Review of population synthesis methodologies

Jeppe Rich[#], Gunnar Flötteröd[¤], Sergio Garrido[#], Francisco Pereira[#]

# DTU Transport, Technical University of Denmark

¤ Swedish National Road and Transport Research Institute

Abstract

In the paper we provide a definition of the term 'population synthesis' and discuss it's important and its challenges from a transport demand perspective. Moreover, we offer a classification of the different methodologies and offer the conclusion that, a major distinction between the different methods, whereas these come from the transport demand literature or the spatial simulation literature, is whether these are based on a deterministic or probabilistic approach. The first approach assumes that the sample, on the basis of which the population is constructed, is a ground truth. Contrarily to this, the probabilistic approach assumes that the original sample is one of many realisations from a ground truth distribution. The paper then moves on to further classify probabilistic methods into data-based and model-based simulation methods. The former refers to the case where proxies for probability are formed by drawing from conditionals and margins, whereas the latter refers to the case where a model-layer is introduced between the original data and the re-sampling of agents. As part of a final discussion the paper considers the different methods from a challenge and opportunity perspective and discusses the role of 'zeros', scalability, sample diversity and validation.

Keywords: Population synthesis, Transport modelling, Micro-simulation, Matrix fitting, Forecasting.

1 Introduction

Population synthesis is concerned with the prediction of population composition in geographical and social spaces (Müller and Axhausen, 2010; Harland et al., 2012). In recent years, it has received increasing attention, which can be attributed to its unquestionable importance for transport predictions and the fact that it is far from trivial. In particular, the growing interest in applying quantitative and person-disaggregate (agent-based) transport models for assessing long-term transport policies, coupled with a strong underlying need for analysing social and geographical distribution effects, has called for an increased focus on how to model individual travellers. This development has been further catalysed by an upsurge in new individual-level data sets, i.e. 'big data' which in combination with traditional data sources enable new modelling possibilities.

Population synthesis is not a clearly delineated research area. It has evolved in many areas, often in parallel, and has deployed different statistical and/or numerical methods and has been applied at varying scales and for different purposes (Rich, 2018; Tanton, 2014; Arentze et al. 2007; Beckman et al. 1992; Daly 1998). This motivates the present article's attempt to develop a general framing of the problem and to provide a comprehensive overview of the state-of-the art within this framing. The following definition will serve as a qualitative starting point:

"Population synthesis is the creation of a synthetic population representing individuals and possibly households such that the synthetic population is - in all of its relevant statistical properties - representative for a target population within a given geographical and socio-demographic reference frame."

In its current form, judged by the way it has developed and been applied, **population synthesis has been less of a modelling exercise** than it has been a procedure for disaggregating and consistently relating socio-demographic data from different data sources. Typically, the structural modelling takes place outside the population synthesis, either before, as a mean to produce consistent targets (e.g., from quantitative demographics) and/or as part of a subsequent transport modelling stage that uses the synthetic population as a travel demand representation. **Land-use modelling plays an accompanying role** in this context by providing a model-based representation – often assuming spatial equilibria – for markets that trade locations directly or indirectly through price mechanisms. Prominent examples are the housing market and the labour market.

1.1 Why it is important and increasingly so

Policy makers and hence **transport planners** ask for increasing levels of geographic and socio-demographic resolution in (computerized) transport scenario analysis.

The widely acknowledged linkage between transport demand and social and spatial preferences, in particular between income, household composition and urban form, is considered the main driver of transport demand (Dieleman et al., 2002; Bhat and Srinivasan, 2005). A detailed representation of the traveller population is hence needed: Individual travel demand is derived from individual activity demand, which in turn is highly person- and context specific (Rasouli and Timmermans, 2014). Consequently, different formations of the population will lead to different transport demands and travel patterns (Bento et al. 2005; Stead et al. 2000) and possibly different congestion dynamics (DeSalvo and Huq, 1996; Donovan and Munro; 2013).

From a planning perspective, there is increasing focus on urban areas and a need to analyse soft modes such as walking and biking at a detailed geographical level. More generally, there is a need to evaluate small-scale projects for urban areas, which because of a higher population density requires a more detailed spatial perspective. Also, high-resolution socio-demographic population representations are increasingly relevant as a mean to support and investigate policies related to equity impacts, aging populations, household composition and the combination of age and urbanisation.

The need to answer quite practical questions of the above type is paralleled by an appetite for more details in the spatial social research domain (e.g, Curtis and Perkins 2006). Overall, this has led to recent developments within the **transport modelling** area towards agent-based (person-centered) modelling (Bowman and Ben-Akiva, 2000; Bradley et al. 2010; Nagel and Flötteröd, 2012). This in turn has led to an imminent need for population synthesis techniques to provide (current and forecasted) input data to transport models (Guo and Bhat, 2005; Miller, 2003).

