CEPS
INSTEAD

WORKING PAPERS

Synthetic populations: review of the different approaches

Johan BARTHELEMY, Eric CORNELIS[1]

FUNDP - University of Namur, Belgium

Working Paper No 2012-18
April 2012

Synthetic Population: Review of the existing approaches*

BARTHELEMY Johan and CORNELIS Eric
Groupe de Recherche sur les Transports
FUNDP - University of Namur

March 27, 2012

Abstract

Microsimulations may involve a large number of agents. It is then practically impossible or too expensive to obtain a fully and complete disaggregated data set about these agents of interest. Moreover, if such a dataset was available, its use would be potentially problematic in view of stringent privacy laws. To address this problem one may build an artificial population starting from known aggregate data.
Most of the known generation methods are explained in this paper. Their advantages and limitations are discussed and references are given for further details.

*This research is part of the MOEBIUS project supported by the Luxembourg 'Fonds National de la Recherche' (contract FNR, project C09/SR/07) and by core funding for CEPS/INSTEAD from the Ministry of Higher Education and Research of Luxembourg.

1 Introduction

Micro-simulations, such as activity-based travel demand models in transport simulation, usually involve a large number of agents. As a result, it may be impossible or too expensive to get a fully disaggregated data set about the agents of interest. Moreover, if such a data set was available, its use may also be problematic due to stringent privacy laws applied in some countries.

To address this complex problem, an artificial population can be build starting from known aggregate data about the true one. As it is obvious that the representativeness of the synthetic population is critical for the simulations accuracy, a synthetic population generator should produce a population approximating as accurately as possible the correlation structure of the true population. In other words, the aim is the generation of a population which is statistically close to the true one.

Recently, synthetic population generation has reveived more and more attention in the litterature. To date the techniques that have emerged belong to one of the following categories: either to the Synthetic Reconstruction techniques (SR) or to Combinatorial Optimization (CO) methods. Most of the known generation methods are explained in this paper which is organized as follows. In Section 2, we first present the conventional approach, from which the others Synthetic Reconstruction techniques are derived, used for building a synthetic population. Section 3 then describes the Combinatorial Optimization methodology. An alternative method developed by the GRT is presented in Section 4. Concluding remarks are finally discussed in Section 5.

2 Conventional approach

2.1 Methodology

To date, the conventional approach for building synthetic population is based on the method developed by Beckman et al. (1996) and is a member of the Synthetic Reconstruction techniques. The main idea behind a population synthesizer consists of merging aggregate data from one source covering the complete population with disaggregated data from a sample in order to get a complete and detailed disaggregated data set for the population of interest. Typically the aggregate data set is an aggregate outcome from an existing census and the disaggregated data set is drawn from a survey over a sample of the population. The aggregate data consist of a set of marginal distributions for some of the relevant characteristics of the true population. We refer to these distributions and variables as target and control variables. Furthermore, the disaggregate data provides full information about the attributes of interest, but only for a sample of agents and is referred to as the seed.

The population synthesis procedure usually starts with identifying the relevant (categorical) socio-demographic variables of the agents. Assuming that the seed presents n attributes of interest and denoting by $V = \{v_1, v_2, ..., v_n\}$ the vector of variables representing these attributes, each combination of values of v_i's therefore defines a socio-demographic group. The synthetic population is then generated by a two steps procedure:

1. Starting from the seed, estimate the k-way joint-distribution of the true population, where $k(\leq n)$ is the number of control variables, such that the resulting distribution

is consistent with the marginal distributions (margins) of the target and preserves the correlation structure of the seed.

2. Select agents from the sample and copy them in the synthetic population in a proportion derived from the distribution computed in the previous step.

The most popular way for estimating a $k-$way joint distribution table based on some known marginal distributions and on a sample is the well known iterative proportional fitting procedure (IPFP) originally described by Deming and Stephan (1940). The procedure implies an initial representative sample of the true population being available. This requirement is important since Mosteller (1968) pointed out that the procedure preserves the interaction structure of the sample as defined by the odd ratios. According to Ireland and Kullback (1968), the IPFP also produces the estimated contingency table that minimizes the discrimination information (also called relative entropy), *i.e.* it yields the constrained maximum entropy estimator. Moreover Little and Wu (1991) demonstrated that IPFP results in a maximum likelihood estimator of the true contingency table. The IPFP uses an initial contingency table of the control variables built from the seed as a starting point. The procedure then iteratively updates the cells depending on the marginal distributions of the target until the margins of the table match the target's ones.

