, D., Cooper, G.F., 2004. Model averaging for prediction with discrete Bayesian networks. J. Mach. Learn. Res. 5, 1177–1203.
Deming, W.E., Stephan, F.F., 1940. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann. Math. Stat. 11 (4), 427–444.
Deville, J.-C., Särndal, C.-E., Sautory, O., 1993. Generalized raking procedures in survey sampling. J. Am. Stat. Assoc. 88 (423), 1013–1020.
Farooq, B., Bierlaire, M., Hurtubia, R., Flötteröd, G., 2013. Simulation based population synthesis. Transport. Res. B: Meth. 58, 243–263.
Friedman, N., 1998. The bayesian structural EM algorithm. In: Proceedings of the Fourteenth conference on Uncertainty in Artificial Intelligence. pp. 129–138.
Friedman, N., Koller, D., 2003. Being bayesian about network structure. a bayesian approach to structure discovery in bayesian networks. Mach. Learn. 50 (1–2), 95–125.
Glover, F., Laguna, M., 1997. Tabu Search. Kluwer Academic Publishers, Norwell, MA.
Guo, J.Y., Bhat, C.R., 2007. Population synthesis for microsimulating travel behavior. Transport. Res. Rec. 2014, 92–101.
Heckerman, D., 1998. A tutorial on learning with bayesian networks. In: Jordan, M. (Ed.), Learning in Graphical Models. MIT Press, Cambridge, MA, pp. 301–354.
Heckerman, D., Mamdani, A., Wellman, M.P., 1995. Real-world applications of bayesian networks. Commun. ACM 38 (3), 24–26.
Janssens, D., Wets, G., Brijs, T., Vanhoof, K., Arentze, T., Timmermans, H., 2006. Integrating bayesian networks and decision trees in a sequential rule-based transportation model. Eur. J. Oper. Res. 175 (1), 16–34.
Koller, D., Friedman, N., 2009. Probabilistic Graphical Models: Principles and Techniques. MIT press, Cambridge, MA.
Müller, K., Axhausen, K.W., 2011. Population synthesis for microsimulation: state of the art. In: Transportation Research Board 90th Annual Meeting. Washington, D.C.
Pearl, J., 2000. Causality: Models, Reasoning, and Inference. Cambridge University Press, Cambridge, UK.
Pritchard, D.R., Miller, E.J., 2012. Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. Transportation 39 (3), 685–704.
R Core Team, 2015. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. <https://www.R-project.org>.
Reiter, J.P., 2005. Using CART to generate partially synthetic public use microdata. J. Off. Stat. 21, 441.
Robert, C., Casella, G., 2004. Monte Carlo statistical methods,, .. Springer Texts in Statistics, second ed. Springer, New York, NJ.
Robinson, R.W., 1973. Counting labeled acyclic digraphs. In: Harary, F. (Ed.), New Directions in the Theory of Graphs. Academic Press, New York, pp. 239–273.
Ryan, J., Maoh, H., Kanaroglou, P., 2009. Population synthesis: comparing the major techniques using a small, complete population of firms. Geogr Anal 41 (2), 181–203.
Salvini, P., Miller, E.J., 2005. ILUTE: an operational prototype of a comprehensive microsimulation model of urban systems. Networks Spatial Econ. 5 (2), 217–234.
Schwarz, G., 1978. Estimating the dimension of a model. Ann. Stat. 6 (2), 461–464.
Scutari, M., 2010. Learning bayesian networks with the bnlearn R package. J. Stat. Softw. 35 (3), 1–22.
Sebastiani, P., Ramoni, M., 2001. On the use of bayesian networks to analyze survey data. Res. Off. Stat. 4 (1), 53–64.
Voas, D., Williamson, P., 2001. Evaluating goodness-of-fit measures for synthetic microdata. Geogr. Environ. Modell. 5 (2), 177–200.
Waddell, P., 2002. Urbansim: Modeling urban development for land use, transportation, and environmental planning. J. Am. Plan. Assoc. 68 (3), 297–314.
Xie, C., Waller, S.T., 2010. Estimation and application of a bayesian network model for discrete travel choice analysis. Transp. Lett.: Int. J. Transport. Res. 2, 125–144.
Ye, X., Konduri, K., Pendyala, R.M., Sana, B., Waddell, P., 2009. Methodology to match distributions of both household and person attributes in the generation of synthetic populations. In: Transportation Research Board 88th Annual Meeting. Washington, D.C.
Zhang, K., Taylor, M.A.P., 2006. Effective arterial road incident detection: a bayesian network based algorithm. Transport. Res. C: Emerg. Technol. 14 (6), 403–417.
Zhu, Y., Ferreira, J., 2014. Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. Transport. Res. Rec.: J. Transport. Res. Board 2429, 168–177.