2 Methodological overview

Figure 1 provides an overview of existing methods and a classification of these. In a first classification, methods are distinguished according to whether they adopt a deterministic or a probabilistic perspective on the population synthesis problem; this corresponds to the two main columns in the figure. A three-stage structure can then be identified, which is graphically represented through the three rows “Method/Model”, “Population representation” and “Post-processing”.

The first stage for the class of deterministic models relies on a multi-dimensional array representation of the population, which exhibits one dimension per person attribute. Given a finite number of possible value classes per attribute, each entry in this array represents the number of individuals exhibiting a combination of value classes, one for each attribute. “Sample expansion” then means (i) inserting available census or other sample data into the array, followed by (ii) an up-scaling of these entries such that the total population size is reproduced. This process constitutes the arguably simplest instance of the broader class of procedures classified as “reproduction methods” where supplementary information, in particular one- or multidimensional margins, are accounted for. When limiting the process to a “sample expansion”, all array entries without original sample data remain zero, meaning that one can limit the population representation to a set of (sample, weight) tuples where the weight represents the amount by which the sample has been scaled up.

If one wishes to arrive at a person-based population representation, the result of a deterministic population generation needs to be post-processed by either (i) rounding the array entries and generating according numbers of synthetic persons or (ii) using the array entries as sampling weights according to which a synthetic population is drawn. Naive rounding may lead to a downward bias if there are many near-zero weights. Sampling may introduce unwanted stochasticity (leading, for instance, to incompatibilities with available population marginals), in particular if synthetic individuals are drawn independently, which may require further post-processing.

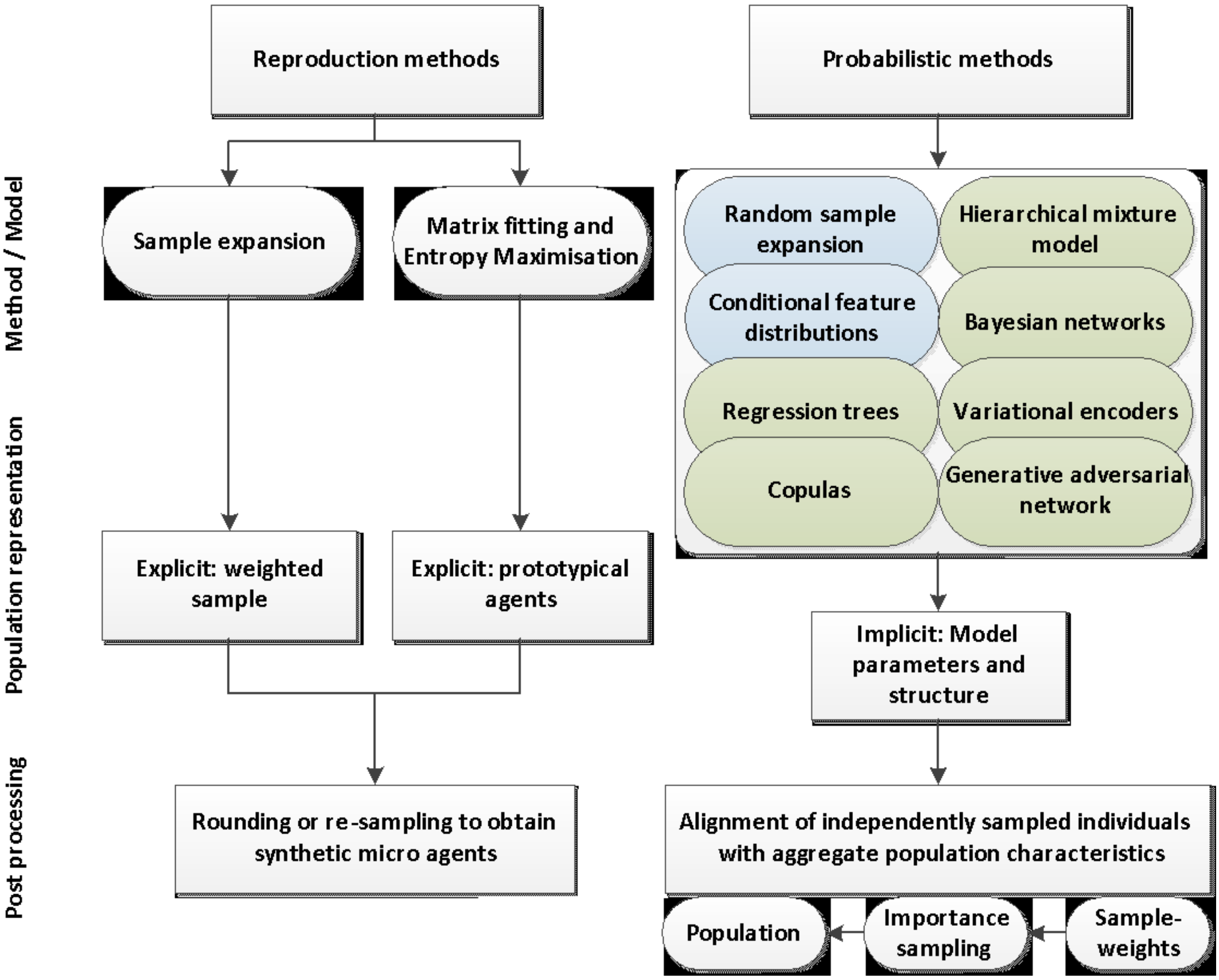
Reproduction methods
Probabilistic methods
Method / Model
Population representation
Post processing
Sample expansion
Matrix fitting and Entropy Maximisation
Random sample expansion
Hierarchical mixture model
Conditional feature distributions
Bayesian networks
Regression trees
Variational encoders
Copulas
Generative adversarial network
Explicit: weighted sample
Explicit: prototypical agents
Implicit: Model parameters and structure
Rounding or re-sampling to obtain synthetic micro agents
Alignment of independently sampled individuals with aggregate population characteristics
Population
Importance sampling
Sample-weights