Once the expected numbers of agents in all the socio-demographic groups are estimated, each sampled agent is associated with a probability of being selected in the synthetic population. This probability typically depends on the agent's sampling weight and the expected number of similar agents in the true population. Based on these probabilities, agents are randomly drawn from the seed using a Monte Carlo procedure until the expected number of agents is reached for each socio-demographic group. When a sampled agent is drawn, then all its attributes, including the uncontrolled ones, are pasted in a new synthetic agent who is added to the synthetic population.

2.2 Limitations

As expected, the IPFP results largely rely on the quality of the data. In particular, it is important to notice that the method requires consistency of the margins across the targets and representativeness of the initial sample of the true population. For example if a class of agents is not represented in the seed then this particular class will remain unpopulated in the final synthetic population. These two requirements limit the applicability of the IPFP in real situations.

In addition, recent mobility surveys such as EGT (Direction Régionale de l'Équipement d'Île-de-France, 2005), MOBEL (Hubert and Toint, 2001) or NTS (Office of UK National Statistics, 2010) suggest that the travel behaviour of an individual is significantly influenced by the type and composition of his/her household. This illustrates another limitation of the conventional approach: it is very unlikely for analysts to have access to a single dataset detailing the joint-distribution of individuals' and households' attributes simultaneously. Since the estimation step of the algorithm is designed to deal with a single contingency table, the conventional approach can consequently account either for individual-level or for household-level control variables but not for both. In other words this process results in a synthetic population where either the households or individ-

uals joint-distributions match the desired ones but not both. Note that historically, households' distributions accuracy has been preferred (Ye et al, 2009).

2.3 Improvements

The strong limitations of this first approach conducted several authors to propose interesting improvements to this basic algorithm. Guo and Bhat (2007) proposed an algorithm to overcome the last issue reported by controlling simultaneously the individual- and household-level variables. Their algorithm generates a population where the household-level distributions are closed to those estimated using the IPFP, while simultaneously improving the fit of person-level distributions. Arentze et al. (2007) proposed another method using relation matrices to convert distributions of individuals to distributions of households such that marginal distributions can be controlled at the person level as well. Ye et al. (2009) further built on previous work and proposed a practical heuristic approach called Iterative Proportional Updating (IPU), based on adjusting households' weights such that both household- and individual-level distributions can be matched as closely as possible.

3 Combinatorial optimization

The Combinatorial Optimization based methodologies is another family of synthetic population synthetizer, which is far less covered in the litterature than the ones derived from the conventional approach. These techniques have been used by the NATSEM[1] to build some synthetic populations (*e.g.* Harding *et al.*, 2004; Melhuish *et al.*, 2002 and Williams, 2003). Voas and Williamson (2000) and Huang and Williamson (2002) also studied the CO algorithm for population synthesis.

3.1 Methodology

In the CO method the area for which the population is generated is divided in mutually exclusive and exhaustive p zones. Assume that the synthetic agents are characterised by n attributes of interest. Two forms of inputs are then required by the CO approach:

- a sample from the whole population at the desired level of aggregation describing all the n desired attributes variables;
- and (cross-)tabulations for a subset of the desired variables, representing the distribution of those variables over the p zones.

For example, if a synthetic population is to be created with agents having the characteristics of gender, age class and driving licence ownership, the values for each of these variables must be available in the sample, and tabulations for at least one variable is also required.

The method creates the synthetic population zone by zone, by fitting a sub-set of the sample to the tabulations for each zone:

[1]National Centre for Social and Economic Modeling, University of Canberra, Canada

1. Agents are randomly selected from the sample such that the population size of the current zone is matched. A statistic measure is computed to measure the fit of the generated population to the desired and known distributions of agents' characteristics in the zone.

2. One of the generated agents is then switched randomly (with replacement) with an other one from the sample and the statisctic is computed again. If the fit of the new population is better than the original one, then the switch is maintained, otherwise, the original subset is preserved. This process is repeated until the goodness of fit statistic reachs a threshold value, or a defined number of iteration is reached.

