Chapter 3
Synthetic Population Techniques in Activity-Based Research

Sungjin Cho
Hasselt University, Belgium

Tom Bellemans
Hasselt University, Belgium

Lieve Creemers
Hasselt University, Belgium

Luk Knapen
Hasselt University, Belgium

Davy Janssens
Hasselt University, Belgium

Geert Wets
Hasselt University, Belgium

ABSTRACT

Activity-based approach, which aims to estimate an individual induced traffic demand derived from activities, has been applied for traffic demand forecast research. The activity-based approach normally uses two types of input data: daily activity-trip schedule and population data, as well as environment information. In general, it seems hard to use those data because of privacy protection and expense. Therefore, it is indispensable to find an alternative source to population data. A synthetic population technique provides a solution to this problem. Previous research has already developed a few techniques for generating a synthetic population (e.g. IPF [Iterative Proportional Fitting] and CO [Combinatorial Optimization]), and the synthetic population techniques have been applied for the activity-based research in transportation. However, using those techniques is not easy for non-expert researchers not only due to the fact that there are no explicit terminologies and concrete solutions to existing issues, but also every synthetic population technique uses different types of data. In this sense, this chapter provides a potential reader with a guideline for using the synthetic population techniques by introducing terminologies, related research, and giving an account for the working process to create a synthetic population for Flanders in Belgium, problematic issues, and solutions.

DOI: 10.4018/978-1-4666-4920-0.ch003

INTRODUCTION

Since its introduction in transportation, ABM (activity-based model), which purpose is to estimate an individual induced traffic demand derived from activities, have been applied for traffic demand forecasts. The ABM typically uses different types of input data including daily activity-trip survey data and population data. The individual daily activity-trip schedule data describes the different trips, its purpose, locations, transport modes, as well as its temporal dimension. The population data, including socio-demographic features, are used to estimate population characteristics such as gender, household composition, income, home location, etc. In general, it seems to be hard to use those datasets because they are rather expensive and normally protected by a privacy law. Thus, it is indispensable to find a solution to substitute population data in a synthetic manner.

Several synthetic population generators have been used in the literature to generate synthetic population data in transportation. Examples are techniques like Iterative Proportional Fitting and Combinatorial Optimization. Despite these advancements in research, using those techniques is not easy for non-expert researchers not only due to the fact that there are no explicit terminologies and concrete solutions to some existing issues and problems so far, but also every synthetic population technique handles different types or structures of input and output data.

In this sense, the chapter is supporting a potential reader with a guideline for using synthetic population techniques by introducing terminologies and related research, and giving an account of the working process to create a synthetic population, along with problematic issues and solutions. In detail, the following sections provide common terminologies and related research in this field. Then, section 3 introduces related research. The next section describes the whole process of generating a synthetic population, which consists of three steps: data preprocessing, fitting and drawing (sampling). The section of issues and proposed solutions deals with some issues and solutions addressed by previous research. Finally, the chapter ends with a summary and by suggesting future work in this field.

RELATED RESEARCH

Synthetic population techniques can be largely divided into two groups: IPF and CO. Most techniques in these two groups have a similar concept of fitting seed data to a target marginal distribution, but they generate the required synthetic population in totally different way. This section covers the different ways by introducing related research in each group.

IPF

Deming and Stephan (1940) developed a basic algorithm in *IPF* (Iterative Proportional Fitting), which has been widely applied for synthetic population research in several fields, including transportation. The basic algorithm, which is called ‘a least squares adjustment’, is based on the assumption that the source and target have the same correlation structure. The correlation structure is defined by odds ratios, for example the odds ratio in a 2 x 2 cross-table is calculated as follows:

$$\varnothing = \frac{p_{1,1}p_{2,2}}{p_{1,2}p_{2,1}}$$

where $p_{i,j}$ is a cell proportion of the cell (i, j). Based on that assumption, the IPF adjusts seed data to target marginal distribution to keep the correlation between source and target. We do not explain further details of the IPF algorithm in this chapter, but we are dealing with how it can be applied within the synthetic population process in the next section.

In general, Beckman *et al*. (1996) are cited as the first scholar who generated synthetic populations using the IPF. They applied the IPF to predict synthetic populations of households and persons in a census tract using 1990 census data (summary tables and PUMS). The summary tables provide target marginals, and the PUMS (public use microdata sample) are a representative 5% sample of households and persons, used as seed data. There are two steps in building synthetic populations: a fitting step and a drawing step (also referred to this chapter). At first, a cross-table in the source is made of seed data. A cross-table in the target area is also formed, which is not a complete table because the cell values are not known for the target area. The IPF is used to complete the target cross-table by iteratively revising the cell values in the source area based on the target marginal distribution. Once the cross-table in the target area is completed, the synthetic population of households is generated by sampling the desired number of households from the seed data in the source by means of a household selection probability. The household selection probability is calculated using the complete cross-table in the target area. After Beckman *et al.* (1996), several researches have been conducted for the synthetic population using the IPF algorithm.

Guo and Bhat (2007) generated synthetic population for the Dallas/Fort-Worth area in Texas based on the conventional IPF, proposed by Beckman *et al.* (1996). They also advanced the conventional IPF by dealing with two issues: zero-cell value and multi-level fitting. We will discuss these issues in the following sections.

Arentze *et al.*, (2007) proposed a relation matrix as a new solution to multi-level fitting in synthetic population research. The relation matrix is constructed by converting *individual* marginal distributions to *household* distributions by assigning individuals to household positions, e.g. '2-adult households', '1-male households', 'males living in' and so on. In addition, they introduced data segmentation algorithms (CART and CHAID) for analyzing spatial heterogeneity in population.

Ye *et al.*, (2009) developed synthetic population software (named 'PopGen') with a new algorithm, IPU. The IPU (Iterative Proportional Updating) algorithm is for matching both household and person marginals by updating sample household weights. In the fitting step, household and person type constraints are estimated using the IPF procedure, followed by the calculation of sample household weights by the IPU algorithm. Then, the synthetic population is predicted by a household sampling process that expands sample households according to household selection probabilities. The household selection probabilities are computed by sample household weights calculated in the fitting step. This drawing step is repeated until a best-fit synthetic population is obtained. Note that the next section will cover details of these processes using the IPU algorithm.

Pritchard and Miller (2009) implemented the IPF with a sparse list-based structure, which is composed of a large number of records with household and person attributes, in the fitting step. Then, a conditional Monte-Carlo simulation is used for the drawing step to fit both household and person marginal distribution simultaneously. Moreover, their study insisted a rounding issue on target marginal and seed cross-table. The rounding issue will be discussed later in this chapter.