Figure 1: Classification of population synthesis methods.

Probabilistic methods, shown in the second column, bypass the array-based population representation and use other, often more compact approximations of the population's feature distribution. Here, the spectrum of available methods is larger, given that one has moved on from the array-based "counting of instances" to the more general and broadly studied problem of approximating high-dimensional probability distributions. The resulting population representation then depends very much on which approximation method is used. Postponing an enumeration of existing techniques to the subsequent section, it is here merely noted that most, if not all operational techniques, model the feature distribution of individual persons or households but do not include information about the joint distribution of multiple individuals/households, as it is represented by one- or multi-dimensional marginals. Here, a person-based population representation is obtained by drawing individual/household-features from the approximate population distribution, followed again by a possible post-processing step if one wishes to remove sampling noise and/or ensure the reproduction of marginals.

3 Review of deterministic and probabilistic methods

Having set the stage with respect to an overall classification of methods in the previous section, this section discusses and compares the deterministic branch of synthesis methods in greater detail.

The two most well-known approaches for accomplishing deterministic population synthesis are (i) matrix fitting (Tanton et al., 2014) and (ii) sample expansion. Often the two methods are distinguishable from the input data they are applied to. Matrix fitting methods are, as the word suggest, typically applied to *matrices of prototypical individuals* with the stratification of the population being defined from the dimensions of the matrix. Sample expansion comes in different flavours and has been referred to as *prototypical sample enumeration* (Daly, 1998), *synthetic reconstruction* (Ballas et al., 2005) and *deterministic re-weighting* (Harland et al. 2012). The simplest approach is when applied to a survey of agents for which expansion factors are calculated to align the weighted population according to aggregated population targets. This approach, which we will refer to as *prototypical sample enumeration* shares many similarities with matrix fitting and represent in essence, an *up-scaling* of the sample data, based on a limited number of weights (Smith et al., 2009). Alternative approaches from the micro simulation literature typically generate individual specific weights through iterative schemes.

Probabilistic models assume a distributed ground truth population and aim at drawing possible populations from the underlying distribution. Probabilistic methods can be classified into i) methods that are re-sample from a given data set, and ii) methods that sample from a model that has been estimated from a given data set. We will refer to this as *data-based resampling* and *model-based resampling*.

Generally speaking, this classification is also an indirect distinction between historical probabilistic methods, which has almost solely used *data-based resampling*, and the newer branch of model-based resampling techniques.

Model-based resampling methods for population synthesis represent, until now, a largely unexplored field when it comes to applications for population synthesis. However, from the perspective of the authors, it is a field that holds great potential for this particular problem. In the following the papers by Sun et al (2015), Sun et al. 2018, Borysov et al. (2018) and Albert et al. (2018) will be discussed, however, as the field represent a new direction within the field of population synthesis an introduction is offered.