As one can easily notice, the CO method still requires a initial sample of the population, but not necessary at the most dissagregated level, namely the zones. Moreover no consistency assumption are made for the known tabulation at the zones levels. As a result the data requirements for this method are less restrictive than the ones needed for the conventional approach.

3.2 Comparison with the conventional approach

Huang and Williamson (2002) and Ryan *et al.* (2009) compared the CO and the SR techniques by testing their ability to produce accurate synthetic populations, *i.e* statistically similar to the true one. These papers indicate that even if the two approaches are able to build reliable synthetic populations, the CO method tends to show less variations amongst the populations generated. In other words, if one generates m populations using the same set of input data, then the populations generated using a CO method will be much more alike from one run to another, than the one using a SR method. However, as the initial sample is not available at the zone level, the correlation structure of the true population in the zones may not be preserved by this method.

4 GRT approach

As already stated, results of the approaches belonging to the SR family largely rely on the quality and the consistency of the data. The use of different date sources can thus be problematic as inconstencies between them can occur. Moreover, the CO and SR methods both requires an initial sample of the population at a very high dissagregate level, which can be unavailable for synthetizing a large population, *e.g.* a population of a whole country at the municipality level (NUTS-5) consisting of individuals gathered in households. In order to obviate these limitations, the GRT[2] developped an alternative method whose general philisophy is to construct individuals and households by drawing their characteristics or members at random within the relevant distribution at the most disaggregate level available while maintening known correlations as well as possible. The next subsection outline the main step of the procedure, but a complete and formal description can be found in Barthélemy and Toint (2010).

[2]Groupe de Recherche sur les Transports, University of Namur, Belgium

4.1 Methodology

The algorithm consists of a 3-steps procedure for each zone/municipality.

1. a pool of available individuals is generated for the current zone, namely the individuals' attribute joint-distribution denoted by Ind;

2. the households' joint-distribution is estimated and stored in the contingency table Hh;

3. the synthetic households are constructed by randomly drawing individuals from Ind. This is achieved while preserving the distribution computed in the second step. Once a household has been built, it is added in the synthetic population.

Estimating the individuals' distribution

The first step aims at building the Ind pool of available synthetic individuals. This pool is built individual by individual and the contingency table updated accordingly. If disaggregated data is unavailable at the municipality level for some attributes (while margins are), a more aggregate level is used to obtain (approximate) information on the missing attributes.

Since draws from the district-level joint-distributions were used to assign some characteristics, the margins of Ind for these particular variables can be inconsistent with the known true one. A correction is then made to Ind to make it consistent with the margins at municipality level. This correction is computed by suitably shifting some of the attributes' values of certain individuals. Only shifts between two contiguous modalities are allowed, *e.g.* if an individual's age class is 5, then the shift allowed are either 4 or 6.

Estimating the households' repartition

Now that a pool of individuals has been built, the next step is to find an estimator of the households' type contingency table denoted by Hh given some known data provided by different sources. Each cell of Hh corresponds thus to a number of a particular household type.

The estimation of Hh's cells given known data is obtained as the rounded solution of an optimization problem, where the entropy is maximized under the (linear) constraints implied by the known margins on household types. This approach has the advantage of producing a more reasonably spread-out distribution amongst all household's types with respect to the constraints than the one produced by a least-squares formulation.

This solution is finally used as a starting point of a combinatorial optimization problem using a tabu-search algorithm (see Cvijovic et al. 1995, Glover 1989, Glover 1990 and Glover et al. 1997), in order to get a final estimation of Hh. More details on this process are provided in Barthélemy and Toint (2010), but it is enough to note here that the household-type distribution is computed, for each municipality, as the approximate solution of a maximum entropy problem in discrete variables under constraints given by statistics known for the municipality.

Household's generation

Ind and *Hh* being estimated, the last step consists of gathering individuals into households by randomly drawing households' constituent members. Households' types are considered sequentially and household's members generated as follows: a household head is first drawn, without replacement, from the pool of individuals, and then, depending on the household's type, a mate, children and additional adults are also drawn from the pool if relevant. The members' attributes are either directly derived from the household type (*e.g.* the head's and mate's gender) if possible or randomly drawn accordingly to known distributions (*e.g.* the mate's age class can be drawn from the *head's age class* $\times$ *mate's age class*).