Auld and Mohammadian (2010) proposed a new methodology for synthesizing population on multiple levels of household and person using household selection probabilities. On the one hand, the existing household selection probability is calculated by a certain type of households' weight that is divided by the sum of the weights of all other households having the same type in the seed. On the other hand, a new household probability they proposed considers both household marginal distributions and person marginal distributions by combining the person probabilities with the existing household selection probabilities. The details of the selection probability will be handled in the following section.

CO

For the second group of the synthetic population techniques, *CO* (combinatorial optimization) is an iterative algorithm, which is also known as '*entropy maximization*' and '*hill climbing*'. The CO algorithm begins with a random assignment of sample households, and then iteratively replaces an assigned household with another one until reaching a given termination criterion to find a best-fit synthetic population. Compared to IPF, the number of related research using the CO algorithm is limited. Voas and Williamson (2000) proposed a 'sequential fitting procedure' as a new solution to the improvement in the estimation accuracy of synthetic population. In addition, they discussed error measurements for evaluating the quality of the synthetic population. Huang and Williamson (2001) compared the CO with the IPF algorithm by comparing the result of the synthetic population in a small area. At the end, they concluded that both techniques generated a well-fitted synthetic population, but the CO shows a better result in the variability of synthetic population. Melhuish *et al.*, (2002) generated synthetic households using the CO to build the socio-demographic profiles of each CDs (census collection district)[1] in Australia. Then, they evaluated the accuracy of the socio-demographic profiles by comparing with data from the census BCP (basic community profile)[2].

Ryan *et al.*, (2009) applied the CO to predict synthetic population of firms for the City of Hamilton, Ontario, in 1990, in order to compare the performance of the two algorithms. As a result, they concluded that the performance of the CO is better than the IPF. Through the comparison test, they found that the quality of synthetic population depends more on a tabular detail, rather than sample size.

Discussion

The heuristic methods used in IPF and CO techniques show some similarities, but there are also some differences in processing and application perspective. First, the IPF sequentially adjusts a seed marginal to a target marginal distribution, but, on the other hand, the CO first assigns individual sample and then iteratively changes the individual sample with another to find a best solution of synthetic population (Kurban et al., 2011). Second, the IPF has been generally applied in transportation, but the CO has been used for research in geography.

According to related research, there does not seem to exist one single best technique in synthetic population research, because every technique has pros and cons depending on its theoretical feature and purpose. For example, the IPF approach strongly depends on representativeness of the seed data and consistency of the target marginals (Barthelemy and Cornelis, 2012). The CO has also some weaknesses: inability of preserving consistency between attributes and an expensive computation. Therefore, it is important to figure out what data we have (input data) and what shape of output we want (output data) in selecting and applying synthetic population techniques. Note that this chapter only treats the IPF, not the CO algorithm. Table 1 in the appendix lists some of the synthetic population techniques using the IPF with information about using data and features.

WORKING PROCEDURE AND EMPIRICAL CASE

There are two main steps in the process of generating synthetic population using the IPF: a fitting step and a drawing step. The fitting step is to adjust seed data to target marginal distribution by iteratively re-scaling the value of a cell in the seed cross-table. The drawing step is to draw synthetic population by adding individuals/households to the population up to the desired size of the population. In addition to those two steps, there are also other steps before and after the above process: data-processing step and validation step.

This section explains the whole procedure from the data preprocessing to the validation test in the synthetic population research with some examples. Figure 4 in the appendix depicts the whole procedure considering both cases using the IPF and the IPU algorithm, because those algorithms do not require significantly different data structures of input. It also illustrates two different drawing methods, MC sampling and random sampling.

For our empirical case, we generated a synthetic population for Flanders (Belgium) using the IPU algorithm to provide examples in each step of the working procedure. Marginals corresponding to entities at the lowest spatial level (e.g. Building Block or SUBZONE) are assumed to be mutually independent. Furthermore each household belongs to exactly one such spatial area and each person belongs to exactly one household. As a consequence each spatial entity for which marginals need to be approximated constitutes an independent case. Remember that such area either is a basic area (Building Block, SUBZONE and so on) or an aggregate of basic entities used to resolve the zero marginal problems. The computational structure of the problem is embarrassingly parallel due to the spatial independency. Hence we partitioned the input data and were able to run cases in parallel on the VSC (Vlaams Supercomputer Centrum) cluster machines.

In our study area, there is a spatial hierarchy that is composed of three levels; SUPERZONE (a highest level and compatible with a municipality), ZONE (a higher resolution), and SUBZONE (a lowest level with highest resolution) (see Figure 1). We used seed data from a travel survey (OVG) and target marginals from a population in 2010. For the synthetic population technique, we considered both household attributes (income and size) and person attributes (age, job, gender and driver license) (see Table 2 in the appendix).

Data Preprocessing

Obviously the first step in the process is a data preprocessing step in order to prepare input data for the synthetic population technique. To do this,

