Rich *European Transport Research Review* (2018) 10:63
https://doi.org/10.1186/s12544-018-0336-2

European Transport Research Review

ORIGINAL PAPER Open Access

Large-scale spatial population synthesis for Denmark

Jeppe Rich

Abstract

The recent development in micro-based transport models is a major step towards an improved understanding of transport demand and its underlying drivers. By adapting a detailed geographical resolution level and a fine-grained social description of individuals it becomes possible to investigate distribution effects across social classes and geographical spaces, elements which were not possible to take into account until recently. However, the increasing amount of details comes at a cost. As the prediction-space is enlarged, models become increasingly dependent on the quality of inputs and exogenous model assumptions of which the formation of synthetic population forecasts is by far the most important one. The paper presents a coherent description of a large-scale population synthesis framework involving all relevant steps in the synthesis stages from target harmonisation, matrix fitting, post simulation of households and agents and reweighting of the final population. The model is implemented in the Danish National Transport Model and is aimed at predicting the entire Danish population at a very detailed spatial and social level. In the paper we offer some insight with respect to the propagation of sampling noise caused by the household simulation stage and a brief validation of the model when comparing a modelled 2015 population with observed data.

Keywords: Population synthesis, Travel demand forecast, Model validation, Spatial micro-simulation

1 Introduction

A fundamental input to any agent-based model, that being a large-scale transport model or a network simulation model, is a population of agents. The agents should ideally be formed in such a way that they represent a realistic picture of the population at the time of the scenario and such that they are adequately detailed with respect to socio-economic classes and the geography in which they are measured. In the literature, this problem is referred to as *'population synthesis'* and has, with the upsurge of agent-based modelling, received increasing attention in recent years. In this paper, a large-scale population synthesis approach is presented for Denmark. The model is applied in the Danish National Transport model [29] and extends previous work [28] by giving attention to all stages of the population synthesis from harmonisation of constraints to the micro-simulation stage where individuals are allocated to households.

The process by which populations are constructed, is by no mean trivial as it needs to; i) form a representative picture of the population in the modelling area for a given base year, ii) facilitate forecasting according to projected population targets, iii) be sufficiently detailed at the geographical and socio-economic level to support the requirements of a detailed transport model and allow for heterogeneous preferences across the population, and iv) recognise individuals as being part of a household [20]. In addition, from a technical perspective, population synthesis is challenging because uncertainties in the population synthesis model will propagate through subsequent modelling steps in the transport model and bias the final equilibrated model output [36]. As a result, attention should be given to how these populations are formed and the implications of error propagation.

In recent years there has been an increasing awareness of the importance of the population stage which has led to methodological developments and an increase in applications [25, 34]. Reasons for this increased attention includes, among other things, increased awareness of the

Correspondence: rich@dtu.dk
Department of Management Engineering, Technical University of Denmark, 2800 Kgs. Lyngby, Denmark

importance of distribution effects, e.g. equity impacts, generation effects and gender effects of transport policies [9]. Hence, transport models of today should not only address *'size effects'* but also who is likely to be affected by certain policies. Also, there is an increasing awareness that the spatial and social formation of the population has a significant impact on the demand for transport [6, 33]. The ongoing agglomeration trend with people moving to the cities leads to potentially increasing congestion, decreasing trip distances, and changed mode-shares all of which are catalysed by population changes [14, 15]. Figure 1 below illustrates the speed of which urbanisation takes place in Denmark according to projections carried out by Denmark Statistics. These projections are included as baseline population forecasts in the population synthesis model, although at much more detailed levels as illustrated in Tables 4, 5, 6 and 7 in Section 3. It is also worth stressing that the urbanisation process in Denmark is also a socially diverse process where younger people move to the urban areas whereas older people move to the suburban areas or even out of the cities.

1.1 Literature review

Population synthesis has been approached from different methodological perspectives [19, 25], including reweighting approaches, matrix fitting approaches and simulation-based approaches.

Approaches based on reweighting typically aim at estimating weights or "expansion factors" which can be used to expand surveys or smaller samples such that these are representative for the populations. Reweighting based on quadratic optimisation has been proposed in Daly [10], whereas others have used combinatorial optimisation [1, 30, 35] and maximum cross-entropy [3, 17, 22]. It should be noted that reweighting and matrix fitting are closely connected as a matrix fitting indirectly generates weights relative to a starting solution. Matrix fitting approaches typically apply different variants of Iterative Proportional Fitting [13], which may include multi-level fitting [27] and intermediate stages to circumvent missing values in the starting solution. The correspondence between cross-entropy and IPF under convex constraints has been covered in McDougall [23], Dykstra [12] and Darroch and Ratcliff [11] who showed that IPF throughout the iteration scheme increases entropy monotonously. Applications of IPF have been presented in Beckman et al. [5], Arentze et al. [2], and Simpson and Tranmer [31]. Other slightly more advanced applications can be found in Pritchard and Miller [27] in which Hierarchical IPF procedures were proposed for fitting household and individuals jointly. Another application is in Rich and Mulalic [28] which proposed a pre-harmonisation procedure in order to account for inconsistent targets. Recently there has been an increased interest for using simulation based approaches to generate synthetic populations. Mostly these approaches has been concerned with the problem of generating synthetic 'pools' of individuals by sampling from an original data source. Farooq et al. [16] present an application of a Gibbs sampler whereas Borysov et al. [4] suggest using

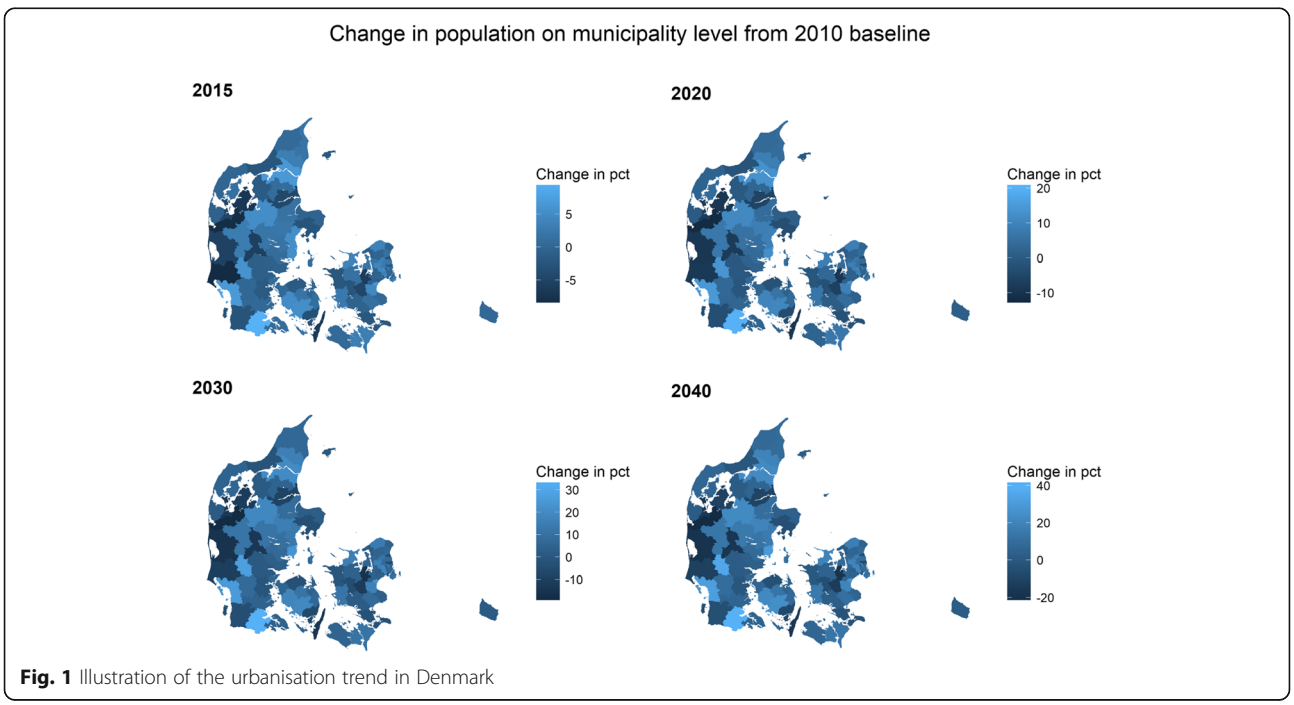

Fig. 1 Illustration of the urbanisation trend in Denmark

deep generative modelling in order to address scalability issues and sparsity in the origin data. However, simulation-based approaches are not new to population synthesis [7] and have often been used in post allocation stages, e.g. to generate agents and household entities from aggregated population matrices. In this sense, most applied frameworks for population synthesis are "hybrids" that combine matrix fitting with simulation stages although these combined applications are rarely described in the literature. Even pure simulation-based population synthesis as presented in Farooq et al. [16] and Borysov et al. [4] typically require a re-sampling stage where individuals are 'importance-sampled' such that the resulting population are consistent with population targets.

While referring to the literature and the many different approaches that have been applied, it should be stressed that choice of strategy for the population synthesis is closely connected with the amount, the type and quality of data at hand. Generally speaking, if detailed high quality data are available as a basis for the starting solution it is important to utilise this information as much as possible. In that case, IPF/Cross Entropy or Markov Chain Monte Carlo approaches are natural choices because they maintain the correlation structure captured in the starting solution well [8] by maintaining odd ratios. Another determining criterion is whether 'hard' or 'soft' targets are required. Certain methods, such as quadratic optimisation [10], do not facilitate 'hard targets' and may not be acceptable from an application perspective or at least require a subsequent quota-based sampling stage.

