Construction and forecast of synthetic households - Experience from the Danish National Model

Jeppe Rich *

Stefan Lindhard Mabit

Department of Transport,

Technical University of Denmark, Kgs. Lyngby, Denmark

* Email: mailto:jr@transport.dtu.dk

1 Introduction

One of the most fundamental elements of any demand analysis is to be able to properly represent the underlying population. This includes a representation of the socio-economic profile of respondents and possibly how respondents are linked within households. This representation is needed in the baseline as well as in the application of the model, which in case of the latter typically involves forecasting of the population. In transport demand modelling the predominant trend during the last two to three decades has been to work on ever more detailed data representing individuals rather than aggregated prototypes. The introduction of activity-based principles in transport demand analysis as well as the investigation of intra-household linkages among the respective household members have emphasized the need of being able to synthesise individuals as well as the entire composition of households[1].

2 Methodology

In the present paper, we describe a new household population synthesiser, which is developed as part of the Danish National Transport Model (NTM). The methodology is based on two main modules; i) a population fitting module, and ii) a sampling module. Based on the first module, a master table for the entire population is created. This table represents a detailed description of all individuals in the population represented as prototypical individuals. The coverage of the table is very detailed as each entry in the master table corresponds to approximately one individual (e.g., approximately 5.2 million entries). The table is fitted

---

[1] A state-of-art summary of population synthesis can be found in [2].

using iterative proportional fitting conditional on flexible linear constraints to restrict the generated population to detailed targets. In the sampling module, we apply random sampling and use additional probability margins describing spouse/husband probability match (SMT table), type of kid matching (TKT table) and number of kids (NKT table) to link individuals to households. A brief description of the most important input tables is given in the Abbreviation list.

The population fitting module [1] produces a master table for the population. The master table is spanned by seven variables (four household specific variables and three individual specific ones) as shown in Table 1 below.

Variables	Description	Type
ZoneID	Residential zone (907 NTM zones)	Household
HHIncome	Household income, 12 classes	Household
Single	2 classes (single or not)	Household
Kids	2 classes (kids or not)	Household
Gender	2 classes (male or female)	Individual
Age	10 age classes	Individual
LMA	Labour market association, 6 classes	Individual

Table 1: Dimensionality of the master population table.

The master table can be forecasted by adjusting the targets to match future levels and then subsequently running the fitting routing. The next step is to link the prototypical individuals in the master table to households. Clearly, the simplest approach would be to extend the master table to include e.g. the age of all members in the household and then fit this matrix accordingly. However, this is infeasible as the dimensionality of the matrix gets too large. A feasible way is to apply a sampling scheme in which individuals are linked across households based on random sampling.

The sampling scheme is based on the following overall steps:

1. Consider the master table from Table 1 and extend the table with a variable representing the adult status of the category. This is based on a sampling based on the Adult Probability Table (APT). Refer to the output table as the Extended Master Table (EMT).

2. Construct an Aggregate Household Table AHT by summing Table 1 according to <ZoneID, HHIncome, Single, Kids>. The counts representing the sum for each household class is rounded to nearest integer.

3. Let $k = 1, \ldots, K$ represent the different aggregated household classes and N_k the number of households within each class. Initialise $k = 1$.

4. Let $i = 1, \ldots, N_K$ represent the individual households within each class. Initialise $i = 1$.

5. For $\{i, k\}$ do the following;

    a. Sample first adult.

    b. If Single = 0 sample based on the first adult a second adult.

    c. If Single = 1 go to 5d).

    d. If Kids = 1 sample based on the household characteristics and the characteristics of the adults (there may be one or two) the expected number of kids $NK_{i,k}$ represented as an integer value.

    e. If $NK_{i,k} > 0$, sample $1, \ldots, NK_{i,k}$ kids.

6. While $i < N_k$ let $i = i + 1$ and go to 5). If $i = N_k$ go to 7).

7. While $k < K$ let $k = k + 1$ and go to 4). If $k = K$ go to 8).

8. End of sampling.

The paper will give a detailed description of the different sampling modules applied in step 5a), 5b), 5d) and 5e).

The outcome of the above sampling scheme is a population of individuals, which is consistently linked within households. Although the sampling is strongly controlled by the number of households in the AHT table it is still a random population within these boundaries. As a result, it is necessary to generate different populations using Monte Carlo to make sure that those dimensions not controlled in the AHT table replicate the actual population profile.

2 Results

In the paper we investigate the precision of the synthesiser in a back-casting exercise. This is possible as we have full access to Danish register data and (will) know the exact population profile 30 years back in time. As there are many sources to uncertainty; i) the population fitting, ii) the sampling scheme, and iii) use of marginal probability tables in the sampling scheme as well as in the population fitting, we specifically investigate how the different input sources affect the precision of the final population.

References

[1] Rich, J., Mulalic, I., ``Generating synthetic baseline populations from register data'', Transportation Research Part A 46(3), 467-479, 2012.

[2] Müller, K., Axhausen, K.W., ``Population synthesis for micro simulation: State of the art'', Paper Presented at STRC conference, 2010,
url: http://www.strc.ch/conferences/2010/Mueller.pdf, 2010.

Abbreviation list

NTM: (Danish) National Transport Model.

MT: Master Table (Table 1).

EMT: Extended Master Table (MT extended with a row for being adult).

AHT: Aggregate Household Table constructed from the EMT. If we let the number of individuals in EMT be represented by val, the AHT is based on valH defined as;

$if\ adult = 1\ and\ single = 1\ then\ valH = val$

$if\ adult = 1\ and\ single = 0\ then\ valH = val/2$

The AHT is then established by summing the valH over the AHT classes <ZoneID, HHIncome, Single, Kids>

APT: Adult Probability Table. Represent the probability of being "adult" in a given household.

SPM: Spouse Match probability Table. Represent the probability of a spouse/husband "match" between two adults.

NKT: Number of Kids Table. Represent the probability of having x number of kids in a given household when the profile of the adults are known.

TKT: Type of Kid Table. Represent the probability of kid "match" in a given household when the profile of the adults are known.