Figure 1. Study area with zoning system

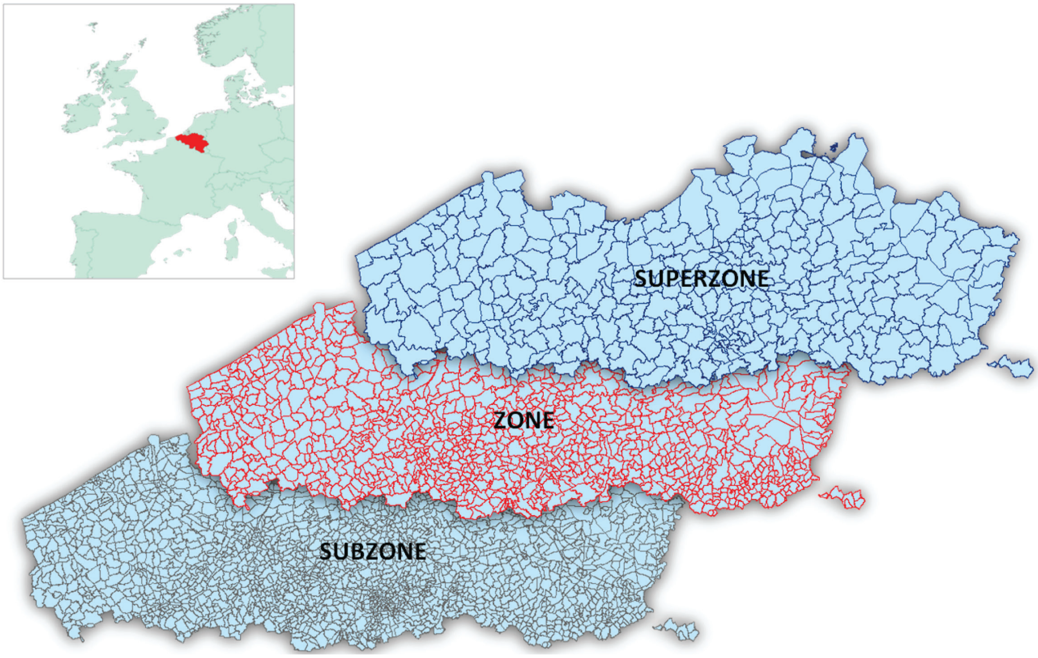
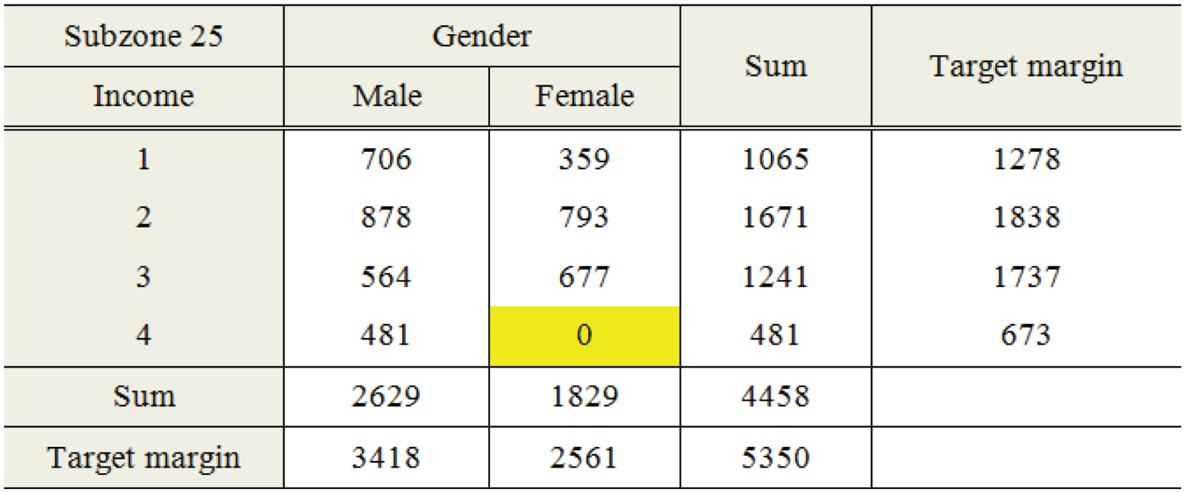


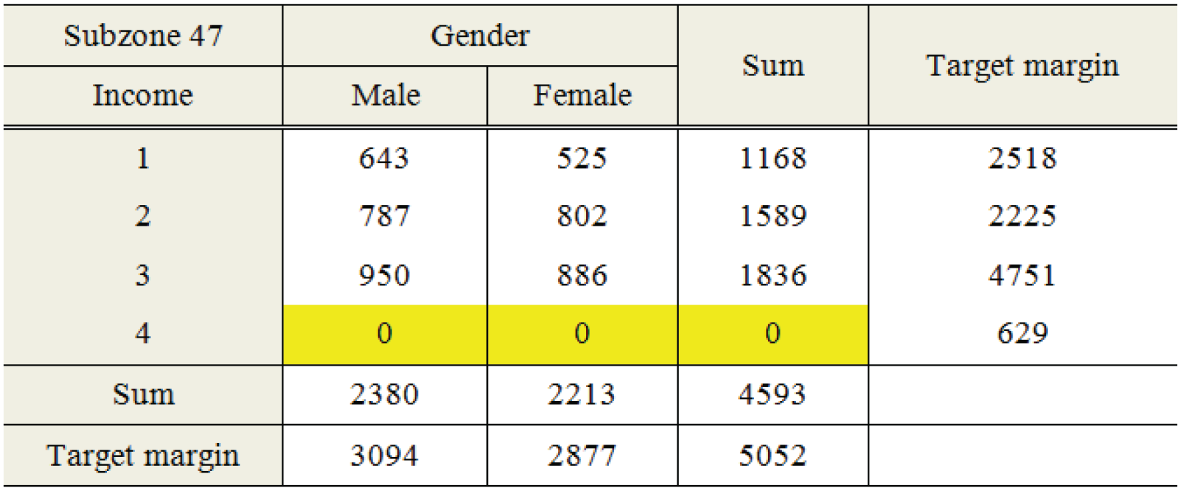
Figure 2. Example of zero-cell problem



Subzone 25	Gender		Sum	Target margin
Income	Male	Female		
1	706	359	1065	1278
2	878	793	1671	1838
3	564	677	1241	1737
4	481	0	481	673
Sum	2629	1829	4458	
Target margin	3418	2561	5350	

we first have to check the data requirement of the synthetic population technique we use, because the data requirement is slightly different depending on the synthetic population techniques. For example, the typical IPF technique needs a cross-table as input, but on the other hand, the IPU uses a list-based data for generating a synthetic population. However, two types of data are typically required for the synthetic population techniques: seed data and target marginal information. The seed data contain disaggregate population which normally describes enough details of the population, but there are only a small number of individual elements in the seed data. On the other hand, the target marginal only has information about the sum of one dimension in an attribute, and not multi-dimensional distribution of several attributes. In general, the seed data can be acquired from a census institute and the target marginals can be easily collected from any institute. The most

Figure 3. Example of zero-marginal problem



Subzone 47	Gender		Sum	Target margin
Income	Male	Female		
1	643	525	1168	2518
2	787	802	1589	2225
3	950	886	1836	4751
4	0	0	0	629
Sum	2380	2213	4593	
Target margin	3094	2877	5052	

important thing in data collection is that attributes in the seed data must be the same as the ones in the target marginals because the basic algorithm in the IPF is to match the marginal distributions of attributes in the seed with the target marginals. Thus, when some attributes in the target marginal are not included in the seed data (or the opposite case), the process does not work due to the mismatching of the attribute dimensions in the marginals. Table 3, Table 4 , and Table 5 in the appendix show an example of source data, which normally includes a small number of household and person data, and target data, which is zone-based and contains a joint set of attributes of the household and person, respectively.

Once input data are available, we have to clean the data if necessary. This is because in the following steps an error may result from such erroneous values, e.g. typo and incorrect value types. After that, the data structure needs to be changed in order to make it feasible for using synthetic population techniques. There are two types of data structures commonly used for synthetic population techniques: a cross-table and a list-based data structure. The cross-table, also called a contingency-table, consists of rows and columns that represent one dimension with cells and marginal totals in each attribute. The ordinary IPF uses a cross-table as input. When the numbers of attributes increase, the size of a cross-table grows exponentially (Muller and Axhausen, 2011). To solve this problem, Williamson et al. (1998) first recommended the sparse list-based data structure as a solution. In addition, it can be applied without any additional processing due to its similarity to a list of attributes in raw data. For these reasons, we used the list-based data structure in our empirical case. Figure 5 and Table 6 in the appendix illustrate an example of the cross-table and the list-based data structure, respectively.