1.2 Contribution of the paper

The paper provides a detailed description of a large-scale population synthesis framework. The model is implemented as part of the Danish National Transport Model and is used for population synthesis for the entire Danish population. Main contributions to the existing literature are as follows. First, to our knowledge, almost no population synthesis frameworks have been presented in the scholarly literature describing all stages from target harmonisation, matrix fitting to the final allocation stage. We have deliberately included a complete and coherent description in order to provide a standing example of a self-contained model for population synthesis for an entire country. Secondly, on the more technical side, the paper proposes a simulation-based household allocation procedure which involves the concept of "spouse matching" and "kids matching". Thirdly, we introduce a pre-harmonisation stage of population targets combined with a two-stage IPF fitting procedure to enable adjustments of the fitted matrix and to enforce consistency among targets. Finally, we provide some validation insight by analysing the degree of accumulated household simulation noise and by comparing a prediction from 2010 to 2015 with observed data.

In Section 2 the model framework is presented, which includes two fitting stages, a harmonisation stage and a simulation stage. Section 3 is concerned with model application and provides an overview of data and notation and considers model validation. Finally in Section 4, we offer a conclusion, including a research outlook.

2 Methodology

The population synthesis framework consists of three main stages: i) a harmonisation stage, which is run only once for every set of constraints, ii) a fitting stage where the population matrix is fitted using IPF and iii) a simulation stage where prototypical agents are translated into micro agents and subsequently grouped into households. An illustration of the different stages is provided in Fig. 2 below.

The household simulation stage involves going from an aggregated matrix perspective, e.g. the concept of *"prototypical individuals"*, to a list that essentially includes all 5.5 Million individuals in Denmark for the base year 2010. The objective of the simulation stage is to classify individuals into households in order to support household based decisions in the demand model. A challenge is that the household simulation stage renders stochasticity into the final solution. In order to make sure that the final population corresponds to the targets for the true representative population, a re-weighting approach is applied. This is considered in more detail in Section 2.3.

Notation is introduced while describing the different parts of the model. However, a complete list of notation is offered in Table 14 in Appendix 1.

2.1 Harmonisation of constraints

The harmonisation stage is a pre-processing stage with the sole objective of ensuring consistency between input target constraints. The issue has not received much attention in the scholarly literature where it is common to assume that inputs are correct and consistent. It is also a principle discussion whether we should either correct inputs (via harmonisation) or require inputs to be correct before running the model.

However, from a user and application perspective the auto-harmonisation of targets makes the process of generating scenarios much easier and the application of the model less prone to errors. As an example, if a scenario is concerned with a generic income increase then the user can focus on shifting the conditional income distribution without worrying about the actual population

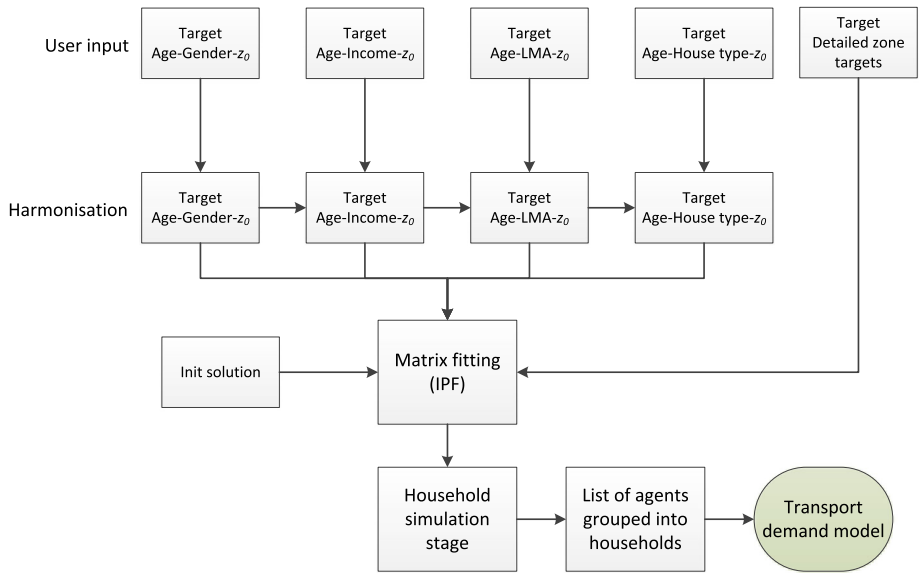

Fig. 2 Overall illustration of the population synthesis structure in The Danish National Transport Model

counts. Another advantage of pre-harmonised targets is that it solves the problem of convergence problems that may occur as a result of rounding errors when defining the targets. As we typically have a convergence threshold below E-6 in the matrix fitting stage this can easily turn into a problem. The harmonisation stage prevents this problem by ensuring consistency at the level of the machine precision. As the harmonisation process has not received much attention in the literature we discuss this in more details below.

Firstly, why is it a problem? When having many targets constraints of which some are cross-linked in that they share common attributes, it is not trivial to make sure that these are consistent across all dimensions. This is a problem that can easily arise when users of the model are editing targets for a given scenario. As a result, summing different targets across attributes that are similar (across targets) may render sums that are different. This in turn will lead to inconsistent input data, which will affect convergence and bias outputs. This is not just a problem that relates to the IPF algorithm but for any approach which is supposed to align a population with future targets one way or the other. Clearly, the simple solution to the problem is to reduce the number of targets and their complexity. However, this generally conflict with the increasing need for more detailed long-term forecasts. Having few and simple targets will make it difficult to capture social processes and trends in the population synthesis stage.

A solution was proposed in Rich and Mulalic [28] in which constraints were harmonised according to a ranking procedure and this procedure has been applied in the Danish National Transport Model. The idea is that the different targets are ranked according to reliability and subsequently adjusted such that lower ranked constraints always comply with higher ranked constraints. The ranking that is used in the National Transport Model is presented in Table 1 below.

The idea is that rather than using targets directly, a harmonised version of the targets is used instead. As a result, it is the harmonised targets that are passed on to the IPF algorithm in order to ensure consistency across inputs and convergence of the algorithm.

Per definition, the target with rank 1 is the 'ground truth'. This target will therefore not be harmonised but is used as a baseline target. The other targets are harmonised based on the levels of the highest ranked target. Below $\tilde{T}_{ai}(z_0, a, i)$ represent the harmonised targets for $T_{ai}(z_0, a, i)$ and similarly for the other targets.

$$\tilde{T}_{ag}(z_0, a, g) = T_{ga}(z_0, g, a) \tag{1}$$

Table 1 Ranking of main targets in the National Model

Target	Rank	Description
$T_{ga}(z_0, a, g)$	1	Targets for municipality (z_0), age classes (a) and gender (g)
$T_{ai}(z_0, a, i)$	2	Targets for municipality (z_0), age classes (a) and income classes (i)
$T_{al}(z_0, a, l)$	3	Targets for municipality (z_0), age classes (a) and labour market association classes (l)
$T_{af}(z_0, a, f)$	4	Targets for municipality (z_0), age classes (a) and family structure (f)

$$P(i|a,z_0) = \frac{T_{ai}(z_0,a,i)}{\sum_i T_{ai}(z_0,a,i)} \quad (2)$$

$$\tilde{T}_{ai}(z_0,a,i) = P(i|a,z_0)\sum_g T_{ga}(z_0,g,a) \quad (3)$$

$$P(l|a,z_0) = \frac{T_{al}(z_0,a,l)}{\sum_l T_{al}(z_0,a,l)} \quad (4)$$

$$\tilde{T}_{al}(z_0,a,l) = P(l|a,z_0)\sum_g T_{ga}(z_0,g,a) \quad (5)$$

$$P(f|a,z_0) = \frac{T_{af}(z_0,a,f)}{\sum_f T_{af}(z_0,a,f)} \quad (6)$$

$$\tilde{T}_{af}(z_0,a,f) = P(f|a,z_0)\sum_g T_{ga}(z_0,g,a) \quad (7)$$

As can be seen, we apply only the relative distribution of $T_{ai}(z_0,a,i)$, $T_{al}(z_0,a,l)$ and $T_{af}(z_0,a,f)$ by i, l and f. As a result, summing over any dimension in all targets will reproduce the same sum. In the baseline all targets are naturally harmonised as they are drawn directly from harmonised register data. However, the harmonisation is mainly supposed to support users of the model who alter the targets for scenario analysis.

In the case described in Table 1, the cross-linking of targets is relative simple in that at most two dimensions are shared. Moreover, both of these two dimensions are included in the highest ranked target which makes it straightforward to harmonise subsequent targets as described. However, if the cross-linking is more complex the harmonisation stage will be equally complicated. Sometimes it may involve running separate IPF steps in the harmonisation. To illustrate this potential problem consider a slightly modified set of targets in Table 2.

Although only three targets are considered with only two attributes it is not possible to apply the ranking procedure introduced above. The problem arise in the harmonisation of $T_{il}(i,l)$ as this target shares attributes with $T_{ai}(a,i)$ and $T_{al}(a,l)$. In other words, we cannot represent $T_{il}(i,l)$ as scaled marginal probabilities as before. However, in this particular case potential inconsistencies that may arise from $T_{il}(i,l)$ can be solved by running a pre-stage IPF.