4 Summary and conclusions

The paper offers a detailed description of the current state-of-art for population synthesis. The following aspects are discussed in the paper:

i) the delineation across fields
ii) why it is important and increasingly so
iii) a classification of methodologies
iv) Distinction between deterministic and probabilistic methods
v) A challenge and solution perspective
vi) Future research perspectives

With respect to the challenges and solutions, we discuss issues related to structural zeros, scalability, constraints, agents versus prototypical agents, diversity and prediction reliability. With the upcoming focus on agent based modelling and the accelerated need for details in the social and geographical the population synthesis problem becomes increasingly important and this paper is an attempt to describe the current state of art and provide of a direction of the recent research.

Literature

Albert, A. Strano, E. Kaur, J., Gonzalez, M. (2018). MODELING URBANIZATION PATTERNS WITH GENERATIVE ADVERSARIAL NETWORKS. https://arxiv.org/pdf/1801.02710.pdf

Arentze, T., Timmermans, H. and Hofman, F. (2007) Creating Synthetic Household Populations - Problems and Approach. Transportation Research Record, No. 2014, 85-91.

Ballas, D., Clarke, G., Dorling, D., Eyre, H., Thomas, B., & Rossiter, D. (2005). 'SimBritain: a spatial microsimulation approach to population dynamics', Population, Space and Place, 11(1), 13–34, doi:10.1002/psp.351

Beckman, J.R., Baggerly, K.A. and McKay, M.D. (1996) Creating Synthetic Baseline Populations. Transportation Research Part A, 30(6), 415-429.

Bowman J.L., M.E. Ben-Akiva, 2000. Activity-based disaggregate travel demand model system with activity schedules. *Transportation Research Part A* 35(1):1–28.

Borysov, S., Rich, J., and Pereira, F. (2018). Scalable population synthesis with deep generative modelling. Url: https://arxiv.org/abs/1808.06910

Bradley, M., Bowmanb, J.L., Griesenbeckc, B. (2010). SACSIM: An applied activity-based model system with fine-level spatial and temporal resolution. *Journal of Choice Modelling* 3(1):5-31.

Daly, A. (1998) Prototypical Sample Enumeration as a basis for forecasting with disaggregate models. Proceedings of the European Transport Annual Meeting. PTRC, London, pp. 225-236.

Dieleman, F. M., Dijst, M., & Burghouwt, G. (2002). Urban Form and Travel Behaviour: Micro-level Household Attributes and Residential Context. Urban Studies, 39(3), 507 - 527.

Farooq, B., Bierlaire, M., Hurtubia, R., Flötteröd, G. (2013) Simulation Based Population Synthesis. *Transportation Research Part B: Methodological* 58, (2013): pp 243-263.

Guo, J.Y., Bhat, C.R. 2005. Population Synthesis for Micro Simulating Travel Behaviour. CAEE Working paper. Url: http://www.caee.utexas.edu/prof/bhat/ABSTRACTS/SPG_Guo_Bhat.pdf

Harland, K, Heppenstall, A, Smith, D., Birkin. M.H. (2012) Creating realistic synthetic populations at varying spatial scales: A comparative critique of population synthesis techniques. Journal of Artificial Societies and Social Simulation, 15 (1). 1. ISSN 1460-7425

Miller, E. J. (2003). Land Use: Transportation modeling. In K. G. Goulias (Ed.), *Transportation systems planning methods and applications* (Vol. 2, pp. 12-1-22). Boca Raton: CRC Press. doi: 10.1201/9781420042283.ch1

Müller K., Axhausen K.W. (2010) “Population synthesis for micro simulation: State of the art”, Paper Presented at STRC conference, http://www.strc.ch/2010/Mueller.pdf

Rich, J. (2018), Large-scale spatial population synthesis for Denmark. European Transport Research Review. https://doi.org/10.1186/s12544-018-0336-2

Rich, J., Mulalic, I., 2012. Generating synthetic baseline populations from register data. Transportation Research Part A 46, 467–479.

Smith, A., Lovelace, R. and Birkin, M. (2017). Population Synthesis with Quasirandom Integer Sampling. Journal of Artificial Societies and Social Simulation 20 (4) 14.

Url: http://jasss.soc.surrey.ac.uk/20/4/14.html

Sun, L., Erath, A., Cai, M., 2018. A hierarchical mixture modeling framework for population synthesis. Transportation Research Part B: Methodological 114, 199–212

Sun, L., Erath, A., 2015. A Bayesian network approach for population synthesis. Transportation Research Part C: Emerging Technologies 61, 49–62.

Tanton, R. 2014. A Review of Spatial Microsimulation Methods. INTERNATIONAL JOURNAL OF MICROSIMULATION (2014) 7(1) 4-25