Fitting Step

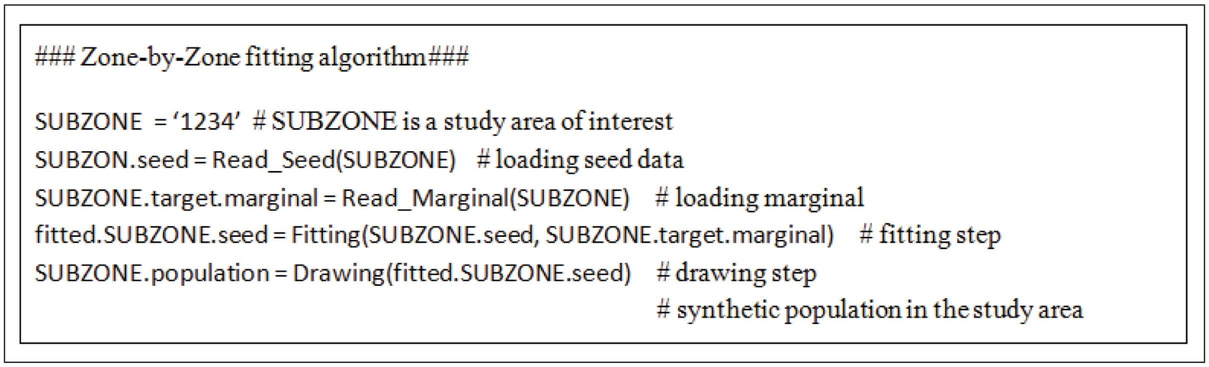
The fitting process between the IPF and the IPU algorithm is different. First, an IPF fitting process is applied to the household data to make them comply with household marginals. Then, a person distribution is determined in the same way as the household distribution. Lastly, the two distributions are combined using a person-per-household ratio at the end of the fitting step or using conditioned MC sampling in the drawing step (or in the preprocessing step; Arentze *et al*., 2007). On the other hand, the IPU first initializes household weights as ‘1’. Next, a (household) adjustment is calculated by computing the proportion to a target marginal. Finally, the household weights are updated by the adjustment. During the updating process, person marginals automatically match with household marginals (also see Figure 4 in the appendix).

The fitting step can be differently operated according to the synthetic population we generate. For example, depending on the levels in the synthetic population, one can chose a single-level or multi-level fitting. However, one cannot apply a multi-level fitting without multi-level data (e.g. household and person data).

Another classification of the fitting process distinguishes between zone-by-zone and multi-zones fitting. In the zone-by-zone fitting, one zone is adjusted to the marginal of that zone alone at a time. On the contrary, the multi-zone fitting matches all zones by aggregating all marginals. While the multi-zone fitting typically shows a better performance than the zone-by-zone fitting, it requires more data storage than the other (Muller and Axhausen, 2011). Therefore, the fitting approach selected depends on the zoning system and data status (both target and source). For example, when the study area has a few small zones, then the multi-zones fitting is a feasible solution. However, in the opposite case, when the study area consists of a few big zones, then the other approach, multi-zones fitting probably is a better solution. This approach is also related to the zero-cell and zero-marginal problems. If the sample size for some of the zones is too small to apply the IPF, then one has to use the multi-zones fitting by aggregating data in those small zones to avoid the zero-cell or zero-marginal problem. Algorithms 1 and 2 indicate the algorithms of zone-by-zone fitting and multi-zone fitting (In the algorithm, a sentence after '#' in each line is a comment on the function or the parameter used in the algorithm. Figures 6 and 7 in the appendix describe an example of zone-by-zone and multi-zones fitting, respectively.

1. Reading seed data in a study area of interest, on a SUBZONE level.
2. Reading target marginal in a corresponding area.
3. Fitting the seed data to the target marginal.
4. Drawing a synthetic population for the study area.

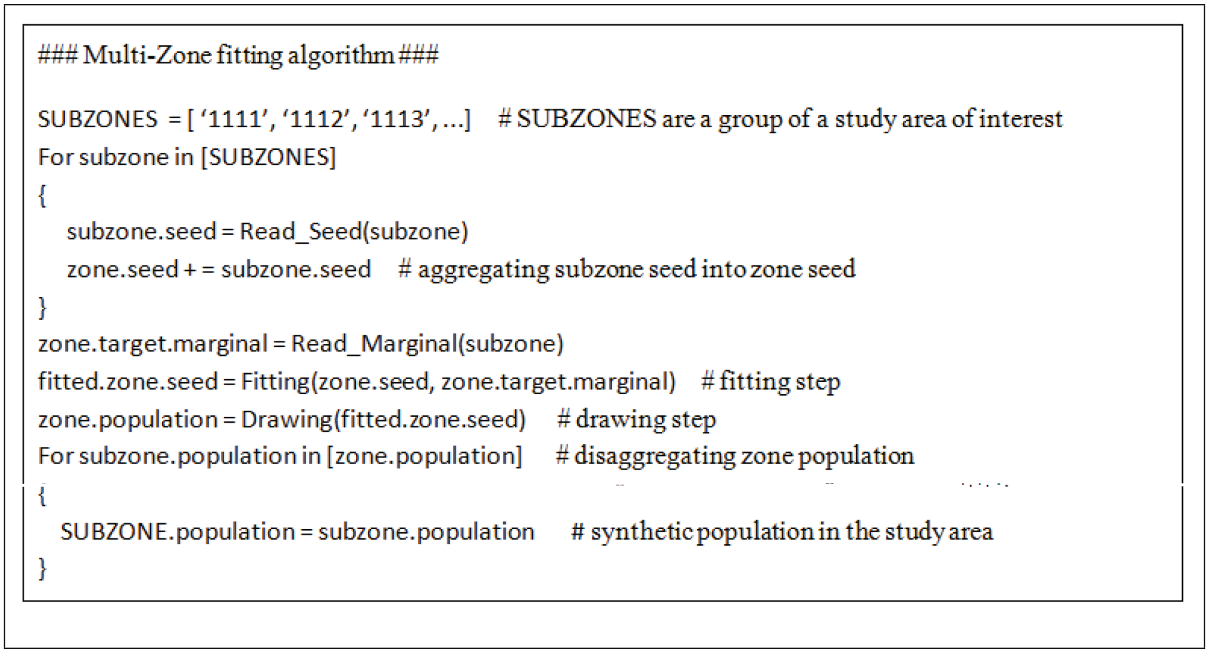
Algorithm 1. Zone-by-zone fitting algorithm



```
### Zone-by-Zone fitting algorithm###

SUBZONE  = '1234'  # SUBZONE is a study area of interest
SUBZON.seed = Read_Seed(SUBZONE)   # loading seed data
SUBZONE.target.marginal = Read_Marginal(SUBZONE)    # loading marginal
fitted.SUBZONE.seed = Fitting(SUBZONE.seed, SUBZONE.target.marginal)    # fitting step
SUBZONE.population = Drawing(fitted.SUBZONE.seed)    # drawing step
                                                     # synthetic population in the study area
```