More specifically, the harmonised target $\tilde{T}_{il}(i,l)$ could be calculated using an IPF algorithm with starting value $T_{il}(i,l)$ and constraints formed by $T_1(i) = \sum_a T_{il}(a,i)$ and $T_2(l) = \sum_a \tilde{T}_{il}(a,l)$. The harmonised $\tilde{T}_{il}(a,l)$ is constructed in a similar way as before. That is, $\tilde{T}_{il}(a,l) = P(l|a)\sum_i T_{ai}(a,i)$.

$$\begin{aligned}\tilde{T}_{il}(i,l) &= IPF(init = T_{il}(i,l), T_1(i) \\ &= \sum_a T_{il}(a,i), T_2(l) \\ &= \sum_a \tilde{T}_{il}(a,l))\end{aligned} \quad (8)$$

The intuition behind this approach is straightforward. The best estimate for $\tilde{T}_{il}(i,l)$ is $T_{il}(i,l)$ except that the matrix might fail to be consistent with $T_{il}(a,i)$ and the previously harmonised target $\tilde{T}_{il}(a,l)$. Hence, we would like to preserve the correlation structure in $T_{il}(i,l)$ as closely as possible while adjusting the row and column sums to be consistent. This *minimal deconstruction* of the probability distribution is preserved under iterative proportional fitting in that odds log ratios are preserved [8].

A final remark that relates to the definition of targets and the harmonisation of these is that it is sometimes observed that the spanned target space includes cells that are not included in the starting solution or the other way around. This essentially corresponds to a situation where targets are inconsistent and will generally lead to convergence problems. It is therefore important to align the dimensionality of the target space and the solution space before starting the iteration process. Typically this is solved as part of a pre-processing stage where the input space and target space are aligned. In practise it may lead to an extended input space to allow for new cell-entries.

Table 2 Ranking of highly cross-linked targets

Target	Rank	Description
$T_{ai}(a,i)$	1	Municipality, age and gender targets
$T_{al}(a,l)$	2	Municipality, age and income targets
$T_{il}(i,l)$	3	Municipality, age and LMA targets

2.2 Fitting of master table for individuals

The fitting of the master table translates into finding the maximum cross-entropy of $T(a,g,i,l,f,c,z)$ provided we have a starting solution $T_0(a,g,i,l,f,c,z)$ and subjected to a set of pre-defined constraints.

The corresponding maximum cross-entropy for the problem at hand is provided in (9) below.

$$\max_{\{T(a,g,i,l,f,c,z)\}} Z = -\sum_{\{a,g,i,l,f,c,z\}} T(a,g,i,l,f,c,z) \ln\left(T(a,g,i,l,f,c,z)/T_0(a,g,i,l,f,c,z)\right) \quad (9)$$

The constraints are provided in (10)–(13).

$$\sum_{i,l,f,c,z\in z0} T(a,g,i,l,f,c,z) = \tilde{T}_{ga}(z0,g,a) \quad (10)$$

$$\sum_{g,l,f,c,z\in z0} T(a,g,i,l,f,c,z) = \tilde{T}_{ai}(z0,a,i) \quad (11)$$

$$\sum_{g,i,f,c,,z\in z0} T(a,g,i,l,f,c,z) = \tilde{T}_{ai}(z0,a,l) \quad (12)$$

$$\sum_{g,i,l,c,,z\in z0} T(a,g,i,l,f,c,z) = \tilde{T}_{af}(z0,a,f) \quad (13)$$

As seen, the right-hand side is the harmonised targets from (1)–(7) in the previous section. The matrix fitting is solved using an IPF algorithm which is carried out in two stages (refer to Appendix 2 for a more elaborate description of the fitting stage). In a first stage the problem that corresponds to the above cross entropy problem is solved. This problem returns the fitted matrix $T^{*}(a,g,i,l,f,c,z)$. In the second stage fitting we allow users to add additional constraints at a more detailed zone level. This introduces a possibility of 'aligning' the fitted solution to additional information that may be available as local projections of zone population. The outcome of the second-stage fitting is a $T^{**}(a,g,i,l,f,c,z)$ matrix.

2.3 Simulation of household entities

The outcome of the population fitting (as we considered in Section 2.2) is a new master table, which conforms to the new targets and to the structure of the master table for individuals (refer to Table 3 in Section 3).

The next stage of the population synthesis is to convert prototypical individuals into micro agents and to allocate these individuals to households. This task is necessary because the master table is not a micro representation of the population but simply a weighted list of prototypical individuals grouped into socio-groups. Although the socio-economic groups are very detailed, certain entries in the master table may represent as much as 30–50 prototypical individuals and others very small fractions of individuals. If we apply micro-simulation directly to the master table it would mean that all of these 30–50 individuals would be treated in a similar way as regards the household sampling, which is not desirable. We therefore create an enumerated list of all individuals in the population and process those in a micro-simulation algorithm where individuals are grouped into households.

Table 3 Structure of starting solution and final master table

Variable	Description	Classes
AgeID (*a*)	Age	10
GenderID (*g*)	Gender	2
IncomeID (*i*)	Personal income	11
LmaID (*l*)	Labour market association	8
FamID (*f*)	Family structure	2
NumChildID (*c*)	Number of children	4
ZoneID2 (*z*)	L2 zone level.	907
Val	Number of individuals	

Many have considered the challenge of how to translate a numerical representation of individuals into an integer representation [18, 32]. In the Danish National Transport Model this is accomplished by allowing fractions of individuals and fractions of households to be processed. Hence, in the demand model that precedes the population synthesis framework an internal numerical weighting of each individual and household is facilitated. Hence, the enumeration of a households and individuals is allowed to be strictly different from 1. So, if in a cell, there are 11.3 households or individuals, we process the decimal part as a 0.3 fraction of a household and introduce a reweighting in the demand model. In case the demand model cannot process fractions a truncation or rounding process is required where the residual fractions are recirculated in the simulation algorithm.

The micro-simulation scheme is based on the following overall steps;

1) Extend the master table with a variable representing the adult status of the individual. This is based on a deterministic probability P_a of being adult based on the full set of socio-economic variables (e.g., income, age, labour market association and more). This table will be referred to as an "extended master table".
2) Construction of an aggregate household table by summing the previously generated "extended master table" according to $\{z,f,c\}$, the counts then representing the sum for each household class.
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 an index over households within each class. Initialise $i = 1$.
5) For $\{i,k\}$ do the following;
   a. Sample first adult.
   b. If $f <> "single"$ sample a second adult based on the characteristics of the first adult.
   c. If $f = "single"$ go to 5d).
   d. If $c = "kids"$, (hence $c > 0$) enter a loop where all kids are sampled based on the household characteristics and the characteristics of the adults (there may be one or two).
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 most important stage in the simulation scheme is Step 5. It involves several stages as illustrated in Fig. 3 below.

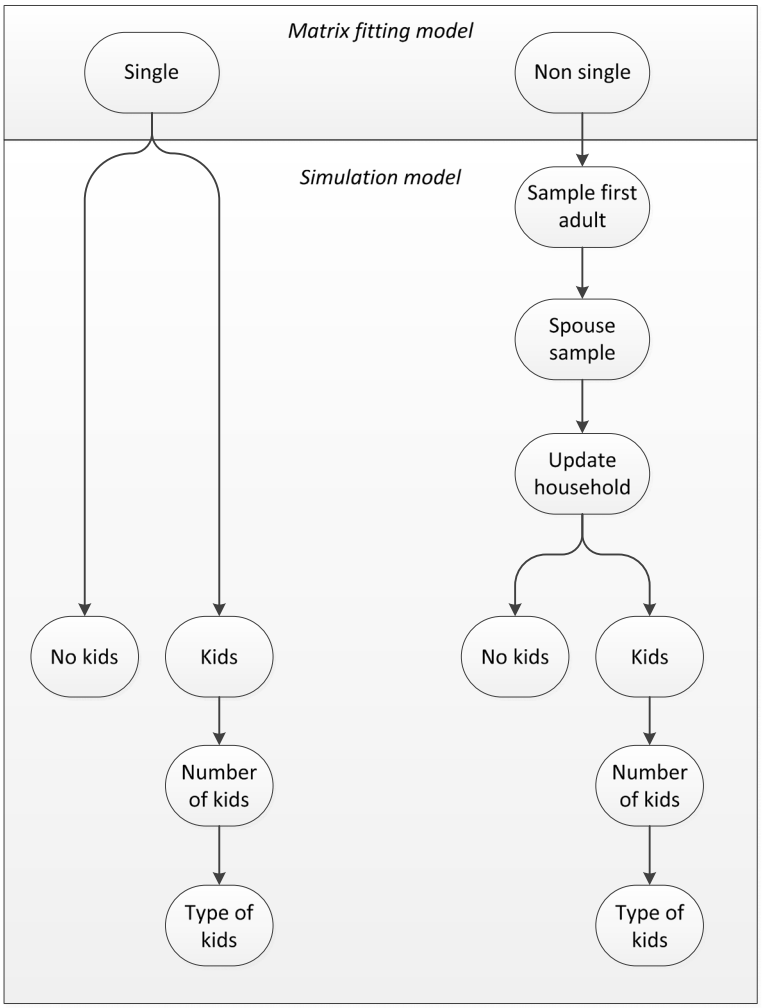

Fig. 3 Illustration of main steps in the micro-simulation stage

From the matrix fitting stage the type of household is known. Hence, for households classified as 'singles' the only decisions to consider in the simulation stage is related to the number of kids and the characteristics of these. For households with two adults we start by selecting a first adult for the household. Subsequently, we select a partner based on a spouse-match model (Step 5b). The probability of selecting a given spouse depends on the income, age, labour market association and gender of both. The conditional "spouse matching" probability is defined on the basis of register data, e.g.

$$P_n\left(a_{s1}, i_{s1}, g_{s1}, l_{s1} | a_{s0}, i_{s0}, g_{s0}, l_{s0}\right) = \frac{P\left(a_{s1}, i_{s1}, g_{s1}, l_{s1}, a_{s0}, i_{s0}, g_{s0}, l_{s0}\right)}{\sum_{a_{s0}, i_{s0}, g_{s0}, l_{s0}} P\left(a_{s1}, i_{s1}, g_{s1}, l_{s1}, a_{s0}, i_{s0}, g_{s0}, l_{s0}\right)} \quad (14)$$

The conditional sampling from this joint distribution takes account of the strong correlation when matching people into households. Although the marginal distribution in (14) excludes the spatial dimension its dimensionality is very large and consists of more than 3 million potential cells. To circumvent the problem of sparsity, less detailed sampling schemes are enforced depending on the "sparsity" of a given sampling (refer to Table 9).