4.2 Comparison with the conventional approach

This method presents interesting properties. First of all this synthetic population generator obviates the need for a significant sample of households and individuals at the desired disaggregate level. Moreover, as in conventional-based approaches, the procedure attempts to maximize the entropy of the unknown contingency tables. It also has the advantages of allowing the merging of several data sources and of handling reasonable inconsistencies between them. The preliminary results conducted in Barthélemy and Toint (2010) indicates that the new methodology has potential for generating large synthetic populations.

5 Conclusion

As microsimulations become more and more used, the development of synthetic population generation methods become a growing field of interest as it is an important step of these models. Several algorithm are available in the litterature, and the choice of one of them depends on the final application, the available data and the size of the population to synthetize. For example, if one has to produce a small synthetic population consisting of individuals, with a significant sample available, then a conventional approach would be a reasonnable choice. At the opposite, if the goal is a large population of individuals gathered in households, with no sample available, the GRT method should be preferred.

References

[1] T. Arentze, H.J.P. Timmermans, and F. Hofman. Creating synthetic household populations: Problems and approach. *Transportation Research Record: Journal of the Transportation Research Board*, 2014:85–91, 2007.

[2] J. Barthélemy and Ph. Toint. Synthetic population: a generator obviating the conventional approach limitation. *to be published*, 2010.

[3] R.J. Beckman, K.A. Baggerly, and M.D. McKay. Creating synthetic baseline populations. *Transportation Research Part A*, 30(6):414–429, 1996.

[4] Direction Régionale de l'Équipement d'Île-de France. Les cahiers de l'enquête globale de transport. http://www.ile-de-france.equipement.gouv.fr, 2005.

[5] W.E. Deming and F.F. Stephan. An a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, 11:428–444, 1940.

[6] Department for Transport, editor. *Transport Statistics Great Britain 2009 Edition*. Office of UK National Statistics, London, United Kingdom, 2009.

[7] J.Y. Guo and C.R. Bhat. Population synthesis for the microsimulating travel behavior. *Transportation Research Record: Journal of the Transportation Research Board*, 2014:92–101, 2007.

[8] A. Harding, R. Lloyd, A. Bill, and A. King. Assessing poverty and inequality at a detailed regional level - new advances in microsimulation. Research Papaer No. 2004/26, originally prepared for the UNU-WIDER Conference on Inequality, Poverty and Human Well-Being, 30-31 May 2003, Helsinky, 2004.

[9] Z. Huang and P. Williamson. A comparison of synthetic reconstruction and combinatorial optimization approaches to the creation of small-area microdata. Working Paper. Department of Geography, University of Liverpool, 2002.

[10] J.-P. Hubert and Ph. Toint. *La Mobilité quotidienne des Belges*. Presses Universitaires de Namur, Namur, Belgium, 2002.

[11] C.T. Ireland and S. Kullback. Contingency tables with given marginals. *Biometrika*, 55(1):179–199, 1968.

[12] R.J.A. Little and M.-M. Wu. Models for contingency tables with known margins when target and sampled population differ. *Journal of the American Statistical Association*, 86(413):87–95, 1991.

[13] T. Melhuish, M. Blake, and S. Day. An evaluation of synthetic populations for census collection districts created using spatial microsimulation techniques. Gold Coast, Queensland, Australia, 2002. 26th Australia & New Zealand Regional Science Association International Annual Conference.

[14] F. Mosteller. Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63:1–28, 1968.

[15] J. Ryan, Maoh H., and Kanaroglou P. Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis*, 41:181–203, 2009.

[16] D. Voas and P. Williamson. An evaluation of the combinatorial optimization approach to the creation of synthetic microdata. *International Journal of Population Geography*, 6(6):349–366, 2000.

[17] P. Williams. Using microsimulation to create synthetic small-area estimates from australia's 2001 census. NATSEM working paper, 2003.

[18] X. Ye, K. Konduri, R.M. Pendyala, B. Sana, and P. Waddel. A methodology to match distributions of both household and person attributes in the generation of synthetic populations. Washington, U.S.A., 2009. Transportation Research Board - 88th Annual Meeting.

CEPS
INSTEAD

3, avenue de la Fonte
L-4364 Esch-sur-Alzette
Tél.: +352 58.58.55-801
www.ceps.lu