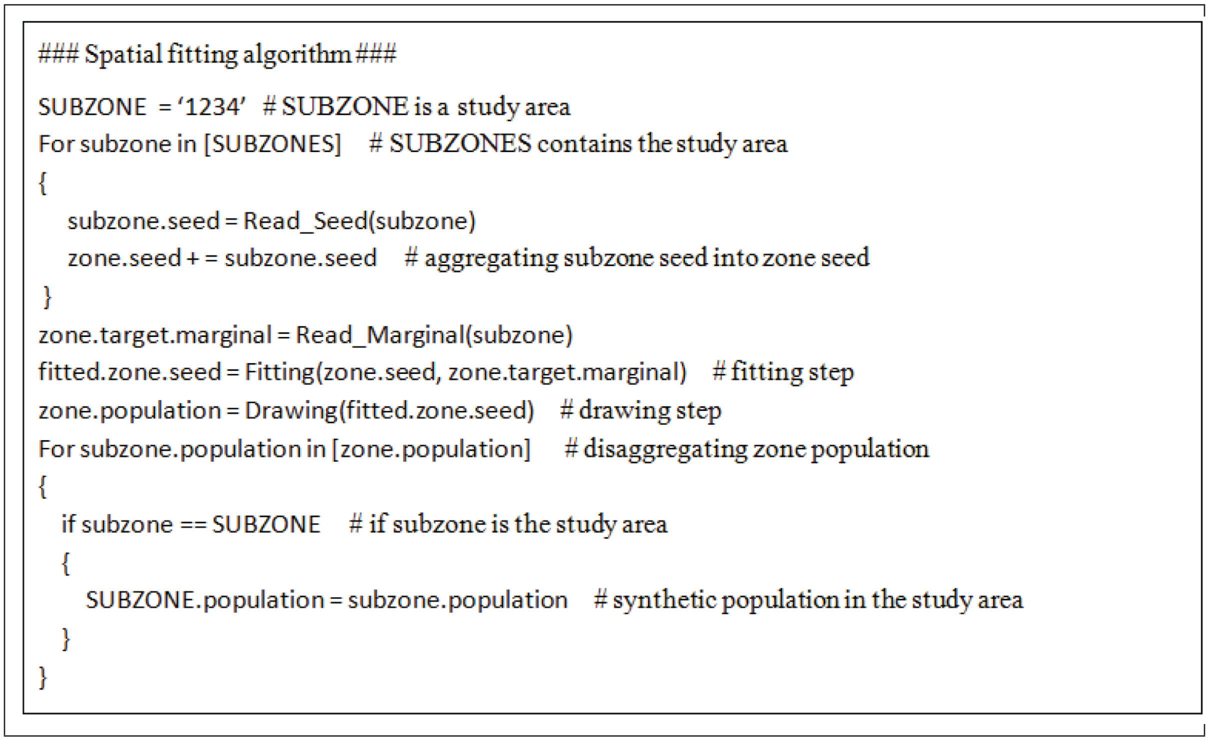
Algorithm 2. Multi-zone fitting algorithm



```
### Multi-Zone fitting algorithm ###

SUBZONES  = [ '1111', '1112', '1113', ...]    # SUBZONES are a group of a study area of interest
For subzone in [SUBZONES]
{
   subzone.seed = Read_Seed(subzone)
   zone.seed + = subzone.seed    # aggregating subzone seed into zone seed
}
zone.target.marginal = Read_Marginal(subzone)
fitted.zone.seed = Fitting(zone.seed, zone.target.marginal)    # fitting step
zone.population = Drawing(fitted.zone.seed)    # drawing step
For subzone.population in [zone.population]    # disaggregating zone population
{
  SUBZONE.population = subzone.population     # synthetic population in the study area
}
```

Algorithm 3. Spatial fitting algorithm



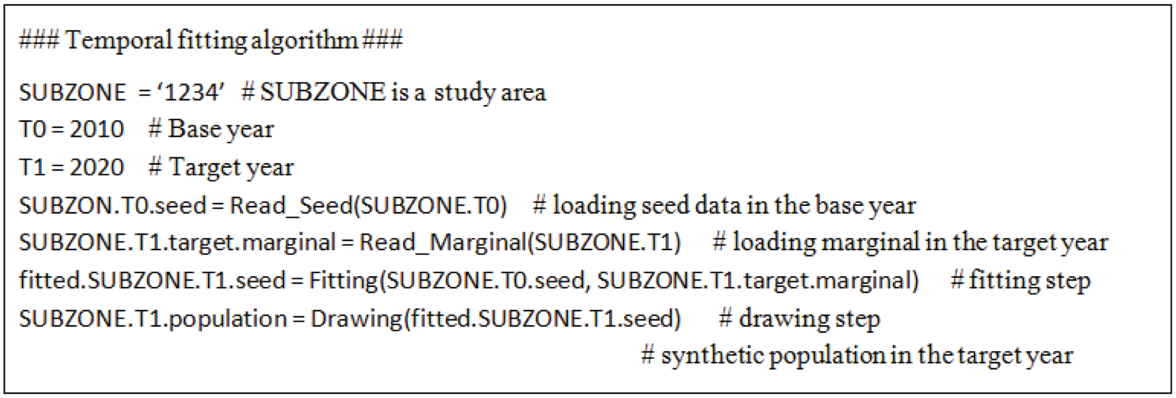
```
### Spatial fitting algorithm ###
SUBZONE = '1234'  # SUBZONE is a study area
For subzone in [SUBZONES]   # SUBZONES contains the study area
{
   subzone.seed = Read_Seed(subzone)
   zone.seed + = subzone.seed    # aggregating subzone seed into zone seed
 }
zone.target.marginal = Read_Marginal(subzone)
fitted.zone.seed = Fitting(zone.seed, zone.target.marginal)   # fitting step
zone.population = Drawing(fitted.zone.seed)   # drawing step
For subzone.population in [zone.population]    # disaggregating zone population
{
  if subzone == SUBZONE    # if subzone is the study area
   {
     SUBZONE.population = subzone.population    # synthetic population in the study area
   }
}
```

1. Reading seed data in all areas, where are belonged to the study area on a SUBZONE level.
2. Merging the see data to a corresponding ZONE, where is an upper-level area of the SUBZONE.
3. Reading target marginal in the ZONE area.
4. Fitting the aggregative seed data to the target marginal.
5. Drawing a synthetic population on a ZONE level.
6. Disaggregating the synthetic population to a SUBZONE level.