Another important allocation in the simulation stage is the allocation of kids and the type of these. Whereas the spouse-matching model were concerned with the choice of individuals, the models for allocating kids is a household based model. The allocation of kids (Step 5d) involves two sequential steps. In a first stage the number of kids c_h is sampled as a function of household income, age, gender and labour market association of each of the adults in the household. That is;

$$P_h\left(c_h|i_h,a_{s1},g_{s1},l_{s1}a_{s0},g_{s0},l_{s0}\right) = \frac{P\left(c_h|i_h,a_{s1},g_{s1},l_{s1}a_{s0},g_{s0},l_{s0}\right)}{\sum_{i_h,a_{s1},g_{s1},l_{s1}a_{s0},g_{s0},l_{s0}} P\left(c_h|i_h,a_{s1},g_{s1},l_{s1}a_{s0},g_{s0},l_{s0}\right)} \quad (15)$$

After having sampled the number of kids we sample a type classification for each of these. This is based on household income and age and labour market association for the two adults as seen below.

$$P_h\left(c_g,c_a,c_l|i_h,a_{s1},l_{s1}a_{s0},l_{s0}\right) = \frac{P\left(c_g,c_a,c_l|i_h,a_{s1},l_{s1}a_{s0},l_{s0}\right)}{\sum_{i_h,a_{s1},l_{s1}a_{s0},l_{s0}} P\left(c_g,c_a,c_l|i_h,a_{s1},l_{s1}a_{s0},l_{s0}\right)} \quad (16)$$

The sampling of type of kids is somewhat simplified in that we do not condition the probability of selecting one kid with other kids. This would significantly complicate the sampling scheme and require very large tables as input data. For the same reason, the gender classification has been excluded as well.

In the current implementation we sample from marginal tables which have been produced on the basis of register data and constitute one of many input tables. However, it is possible to construct any mathematical model for the spouse matching or the sampling of kids to be able to reflect possible future fluctuations in the way households are formed. These external prediction models could be modelled using time-series analysis, however a detailed discussion of this is beyond the scope of the paper.

Although the simulation stage solves the important allocation problem of grouping individuals into households it introduces a potential inconsistency problem. The final list of individuals, after joining these into households, may not be entirely consistent with the master table when aggregating over the different dimensions. This is because the household simulation stage it is based on random draws, which although consistent at the aggregate level due to the law of large numbers may not be entirely consistent with the targets from Tables 4, 5, 6, 7 and 8. There are different solutions to this problem. The simplest solution is to introduce a re-scaling of individuals. It means that for a household, one person might have a weight of e.g. 1.02 when used in the demand model. In case of the Danish National Transport Model, this is the approach taken and the accumulated trip matrices are calculated as a weighted matrix of trips across all households and individuals at the micro level. In that case it is perfectly fine to have weighted individuals.

However, if the model following the population synthesis requires complete agent-based inputs, which at all times needs to be consistent with targets, this is a substantial complication. Although there is a relative developed literature on joint hierarchical matrix fitting of households and individuals (Muller and Axhausen, [26]) it is less clear how to ensure consistency in the micro-simulation stage. One possible approach is to direct the micro simulation to the pool of individuals listed from the IPF. Then join people into households without replacement and essentially continue until the pool is used up. However, from a computational (combinatorial) and book-keeping point of view this is a relative cumbersome process.

Table 4 Target $T_{ga}(z0, g, a)$ for age, gender and municipality level

Variable	Description	Classes
AgeID (a)	Age	10
GenderID (g)	Gender	2
ZoneID0 (z_0)	L0 zone level (Municipality)	98
Val	Number of individuals	

Table 5 Target $T_{ai}(z0, a, i)$ for age, income and municipality level

Variable	Description	Classes
AgeID (a)	Age	10
IncomeID (i)	Income	11
ZoneID0 (z_0)	L0 zone level (Municipality)	98
Val	Number of individuals	

3 Application

In this section we consider the specific application of synthesising the Danish population and focus on notation, data and provide insight with respect to model validation.

3.1 Data and notation

The synthesis is based on Danish register data. This is essentially a micro database for all Danish citizens with a wide range of attributes related to demography, income, social class, job, family structure and location. Most importantly, however, the data is generally of a very high quality as it is the basis for tax payments and income transfers. The data are reported directly from firms and public authorities to Statistic Denmark.

In the fitting of the master table for individuals (refer to Table 3 below) we consider a matrix $T(a, g, i, l, f, c, z)$

Table 6 Target $T_{al}(z0, a, l)$ for age, LMA and municipality level

Variable	Description	Classes
AgeID (a)	Age	10
LMAID (l)	LMA	8
ZoneID0 (z_0)	L0 zone level (Municipality)	98
Val	Number of individuals	

Table 7 Target $T_{af}(z0, a, f)$ for age, family structure and municipality level

Variable	Description	Classes
AgeID (a)	Age	10
FamID (f)	Family structure	2
ZoneID0 (z_0)	L0 zone level (Municipality)	98
Val	Number of individuals	

that is spanned by seven dimensions. The master table represents the fundamental input and output format for the IPF. When fully spanned it represents approximately 17.4 million matrix entries or 19,200 potential socio-economic groups for each of the 907 zones. A more detailed description of the variable definitions is provided in Table 14 in Appendix 1.

The constraints for the IPF is shown in Tables 4, 5, 6, 7 and 8. Most of these operate at the municipality level because this is the lowest level for which official demographic forecasts exists. In all, there are more than 22,000 constraints and these are cross-linked in the sense that they share common variables. In order to account for consistency issues the ranking harmonisation procedure as described in Section 2.1 is used.

The choice of targets is a balance between the precision of the final matrix and what variables we can actually forecast with a reasonable precision and amount of effort. Clearly, if more dimensions and more variables had been introduced, the final matrix would have been more "heavily" constrained, and given that the constraints proved correct, the final matrix would then be more correct. However, it is not trivial to establish spatial forecasts for e.g. 2020 and 2030 for very detailed variables, and the uncertainty of these forecasts is likely to be high. As a result, we have chosen a rather simple set of constraints in which the fundamental constraint, $T_{ga}(z0, g, a)$ in Table 4 is provided as an official forecast.

Whereas the $T_{ga}(z0, g, a)$ constraint can be based on official forecasts, this is not in general the case for income forecasts. Hence, if we are to predict a change in population per income category as a result of an overall increase in income this is non-trivial as the underlying distribution is asymmetric and right skewed. In order to generate future income targets micro-simulation is used. In a first stage, we generate numerical income measures for each individual by random sampling from the income intervals of the targets. These incomes can then be subjected to possible transformations of which a simple uniform upscaling of the average income level is the simplest scenario. Finally, after the income vector has been modified it is converted back to the categorical representation of the target. In practise, it means that, as people gets richer the frequency table representing the income target is shifted to the right as it should.

Table 8 Target $Q(z)$ for L2 zone level

Variable	Description	Classes
ZoneID2 (z)	L2 zone level	907
Val	Number of individuals	

3.2 Household simulation stage

The resampling stage where individuals are matched in households is based on conditional sampling from tables generated on the basis of register data. The spouse matching table is shown below in Table 9 and further illustrated in a 'heat map' type of plot in Fig. 4.

As commented in Section 2.3 we implement different sampling schemes depending on the "sparseness" in the matching. Figure 4 is intended as an illustration of the correlation between the ages of the two adult persons in the household. Although in reality it is a categorical grid it is illustrated in the form of a smoothened heat-map to illustrate the correlation density. Similar correlation patterns can be plotted for income and for the labour market association.

In addition to the sampling of the household composition, the model uses sampling as a mean to allocate kids to households. This is conditional on the spouse matching and considers both the number of kids and a classification of which type of kid to allocate. Tables 10 and 11 represent the marginal probability tables for the sampling of kids and for illustration of the correlation between the number of kids and the age of the adult female in the household, we offer a contour type of plot in Fig. 5.

3.3 Validation

The validation of population synthesis models is non-trivial. It compares to the validation of model fit for high dimensional probability distributions which is an active research area in statistics. The challenge is that usual norms such as root-mean-square (RMSE) or Wasserstein metrics offers little value when evaluated across many dimensions. At least it provides no information at the level of the cells. Also, because of the many dimensions simple illustrations of the "performance" cannot be carried out. As a result, the paper will not represent a complete validation of the model framework but provide important validation insights by focusing on two aspects: i) Uncertainty that result from the household simulation stage and propagate to the final output, and ii) Prediction performance when evaluated at the level of zones and

Table 9 Marginal "Spouse matching" probabilities for allocating individuals into households

Variable	Type	Description
PIncomeID1	Long Integer	Person gross income categories (11 classes)
AgeID1	Integer	Age categories (10 classes)
GenderID1	Integer	Gender dummy, GenderID = 1 for male
LmaID1	Integer	Labour market association variable (8 classes)
PIncomeID2	Long Integer	Personal gross income categories (11 classes)
AgeID2	Integer	Age categories (10 classes) for spouse
GenderID2	Integer	Gender dummy for spouse, GenderID = 1 for male
LmaID2	Integer	Labour market association variable (8 classes) for spouse
Mprob	Double	Probability for person "ID" having a spouse with characteristics given by the "SID" classification variables
Mprob1	Double	Similar as above, but with a less detailed socio-economic classification for person 1 (PIncomeID1 is eliminated).
Mprob2	Double	Similar as above, but with a less detailed socio-economic classification for person 1 (PIncomeID1 and LMAID1 are eliminated).

when considering a prediction horizon between 2010 and 2015.

The model framework is deterministic in the sense that if we select a specific random seed for the household simulation stage, the same population will be replicated given the inputs are the same. However, it is relevant to ask how sensible the final results are with respect to this seed. In other words, how much sampling noise is generated and propagated from the household simulation stage to the final list of agents. An even more important and relevant question however, is to what extent this noise will affect the final output of the transport demand model. To analyse this specific question the entire model framework has been run 20 times with different seed numbers. Hence, for each of these runs, the composition of the households is different due to different seed numbers. In Fig. 6 the overall percentage deviation is illustrated for trips and mileage for 6 different transport modes.

As can be seen, the variation at the aggregated level is small and in most cases comfortably below 0.03%. Clearly, this is the result of the "law of large numbers" and suggests that we do not have to worry about the sampling noise when results are aggregated over many households. However, if we consider more detailed outputs, noise caused by the sampling will increase. In Tables 12 and 13 the share of origin zones where the percentage deviation is below certain thresholds is illustrated.

The column representing the "weighted" distribution has a better profile in that fewer entries are above the 1% deviation. In particularly very few

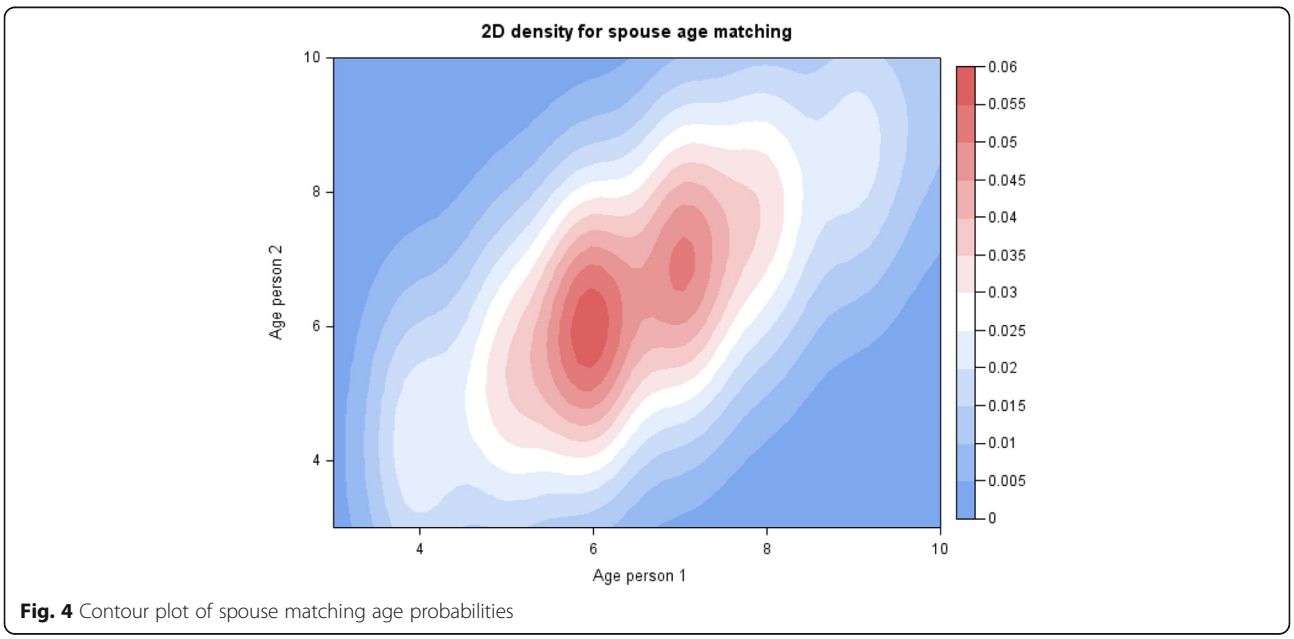

Fig. 4 Contour plot of spouse matching age probabilities

Table 10 Marginal probabilities of having a specific number of kids in certain types of household compositions

Variable	Type	Description
HHIncomeID	Long Integer	Household gross income categories (12 classes)
AgeID	Integer	Age categories (10 classes)
GenderID	Integer	Gender dummy, GenderID = 1 for male
LmaID	Integer	Labour market association variable (8 classes)
AgeSID	Integer	Age categories (10 classes)
GenderSID	Integer	Gender dummy, GenderID = 1 for male
LmaSID	Integer	Labour market association variable (8 classes)
NChildren	Integer	Number of children (1,…,+5)
Cprob	Real	Probability for having the given number of kids (NChildren)

"weighted entries" are above the 5% threshold. Again, this is because of the law of large numbers as smaller zones will have fewer agents which in turn will drive up the relative sampling noise. From an application point of view, what matters is the weighted measurement of trips and mileage and the results suggest that although sampling noise exists, it is at a low level.

It is also relevant to compare the population synthesis with observed register data. More specifically we predict the 2015 population based on a starting matrix form 2010 and by using correct targets for 2015. Hence, per assumption there are no biases in the constraints at the municipality level. This gives an indication of a "best-case" benchmark for the synthesizer. Others have looked at various sensitivity investigations when targets and the starting solutions have been varied [24]. However, this in itself involves a rather comprehensive Monte-Carlo scheme which may not be of any particular interest in general and is certainly outside the scope of the present paper.

Table 11 Marginal probabilities of having a specific type of kids in certain types of household compositions

Variable	Type	Description
HHIncomeID	Integer	Household gross income categories (12 classes)
AgeID	Integer	Age categories (10 classes)
LmaID	Integer	Labour market association variable (8 classes)
AgeSID	Integer	Age categories (10 classes)
LmaSID	Integer	Labour market association variable (8 classes)
GenderCID	Integer	Gender dummy, GenderID = 1 for male
AgeCID	Integer	Age of kid.
LmaCID	Integer	Labour market association variable (8 classes)
CMprob	Real	Conditional probability for a child to be of a given type given the household characteristics (CMprob sums to 1 over GenderCID and LmaCID)

Firstly, in Fig. 7 the weighted percentage deviation at the sub-zones level is presented. This test essentially looks into how well the model recovers the population at more detailed spatial levels (levels that are not supported by the constraints). As can be seen, over the 5-year period there is an average deviation of around 3.5% which equals approximately to 0.7% deviation per year. However, for certain zones it can be substantially higher. The results suggest that there is definitely variation over the years and for some zones this variation may be substantial. This may reflect the development of new urban settlements not reflected in the first-stage fitting. Not surprisingly, the prediction error is significantly smaller per year than has been observed in Krishnamurthy S, Kockelman KM [21] and in McCray et al. [24]. However, these studies considered the total observed error and also included uncertainty in the population targets.

In Fig. 8 we add additional age groups in order to assess the deviation at the subzone level when combined with age groups. As expected, the variation gets bigger and especially for younger generations which are much more mobile compared to older generations. Still, the annual deviation across age groups is in most cases below a 1% deviation per year.

While these results suggest that the synthesis framework give rise to a sizable deviation, two important elements should be considered. First, in Denmark most large cities are experiencing a large inflow of people and to give room for these people new settlement areas are continuously developed. Some of these areas are quite large and will even for a 5 year period represent substantial relative changes to the existing population (examples for Copenhagen are Nordhavn and Sydhavn). These fluctuations are not captured by the synthesis framework as these processes happen inside the municipality. To some extent this uncertainty can be captured in forecasts by aligning the sub-zone population with projected expectations at this level (this is accomplished in the second stage IPF). Such projections typically exist and would substantially improve the performance. The second issue relates to the socio-dynamics of the population over time. As an example, while for

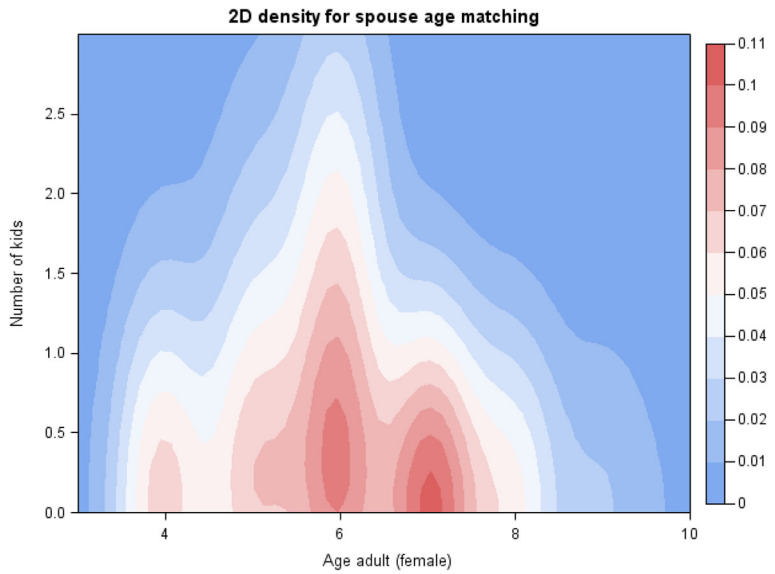

Fig. 5 Contour plot of kids sampling probabilities

Copenhagen there has been a general increase in the population between 2010 and 2015 of approximately 10% the 20–29 years old have increased 17% and kids between the age of 3 and 6 as much as 20%. Hence, predicting this socio-dynamics (which is even more extreme at the level of detailed zones) is quite a challenge.