Depending on target population, there are two types of fitting process: temporal fitting and spatial fitting. In detail, if the target is a synthetic population at a different time (either past or future), the temporal fitting process needs to be applied. Namely, the base year of the seed data is different from the target year. If the target is a synthetic population in a different spatial level (or other region), a spatial fitting process is feasible. In other words, the spatial level (or location) of the seed data is different from that of the target marginal. Algorithms 3 and 4 describe the algorithm of the spatial fitting and the temporal fitting, respectively. Figures 8 and 9 in the appendix show an example of spatial fitting process and temporal fitting process, respectively.

1. Reading all target marginal in all SUBZONE areas, where belong to the same ZONE area.
2. Merging the target marginal to the ZONE area.
3. Reading seed data in the ZONE area.
4. Fitting the see data to the aggregative target marginal.
5. Drawing a synthetic population in the ZONE area.
6. Disaggregating the synthetic population on a SUBZONE level.

Algorithm 4. Temporal fitting algorithm



```
### Temporal fitting algorithm ###
SUBZONE  = '1234'  # SUBZONE is a study area
T0 = 2010    # Base year
T1 = 2020    # Target year
SUBZON.T0.seed = Read_Seed(SUBZONE.T0)    # loading seed data in the base year
SUBZONE.T1.target.marginal = Read_Marginal(SUBZONE.T1)     # loading marginal in the target year
fitted.SUBZONE.T1.seed = Fitting(SUBZONE.T0.seed, SUBZONE.T1.target.marginal)     # fitting step
SUBZONE.T1.population = Drawing(fitted.SUBZONE.T1.seed)      # drawing step
                                                         # synthetic population in the target year
```

7. Reading the seed data on a SUBZONE level, from the synthetic population.
8. (Again) reading the target marginal in the SUBZONE area.
9. Fitting the seed data to the target marginal.
10. Drawing a synthetic population for the SUBZONE area.

1. Reading seed data in a base year.
2. Reading target marginal in a target year.
3. Fitting the seed data to the target marginal.
4. Drawing a synthetic population for the target year.

Once the fitting step is terminated, you will get either a complete target cross-table in using the normal IPF or a household weight list in using the IPU algorithm. Those are the result of adjusting the seed data to the target marginal distributions so that both the seed and the target marginal totals (almost) match each other in all dimensions of attributes. This means that seed data have been successfully expanded to target marginal. Otherwise, the fitting step has failed. Therefore, a consistency between seed and target marginal needs to be checked before the next step.

Drawing Step

The next step is a drawing process which is to generate synthetic population by drawing population. There are two kinds of drawing methods: household selection probability and MC sampling. The household selection probabilities are calculated by the following formula (Auld and Mohammadian, 2010) using the result of either the complete cross-table or household weights list from the fitting step.

$$P_{i,c} = \frac{W_i}{\sum_{k=1}^{N_c} W_k}$$

where $P_{i,c}$ is a probability of selecting household i with household type c, W_i is a household weight for household i, and N_c is remaining households in sample with household type c. This formula accounts that the probability of selecting household with type c is equal to the current household weight divided by the sum of the other households' weights in the sample. It indicates that as the household weight is higher, the household is more often chosen in the synthetic population. The drawback of this method is that it does not conserve compliance with person-marginals. For this reason, Auld and Mohammadian (2010) suggested a solution of an additional person selection probability with the existing formula. However, it is a quite expensive computation because the selection probability should be calculated after each selection processing.

The other method in the drawing step is a MC sampling which is a random sampling method to obtain numerical results. The MC sampling method selects household with the household weights in a random way. The MC sampling allows selecting an almost infinite number of different set of population so that the probability of generating a best-fit synthetic population becomes higher. Like the previous one, the MC sampling also has a drawback. It could happen that there are only few persons to be sampled left with a few desired numbers of person marginals. In this case, it is not possible to add the rest of the desired persons by sampling from few persons left. When you use the MC sampling in generating synthetic population, you need to be careful of the processing time because the MC sampling is quite sensitive to the number of iterations for the sampling. Although more iterations in the MC sampling can produce a better result, they result in a longer processing time. Hence, finding the optimal number of iterations is important in using Monte-Carlo sampling. Therefore, if you have a lot of combination sets of attribute, then the MC sampling is a better solution in terms of an efficient processing. Otherwise, the household selection probability is a better solution to keep a higher chance to produce complete synthetic population without marginals left (also see Figure 4 in the appendix).

Validation Step

After the drawing step, you will acquire synthetic population in your study area of interest. To check the accuracy of estimation, the synthetic population needs to be validated against real population using a goodness-of-fit measure. In statistics, there are three types of error measurements: *traditional* statistics, *information-based* statistics and general *distance-based* statistics.

In the traditional statistics, R^2 and chi-square are the two most commonly-applied goodness-of-fit measurements. However, the R^2 statistics has been argued by several researchers due to its insensitivity to variations in model specification and missing concept to evaluate model performance across different data sets (Black and Salter, 1975; Wilson, 1976). The information-based statistics originated from "information gain statistics" in Kullback and Leibler (1951) includes the phi statistics, the psi statistics, and the measure of absolute entropy difference. The information-based statistics have a limitation in that the information gain is sensitive to the distribution of over and under estimations (Smith and Hutchinson, 1981). The general distance statistics are defined by functions of an element of the observed matrix and estimated matrix. Among the general distance statistics, a standardized root mean square error (SRMSE) has been commonly used for the validation test for synthetic population in transportation because of its theoretical relevance in statistical modeling (Hyndman and Koehler, 2006). The SRMSE is calculated as follows (Pitfield, 1978):

$$SRMSE\sqrt{\sum\sum\frac{(t_{ij}-\hat{t}_{ij})^2}{m\times n}}\ /\ (\sum_{i}\sum_{j}\frac{t_{ij}}{m\times n})$$

where $\hat{t}_{ij}$ is the estimated number of population elements with attributes i and j, and t_{ij} is an observed number of population. m and n are the number of attribute values for attributes i and j, respectively. A value of zero in the SRMSE means a perfect match, and '1' means no matching between estimated and observed data. Even its popularity in error measurements, the SRMSE cannot always serve as a best measurement with some problems. For example, the SRMSE is only feasible in a specific condition that the sum of observed frequencies is exactly the same as the sum of the estimated frequencies (Knudsen and Fotheringham, 1986). In practice, this condition is not always met, especially in evaluating synthetic population data. This is because the only case that perfectly matches between seed and target marginals can satisfy this condition.

$$\sum_i \sum_j t_{ij} = \sum_i \sum_j \hat{t}_{ij}$$

Now, what is the best measurement for a validation test? According to related research, there seems to be no concrete answer to this question because every measurement has merits and drawbacks depending on data and validation purpose. Table 7 in the appendix lists error measurements for testing the performance of synthetic population techniques. Voas and Williamson (2001) proposed some criteria on the decision of error measurements.

> *We seek a goodness-of-fit statistic that:*
>
> - *Can be used for comparisons across tables;*
> - *Produces results corresponding to our intuitive sense of fit;*
> - *Will measure both tabular fit and (via its components) internal fit;*
> - *Will compare counts or totals and not just relative frequencies;*
> - *Is not burdensome to calculate or (where appropriate) to test;*
> - *Has a known, tractable sampling distribution;*
> - *Will be familiar or at least acceptable to the user community. (pp. 196 in Voas and Williamson (2001)*

Summary

In this section, we described every processing step with examples: preprocessing step, fitting step, drawing step and validation step. The preprocessing step is to collect and clean input, and change the data structure if necessary. The fitting step is to adjust seed data to target marginal distribution using different fitting approach (single- and multi-level, zone-by-zone and multi-level, temporal and spatial fitting) feasible for the target synthetic population. Then, the drawing step is to draw synthetic population by expanding the seed to the desired number of population using a sampling tool (household selection probability and MC sampling). At the end, the validation step is to evaluate the performance of the synthetic population technique by computing an estimation error using an error measurement.

ISSUES AND PROPOSED SOLUTIONS

This section deals with some problems that we could experience during the working procedure of generating a synthetic population using whatever techniques. This is because even using a very good technique may have to deal with data problems. We also propose a solution to each problem with an example.

Zero-Cell and Zero-Marginal

A zero-cell problem, also referred to as missing values in the literature, is first addressed by Beckman *et* al. (1996). The zero-cell problem occurs when a target marginal is not zero in an attribute dimension without a corresponding sample in the source. This normally happens when generating synthetic population in rather small regions, because there is a high probability of no representative sample in a certain combination of attributes. In this case, the IPF process cannot converge to a solution because the corresponding cell in the cross-table always takes a zero value during the process due to a zero division error. Figure 2 shows an example of the zero-cell in a two-dimension cross-table.

There are few solutions to the zero-cell problem. The simplest solution is to assign an arbitrarily small value (e.g. 0.01) to the zero cells, which is called as 'tweaking approach' in Beckman *et al*. (1996). However, Beckman et al. (1996) and Guo and Bhat (2006) addressed that the solution may introduce an arbitrary bias. Another solution is to substitute for a zero-cell value by a derived value from overall distribution in the whole sample. Although this solution is quit suitable for the zero-cell problem in the small area, it is possible to over-represent and ignore the local characteristic in that area, for example no population with high income in a rural area (Ye et al., 2009). Another two solutions, using a maximum-iteration value and a category reduction, are indirect solutions to preventing from the zero-cell problem during the IPF process (Guo and Bhat, 2006). The former solution is to avoid non-convergence occurred by the zero-cell problem by terminating the IPF process when reaching the pre-specified maximum-iteration value. The latter one is to lower the chances to get the zero-cells in a cross-table by reducing the sparse categories. For instance, a cross-table with 5 categories has lower chance to get the zero-cell value than a cross-table with 10 categories (the section of category reduction deals with this category reduction in detail). However, those two solutions can affect on other factors, for example computing resource and model performance, and also trigger another problem in the IPF process.

A zero-marginal problem occurs with the same reason as the zero-cell problem, but it is limited to the case of using the IPU algorithm. As for the theoretical features of the IPU algorithm, all households corresponding to the zero-marginal category in the source will get a zero weight. Then, the household with zero weights never get positive weights even if some of the households have to be assigned with non-zero weights. This is because the IPU algorithm can normally update the household weights after each iteration, but it is not possible to update the zero weights due to fact that the denominator for adjusting marginal totals will always take a zero value. Ye *et al*. (2009) suggested a solution to this problem that assigns an arbitrary small number, e.g. 0.001, to the zero-marginal categories. The effect of the arbitrary small margins can be alleviated after little iteration, and the process of updating weights can avoid the zero-marginal problem. Figure 3 shows an example of the zero-marginal problem in the IPU algorithm.

Rounding

The IPF process estimates the value of a cell in a target cross-table by adjusting seed marginal to the target marginal by means of linearly rescaling the value of a cell in the seed cross-table. As a result, the IPF process outputs the value of a cell with a real number, which indicates the number of households (or persons) with certain socio-demographic attributes (e.g. age, gender, income and so on) in the population. Thus, the number should be an integer in the drawing step, otherwise, the number needs to be converted to an integer number by rounding up or down. In rounding a number, an expected problem is that after rounding those numbers, the marginal totals in the source cannot keep consistent with the target marginals. This means that rounding the numbers, also referred to as "*integerization*", in a cross-table can lead to an unexpected biased distribution in the synthetic population process (Bowman, 2004). In addition to simple rounding methods (e.g. rounding up/down and rounding ceiling/floor), there are a number of rounding methods, such as an arithmetic rounding, a bucket rounding and stochastic rounding (introduced by Ye *et al*., 2009). Note that after rounding, the marginal totals in the seed cross-table should be at least consistent with the marginal totals in the target cross-table, as far as possible.

Category Reduction

As mentioned in the section of zero-cell and zero-marginal, the category reduction serves as a solution to two problems, a zero-cell (or zero-marginal) and memory, in using synthetic population techniques. Compared to the IPU algorithm with a sparse list-based data structure, the ordinary IPF with a cross-table requires exponentially bigger memory as the number of categories in attributes increase. For example, consider a two-dimension cross-table that consists of two attributes each with three categories. In this case, there are nine cells (3 by 3) in the cross-table. Next, consider a two-dimension cross table consisting of two attributes with 3+1 categories, and then the cross-table has sixteen cells (4 by 4). As you can see in this example, an increase in the number of cells, which consumes more memory, is exponentially proportional to an increase in the number of categories. In addition, the more cells, the more often zero-valued cells will occur in a cross-table.

As always, the category reduction has not only those advantages but also some disadvantages. First, this solution cannot guarantee a better performance because using less categories means that less control variables are used for the synthetic population process (Auld *et al*., 2008). Furthermore, this solution may ignore a local characteristic, for example reducing an income category would overlook the significant difference between urban area and rural area. Hence, it is important to find the optimal number of categories in the synthetic population research. Auld *et al*. (2008) proposed a user-specified percentage threshold as a category reduction method. At first, a user specifies the percentage threshold. Then, a category which does not exceed the percentage threshold is combined with a neighboring category. For instance, for a given percentage threshold of 20%, there is no attribute with more than five categories. In addition, this solution should be applied to resolve the zero-cell issue together with adapting an input and model itself to fit the local characteristics lost by the category reduction.