4 Summary and conclusions

The paper presents the population synthesis methodology applied in the New Danish National Transport Model. The methodology consists of three main stages involving a target harmonisation stage, a matrix fitting stage and a household simulation stage where individuals from the population matrix are listed as individuals

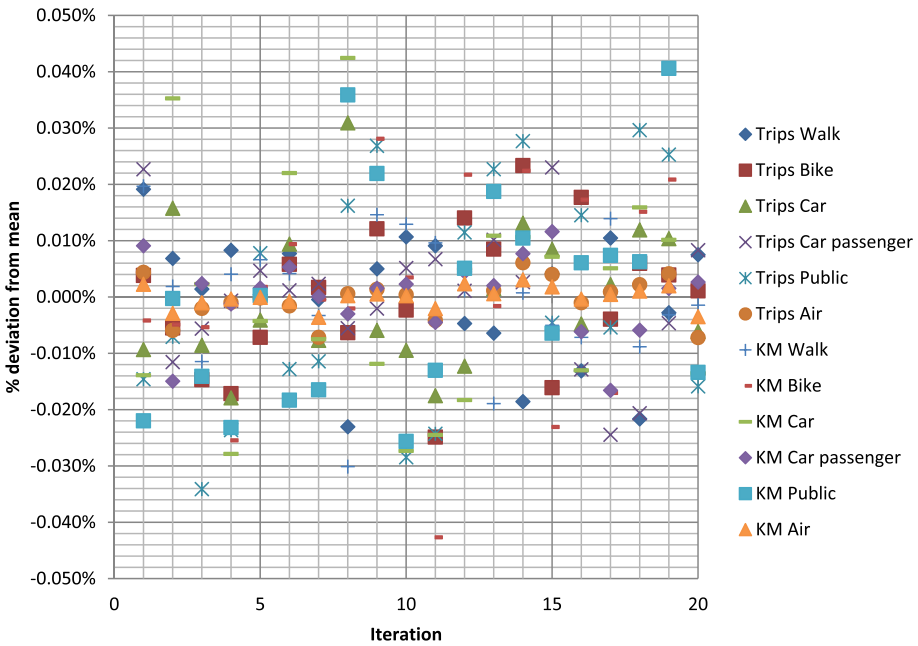

Fig. 6 Accumulated noise in final trips and mileage cause by the household simulation stage

Table 12 The share of origin zones where the percentage deviation is below certain thresholds for car

% diff	Car trips	Car trips (weighted)	Car Mileage	Car Mileage (weighted)
below 1%	97.95%	99.24%	94.08%	95.45%
1–2%	1.53%	0.73%	4.78%	4.24%
2–3%	0.20%	0.03%	0.60%	0.25%
3–4%	0.06%	0.00%	0.18%	0.04%
4–5%	0.03%	0.00%	0.09%	0.02%
5–6%	0.04%	0.00%	0.05%	0.01%
6–7%	0.02%	0.00%	0.06%	0.00%
7–8%	0.02%	0.00%	0.04%	0.00%
8–9%	0.02%	0.00%	0.03%	0.00%
9–10%	0.03%	0.00%	0.01%	0.00%
10–15%	0.04%	0.00%	0.04%	0.00%
15–20%	0.01%	0.00%	0.00%	0.00%
over 20%	0.05%	0.00%	0.04%	0.00%

and grouped into households. Although the presented model applies well known methodologies such as iterative proportional fitting and micro simulation, it excels by providing an end-to-end description of a state-of-the-art model that covers the population synthesis for a whole country. Moreover, the paper considers two aspects of population synthesis which are often not considered in the literature: i) the target harmonisation stage and ii) the household micro-simulation stage.

The harmonisation stage, which hasn't received much attention in the scholarly literature, is related to the problem of ensuring that target inputs are consistent. For complex models with many potential targets it is a non-trivial problem which in specific situations may require a separate matrix fitting stage. The paper provides an example of this and continues to introduce a ranking approach which applies to the target definitions of the current model. Specific attention is also given to the household simulation stage which is concerned with the grouping of individuals into households. The simulation stage consists of different model stages, from the selection of adults within the household, the sampling of spouse and the sampling of kids and the type of kids. All of these stages are is described and examples are provided to illustrate the correlation structure in the underlying data.

The paper provides some insight into the validation of the model framework by; i) evaluating how sampling noise generated in the household simulation stage

Table 13 The share of origin zones where the percentage deviation is below certain thresholds for public transport

% diff	Pub trips	Pub trips (weighted)	Pub Mileage	Pub Mileage (weighted)
below 1%	95.54%	98.81%	95.44%	97.75%
1–2%	2.21%	0.92%	2.55%	1.86%
2–3%	0.56%	0.14%	0.64%	0.19%
3–4%	0.31%	0.06%	0.44%	0.10%
4–5%	0.23%	0.03%	0.24%	0.04%
5–6%	0.17%	0.01%	0.19%	0.02%
6–7%	0.18%	0.01%	0.11%	0.01%
7–8%	0.12%	0.00%	0.04%	0.00%
8–9%	0.12%	0.00%	0.09%	0.01%
9–10%	0.07%	0.00%	0.03%	0.00%
10–15%	0.22%	0.01%	0.11%	0.01%
15–20%	0.13%	0.00%	0.08%	0.01%
over 20%	0.14%	0.00%	0.06%	0.00%

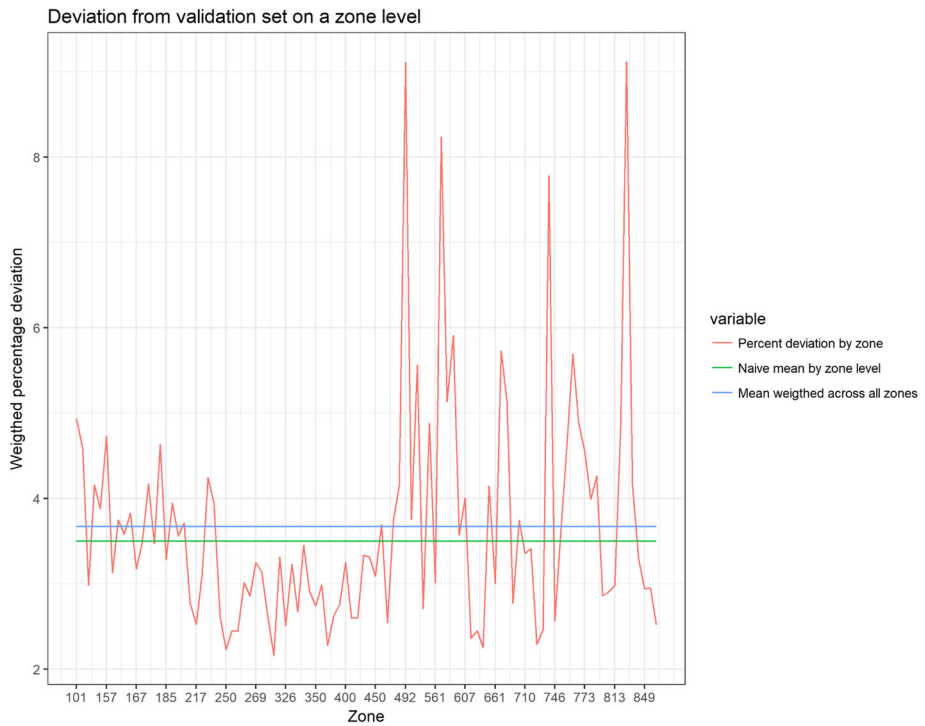

Fig. 7 Weighted percentage subzone deviation of predicted population from 2010 to 2015

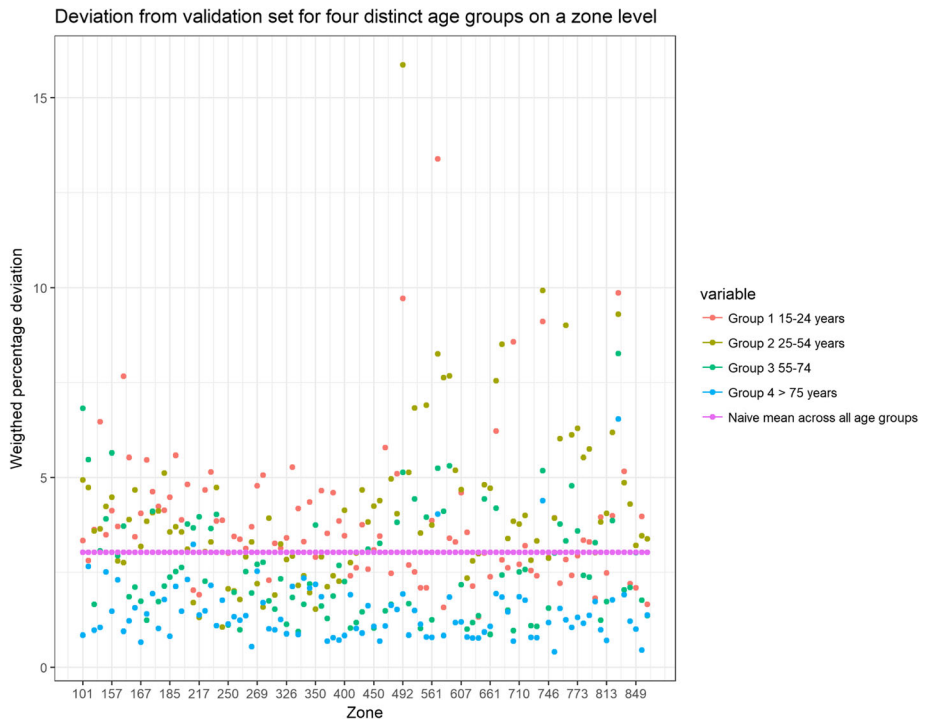

Fig. 8 Weighted percentage subzone deviation of predicted population by age groups from 2010 to 2015

propagate through the entire transport model, and ii) by evaluating prediction performance between 2010 and 2015. The conclusion of these results are summarised below.

- At the aggregated level, the sampling noise can almost be neglected. This is a trivial result of laws of large numbers.
- As results are detailed the sampling noise increases. However, in more than 99% of cases the percentage difference is below 1% even at the level of zones.
- The prediction performance from 2010 to 2015 when evaluated against detailed observed zone targets reveals a deviation of 0.7% per year. Although this is based on correct targets at the level of municipalities it reflect dynamics at the detailed zone level which is particularly apparent for the large cities.

The prediction performance can be substantially improved by adding detailed zone targets to the model. This is not investigated in the paper.

4.1 Future research

Future research seen from an applied perspective may focus on the following directions.

- Methodologies and paradigms for model validation of high-dimensional distributions. Part of this should include back-casting and "forecasting" to years for which data are available.
- The development of parametric models for spouse matching and household composition such that these involve spatial correlation and kids in a joint representation. This requires the use of dynamic models and longitude data to capture trends in single-family households, how this relates to urban migration and the age distribution of children.
- Better models for targets in order to make population synthesis models more robust to the external forecasts.
- Models that can facilitate "reconstructions" of the starting solution in order to circumvent the problem of non-structural zeros.

In general, any improvement to the population synthesis stage will benefit not only the prediction of the population in itself, but all subsequent modelling steps. This is particularly important for transport models as local transport demand is a mirror of the residing population.

5 Appendix 1

5.1 Variable definitions

Table 14 Notation list

Variable	Description	Classes
AgeID (a)	Age	$a = 1, \ldots, 10$
GenderID (g)	Gender	$g = 1, 2$
IncomeID (i)	Personal income	$i = 0, \ldots, 10$
LmaID (l)	Labour market association	$l = 1, \ldots, 8$
FamID (f)	Family structure	$f = 1, 2$
NumChildID (c)	Number of children	$c = 1, \ldots, 4$
ZoneID2 (z)	L2 zone level.	$z = 1, \ldots, 907$
ZoneID0 (z_0)	L0 zone level (municipality)	$z_0 = 1, \ldots, 98$
Val	Number of individuals	
$T_{ga}(z0, g, a)$	Target for age, gender and ZoneID0	$a \times g \times z_0$
$T_{ai}(z0, a, i)$	Target for age, income and ZoneID0	$a \times i \times z_0$
$T_{al}(z0, a, l)$	Target for age, LMA and ZoneID0	$a \times l \times z_0$
$T_{af}(z0, a, f)$	Target for age, family structure and ZoneID0	$a \times f \times z_0$
$\tilde{T}_{ag}(z0, a, i)$	Harmonised target for age, income and ZoneID0	$a \times i \times z_0$
$\tilde{T}_{ag}(z0, a, l)$	Harmonised target for age, LMA and ZoneID0	$a \times l \times z_0$
$\tilde{T}_{ag}(z0, a, f)$	Harmonised target for age, family structure and ZoneID0	$a \times f \times z_0$
$Q(z)$	Target and control vector at the level of ZoneID2	$z = 1, \ldots, 907$
$H(z)$	New temporary target vector at the ZoneID2 level	$z = 1, \ldots, 907$
$\tilde{H}(z)$	New temporary harmonised target vector at the ZoneID2 level	$z = 1, \ldots, 907$
$P_{k+1}(z, i, l, f, k \mid z0, a, g)$	Marginal probability vector for $\{z, i, l, f, k \mid z0, a, g\}$	All
$T_{k+1}(z, z0, g, a, i, l, f, c)$	Solution at iteration $k+1$	All
e	Convergence criteria	1E-6

Table 15 Age classes

AgeID	Description	Count	Probability
1	0–7 years	522,076	9.43
2	8–14 years	479,219	8.66
3	15–17 years	214,187	3.87
4	18–24 years	463,267	8.37
5	25–29 years	310,969	5.62
6	30–54 years	1,919,435	34.68
7	55–64 years	722,636	13.06
8	65–74 years	515,702	9.32
9	75–84 years	277,185	5.01
10	> = 85 years	109,961	1.99

Table 16 Gender classes

GenderID	Description	Count	Probability
1	Male	2,745,318	49.60
2	Female	2,789,420	50.40

Table 17 Children classes

NumChildID	Description	Count	Probability
1	0 children	2,719,295	49.13
2	1 Child	861,105	15.56
3	2 Children	1,279,750	23.12
4	3 Children or more	674,588	12.19

Table 18 Family classes

FamID	Description	Count	Probability
1	Single	998,631	18.04
2	Non-single	4,536,107	81.96

Table 19 Income classes

IncomeID	Description	Count	Probability
0	0 DKK	1,026,498	18.55
1	0–99,999 DKK	746,897	13.49
2	100,000–199,999 DKK	1,396,067	25.22
3	200,000–299,999 DKK	1,111,765	20.09
4	300,000–399,999 DKK	707,445	12.78
5	400,000–499,999 DKK	278,561	5.03
6	500,000–599,999 DKK	115,771	2.09
7	600,000–699,999 DKK	56,468	1.02
8	700,000–799,999 DKK	30,801	0.56
9	800,000–999,999 DKK	29,942	0.54
10	> = 1,000,000 DKK	34,423	0.62

Table 20 LMA classes

LmaID	Description	Count	Probability
1	Full-time employed	1,539,004	27.81
2	Part-time employed (32 h/week)	384,309	6.94
3	Students	1,271,364	22.97
4	Retired	1,002,847	18.12
5	Unemployed	301,015	5.44
6	Other people out of job	558,460	10.09
7	Students with job	272,911	4.93
8	Self-employed	204,828	3.70

6 Appendix 2

6.1 Description of matrix fitting algorithm

6.1.1 First-stage IPF

The first-stage IPF is essentially a traditional IPF iterated over all four constraints to successively update the matrix. The algorithm is presented below.

Step 1: Set $k = 0$ and let $T_k(a, g, i, l, f, c, z)$ be the starting solution.

Step 2: Set $\boldsymbol{k = k + 1}$ and iterate eq. (17)–(24) below.

$$P_{k+1}(i,l,f,c,z|z_0,a,g) = \frac{T_k(a,g,i,l,f,c,z)}{\sum_{i,l,f,c,z\in z_0} T_k(a,g,i,l,f,c,z)} \tag{17}$$

$$T_{k+1}(a,g,i,l,f,c,z) = P_{k+1}(i,l,f,c,z|z_0,a,g)\ \tilde{T}_{ag}(z_0,a,g) \tag{18}$$

$$P_{k+2}(g,l,f,c,z|z_0,a,i) = \frac{T_{k+1}(a,g,i,l,f,c,z)}{\sum_{g,l,f,c,z\in z_0} T_{k+1}(a,g,i,l,f,c,z)} \tag{19}$$

$$T_{k+2}(a,g,i,l,f,c,z) = P_{k+2}(g,l,f,c,z|z_0,a,i)\tilde{T}_{ai}(z_0,a,i) \tag{20}$$

$$P_{k+3}(g,i,f,c,z|z_0,a,l) = \frac{T_{k+2}(a,g,i,l,f,c,z)}{\sum_{g,i,f,c,z\in z_0} T_{k+2}(a,g,i,l,f,c,z)} \tag{21}$$

$$T_{k+3}(a,g,i,l,f,c,z) = P_{k+3}(g,i,f,c,z|z_0,a,l)\ \tilde{T}_{al}(z_0,a,l) \tag{22}$$

$$P_{k+4}(g,i,l,c,z|z_0,a,f) = \frac{T_{k+3}(a,g,i,l,f,c,z)}{\sum_{g,i,l,c,z\in z_0} T_{k+3}(a,g,i,l,f,c,z)} \tag{23}$$

$$T_{k+4}(a,g,i,l,f,c,z) = P_{k+4}(g,i,l,c,z|z_0,a,f)\ \tilde{T}_{af}(z_0,a,f) \tag{24}$$

Step 3: If $\|T_{k+4}(a,g,i,l,f,c,z) - T_{k+3}(a,g,i,l,f,c,z)\| > e\ \forall\ a, g, i, l, f, c, z$ go to Step2. Otherwise stop.

Note that $\|x-y\| = \sqrt{(x-y)^2}$ and $e = 1E-6$. The converged matrix, which will have the same form as in Table 3, is referred to as $T^*(a,g,i,l,f,c,z)$.

6.1.2 Second-stage IPF

The first-stage IPF will render a vector $T^*(a,g,i,l,f,c,z)$, which is on the one hand consistent with the harmonised targets and on the other hand replicates the structure of the starting solution by maintaining the odds ratios (Bishop et al. [8]). However, in specific contexts we may experience that the starting solution is too far away from the true solution. This might be the case if we are looking at long-term forecasts. For instance, if an area is supposed to develop from 0 to 1000 persons

during the forecast period, then the above methodology will be problematic as it will render 0 persons due to the structural zero in the starting solution. Even using heuristic methods to allow for non-empty cells will not solve the core of the problem. A new attempt where machine learning and deep learning is used to reconstruct starting solutions from a compressed distribution reduced the problem of empty cells [5].

In order to account for this problem, additional constraint is allowed, which is processed in a second-stage fitting. The idea is that users can then adjust the projected population for specific detailed zones for which the starting solution may be inappropriate.

The second-stage fitting can be carried out in two ways. One option is to superimpose an additional set of constraints at the most detailed zone level, which is then included as a normal constraint. For zones where no adjustment is required $Q(z) = -1$. For zones z, where $Q(z) \geq 0$, the new values of $Q(z)$ will be used to adjust the solution to these new targets. Hence, if $Q(z) = -1$ for all z, the second-stage IPF can be skipped.

Firstly, calculate (from the IPF fitted master table) a $T^*(z)$ vector, which is simply the estimated population at the zone level z. Hence,

$$T^*(z) = \sum_{g,a,i,l,s,k} T^*(a,g,i,l,f,c,z) \tag{25}$$

Also, define a new temporary target vector $H(z)$ *as*;

$$\textit{if } Q(z) = -1 \textit{ then } H(z) = T^*(z) \tag{26}$$

$$\textit{if } Q(z) \geq 0 \textit{ then } H(z) = Q(z) \tag{27}$$

By introducing the new target vector $H(z)$ it is important to realise that the harmonisation in Section 2.1 will in principle be affected. However, the first option for the second-stage fitting sees the zone correction as a "local correction", which should not overrule the general harmonisation principles. In other words, we will not change the overall population forecast at the municipality level as a result of the changes imposed by $Q(z)$. So if users of the model are adding people to a given zone z without removing people from other zones, the model will do this automatically to maintain the constraints at the municipality level. This means that the $Q(z)$ projection may not be reflected in the final solution.

The second stage of the synthesiser is completed by calculating the harmonised equivalent of $H(z)$ given by $\tilde{H}(z)$

$$\tilde{H}(z) = \frac{H(z)}{\sum_{z \in z_0} H(z)} \sum_{a,g} T_{ga}(z_0,g,a) \tag{28}$$

Finally, we process the IPF by including one additional constraint as given below.

$$P_{k+5}(z|g,a,i,l,f,c) = \frac{T_{k+4}(a,g,i,l,f,c,z)}{\sum_{a,g,i,l,f,c} T_{k+4}(a,g,i,l,f,c,z)} \tag{29}$$

$$T_{k+5}(a,g,i,l,f,c,z) = P_{k+5}(z|a,g,i,l,f,c)\, \tilde{H}(z) \tag{30}$$

That is, the second-stage IPF is carried out by iterating (17)–(24) and (29)–(30) until convergence. It is efficient to use the starting solution from the first stage in the second stage.

Another option which is also supported by the model is to superimpose the $Q(z)$ target such that it is not adjusted in the harmonisation stage. In this case we simply impose the level provided by the target to the conditional distribution for the population for the specific cell. Hence

$$T^{**}(a,g,i,l,f,c,z) = \frac{T^*(a,g,i,l,f,c,z)}{\sum_{a,g,i,l,f,c} T^*(a,g,i,l,f,c,z)}\, Q(z) \tag{31}$$

In this case the overall target at the level of the municipality may be violated. However the additional target at the zone level will be matched exactly. Sometimes this is preferable from the point of view of public authorities if they have additional "local information" that is not embedded in the overall constraints.

Acknowledgements
Not applicable.

Funding
The work has been carried out as part of a large transport model project for the transport ministry.

Availability of data and materials
All data in the paper are register data and can be accessed from www.dst.dk. For most of the data these are publicly available through the data bank: http://www.statistikbanken.dk/
However, specifically for building the starting solution more detailed data is required. These data are protected and cannot be distributed freely. However, the data can be shared upon request to the author.

Authors' contributions
Not applicable as the author is the sole author. The author read and approved the final manuscript.

Competing interests
The author declares that he has no competing interests.

Publisher's Note

Received: 29 June 2018 Accepted: 21 November 2018
Published online: 29 December 2018

References

1. Abraham JE, Stefan KJ, Hunt JD (2012) Population synthesis using 913 combinatorial optimization at multiple levels, transportation research 914 record (2012)

2. Arentze T, Timmermans H, Hofman F (2007) Creating synthetic household populations - problems and approach. Transp Res Rec 2014:85–91
3. Bar-Gera H, Konduri K, Sana B, Ye X, Pendyala RM (2009) Estimating survey weights with multiple constraints using entropy optimization methods. In: Paper presented at the 88th Annual Meeting of the Transportation Research Board, Washington, D.C., January
4. Borysov S, Rich J, Pereira FC (2018) Population synthesis meets deep generative modelling. In: Paper presented to the European Hearts Conference, Athens Url: http://arxiv.org/abs/1808.06910
5. Beckman JR, Baggerly KA, McKay MD (1996) Creating synthetic baseline populations. Transp Res A 30(6):415–429
6. Bento AM, Cropper ML, Mobarak AM, Vinha K (2005) The effects of urban spatial structure on travel demand in the United States. Rev Econ Stat, 87(3):466–78
7. Birkin M, Clarke M (1988) Synthesis—a synthetic spatial information system for urban and regional analysis: methods and examples. Environ Plan A 20(12):1645–1671. https://doi.org/10.1068/a201645
8. Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis – theory and Practise. MIT Press, Cambridge
9. Curtis C., Perkins T. (2006). Travel behaviour: a review of recent literature. Department of Urban and Regional Planning, Curtin University, Working paper no.3. Url: http://urbanet.curtin.edu.au/local/pdf/ARC_TOD_Working_Paper_3.pdf
10. Daly A (1998) Prototypical sample enumeration as a basis for forecasting with disaggregate models. Transp Plann Methods 1(D):225–236
11. Darroch JN, Ratcliff D (1972) Generalized iterative scaling for log-linear models. Ann Math Statist 43(5):1470–1480. https://doi.org/10.1214/aoms/1177692379
12. Dykstra RL (1985) An iterative procedure for obtaining I-projections onto the intersection of convex sets. Ann Probab 13(3):975–984. https://doi.org/10.1214/aop/1176992918
13. Deming WE, Stephan FF (1940) On the least squares adjustment of a sampled frequency table when the expected marginal totals are known. Ann Math Stat 11(4):427–444
14. DeSalvo JS, Huq M (1996) Income, residential location, and mode choice. J Urban Econ 40(1):84–99
15. Donovan S, Munro I (2013) Impact of urban form on transport and economic outcomes. Transport Agency of New Zeeland, Research report 513, p 74
16. Farooq B, Bierlaire M, Hurtubia R, Flötteröd G (2013) Simulation-based population synthesis. Transp Res B Methodol 58:243–263
17. Golan A, Judge G, Miller D (1996) Maximum entropy econometrics: robust estimation with limited data. Wiley, New York
18. Gou J, Bhat S (2007) Population synthesis for microsimulating travel behavior. Transp Res Rec 2014: 92–101. https://doi.org/10.3141/2014-12
19. Harland K, Heppenstall A, Smith D, Birkin M (2012) Creating realistic synthetic populations at varying spatial scales: a comparative critique of population synthesis techniques. J Artif Soc Soc Simul 15(1):1
20. Kao S-C, Kim HK, Liu C, Cui X, Bhaduri BL (2012) Dependence-preserving approach to synthesizing household characteristics. Transp Res Rec 2302:192–200
21. Krishnamurthy S, Kockelman KM (2003) Propagation of uncertainty in transportation land use models: investigation of DRAM-EMPAL and UTPP predictions in Austin, Texas. Transp Res Rec 1831:219–229
22. Lee DH, Fei Y (2011) A cross entropy optimization model for population synthesis used in activity-based micro-simulation models. Transportation Research Board (2011), Washington DC Url: https://doi.org/10.3141/2255-03
23. McDougall, M. (1999) Entropy Theory and RAS are Friends. Working Paper, Center for Global Trade Analysis, Department of Agricultural Economics, Purdue University
24. McCray DR, Miller JS, Hoel LA (2012) Accuracy of zonal socioeconomic forecasts for travel demand modeling: retrospective case study. Transp Res Rec 2302(2012):148–156
25. Müller K, Axhausen KW (2010) Population synthesis for micro simulation: state of the art. In: Paper Presented at STRC conference. http://www.strc.ch/2010/Mueller.pdf
26. Müller, K., Axhausen, K.W (2011). "Hierarchical IPF: Generating a synthetic population for Switzerland," ERSA conference papers ersa11p305, European Regional Science As.
27. Pritchard, D.R. & Miller, E.J. Transportation (2012) Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. 39:685. https://doi.org/10.1007/s11116-011-9367-4
28. Rich J, Mulalic I (2012) Generating synthetic baseline populations from register data. Transp Res A 46:467–479
29. Rich J, Hansen CO (2016) The Danish National Passenger Model – model specification and results. Eur J Transp Infrastruct Res 16(4):573–599
30. Ryan J, Maoh H, Kanaroglou P (2007) Population synthesis: comparing the major techniques using a small, complete population of firms. Geogr Anal https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1538-4632.2009.00750.x
31. Simpson L, Tranmer M (2005) Combining sample and census data in small area estimates: iterative proportional fitting with standard software. Prof Geogr 57:222–234. https://doi.org/10.1111/j.0033-0124.2005.00474.x
32. Smith A, Lovelace R, Birkin B (2017) Population synthesis with Quasirandom integer sampling. J Artif Soc Soc Simul 20(4):1–14
33. Stead D, Williams J, Titheridge H (2000) Land use change and the people -identifying the connections. In: Williams K, Burton E, Jenks M (eds) Achieving sustainable urban form, pp 174–186
34. Tanton R (2014) A review of spatial microsimulation methods. Int J Microsimul 7(1):4–25
35. Voas D, Williamson P (2000) An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. Int J Popul Geogr 6:349–366. https://doi.org/10.1002/1099-1220(200009/10)6:5<349::AID-IJPG196>3.0.CO;2-5
36. Zhao Y, Kockelman KM (2001). The Propagation of uncertainty through travel demand models: An exploratory analysis. In: Presented at the 80th Annual Meeting of the Transportation Research Board, January 2001 Url: http://www.caee.utexas.edu/prof/kockelman/public_html/ARS01ErrorPropagation.pdf