CONCLUSION

As the ABM becomes more popular in transportation, the demand of micro-data increases. Generally, it is difficult to collect such micro-data due to privacy protection and high cost. Thus, more researchers are trying to produce synthetic population as an alternative resource using synthetic population techniques. However, generating a synthetic population is relatively complicated for non-experts or beginners because there is no explicit terminology and there are no concrete solutions to some issues and problems in the field. In this sense, this chapter aims at providing a beginner with a guideline on how to generate synthetic population using some techniques.

The chapter accounts for why we need synthetic population techniques and the chapter goal and structure in the introduction section. For a beginner, some terminologies which are commonly referred to in related research are described in the section of related research. Then, two groups of related research, IPF and CO, are separately introduced in the following two sections. The next section describes the whole process from data collection to the validation test in building a synthetic population (using the IPU algorithm). In the section of issues and proposed solutions, we introduce some issues (e.g. zero-cell and zero-marginal, rounding, and category reduction) in the synthetic population research, and also provide a solution to each issue.

Due to a limited space, this chapter cannot cover everything about the synthetic population techniques, but instead we made effort to give an answer to some practical questions of why, what and how generating synthetic population can be done. The research was supported by means of an empirical case where a synthetic population was generated for Flanders (Belgium). In our future work, we will provide more details of the working process and further issues and challenges in this field, and also introduce new research developing a new synthetic population technique without sample data.

ACKNOWLEDGMENT

For the empirical case we used the infrastructure of the VSC - Flemish Supercomputer Center, funded by the Hercules foundation and the Flemish Government - department EWI.

REFERENCES

Arentze, T. A., Timmermans, & Hofman. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record*, (11): 85–91. doi:10.3141/2014-11.

Auld, J. Mohammadian, & Wies. (2008). *Population synthesis with control category optimization*. Paper presented at the 10th International Conference on Application of Advanced Technologies in Transportation. Athens, Greece.

Auld, J. Mohammadian, & Wies. (2010). *An efficient methodology for generating synthetic populations with multiple control levels*. Paper presented at the the 89th Annual Meeting of the Transportation Research Board. Washington, DC.

Barthelemy, J., & Cornelis, E. (2012). *Synthetic populations: review of the different approaches*. CEPS/INSTEAD..

Beckman, R. J., Baggerly, & McKay. (1996). Creating synthetic baseline populations. *Transportation Research Part A, Policy and Practice*, *30*(6), 415–429. doi:10.1016/0965-8564(96)00004-3.

Black, J. A., & Salter, R. T. (1975). A statistical evaluation of the accuracy of a family of gravity models. *Proceedings - Institution of Civil Engineers*, *2*(59), 1–20. doi:10.1680/iicep.1975.3839.

Deming, W. E., & Stephan. (1940). On the least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Annals of Mathematical Statistics*, *11*(4), 427–444. doi:10.1214/aoms/1177731829.

Guo, J. Y., & Bhat. (2007). Population synthesis for microsimulating travel behavior. *Transportation Research Record*, (12): 92–101. doi:10.3141/2014-12.

Huang, Z., & Williamson, P. (2001). *Comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata* (Working Paper, 2001/2). Liverpool, UK: Department of Geography, University of Liverpool.

Hyndman, R. J., & Koehler, A. B. (2006). Another look at measures of forecast accuracy. *International Journal of Forecasting*, *22*(4), 679–688. doi:10.1016/j.ijforecast.2006.03.001.

Knudsen, D. C., & Fotheringham. (1986). Matrix comparison, goodness-of-fit, and spatial interaction modeling. *International Regional Science Review*, *10*(2), 127–147. doi:10.1177/016001768601000203.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics*, *22*(1), 79–86. doi:10.1214/aoms/1177729694.

Kurban, H., Gallagher, R., Kurban, G. A., & Persky, J. (2011). A beginner's guide to creating small-area cross-tabulations. *Cityscape (Washington, D.C.)*, *13*(3), 225–235.

Melhuish, T., Blake, M., & Day, S. (2002). An evaluation of synthetic household populations for census collection districts created using optimisation techniques. *Australasian Journal of Regional Studies*, *8*(3), 269–387.

Muller, K., & Axhausen. (2011). *Population synthesis for microsimulation: State of the art*. Paper presented at the 90th Annual Meeting of the Transportation Research Board. Washington, DC.

Pitfield, D. E. (1978). Sub-optimality in freight distribution. *Transportation Research*, *12*(6), 403–409. doi:10.1016/0041-1647(78)90028-X.

Pritchard, D. R., & Miller. (2009). *Advances in agent population synthesis and application in an integrated land use and transportation model*. Paper presented at the 88th Annual Meeting of the Transportation Research Board. Washington, DC.

Ryan, J., Maoh, & Kanaroglou. (2009). Population synthesis: Comparing the major techniques using a small, complete population of firms. *Geographical Analysis*, *41*(2), 181–203. doi:10.1111/j.1538-4632.2009.00750.x.

Smith, D. P., & Hutchinson, B. G. (1981). Goodness-of-fit statistics for trip distribution models. *Transportation Research*, *15*(4), 295–303. doi:10.1016/0191-2607(81)90011-X.

Voas, D., & Williamson, P. (2000). An evaluation of the combinatorial optimisation approach to the creation of synthetic microdata. *International Journal of Population Geography*, *6*(5), 349–366. doi:10.1002/1099-1220(200009/10)6:5<349::AID-IJPG196>3.0.CO;2-5.

Voas, D., & Williamson, P. (2001). Evaluating goodness-of-fit measures for synthetic microdata. *Geographical and Environmental Modelling*, *5*(2), 177–200. doi:10.1080/13615930120086078.

Williamson, P., Birkin, M., & Rees, P. H. (1998). The estimation of population microdata by using data from small area statistics and samples of anonymised records. *Environment & Planning A*, *30*(5), 785–816. doi:10.1068/a300785 PMID:12293871.

Wilson, S. R. (1976). Statistical notes on the evaluation of calibrated gravity models. *Transportation Research*, *0*(5), 343–345. doi:10.1016/0041-1647(76)90114-3.

Ye, X. Konduri, Pendyala, Sana, & Waddell. (2009). *A methodology to match distributions of both household and person attributes in the generation of synthetic populations*. Paper presented at the 88th Annual Meeting of the Transportation Research Board. Washington, DC.

KEY TERMS AND DEFNITIONS

CO (Combinatorial Optimization): A synthetic population technique that adjusts seed data to target marginals by swapping the household randomly selected from the seed data to another until getting a best-fit output.

Drawing: Sampling procedure in synthetic population research that synthesizes population by adding a household up to the desired number of households in the target marginal.

Fitting: Matching procedure in synthetic population research that adjusts seed data to target marginal distribution.

IPF (Iterative Proportional Fitting): A synthetic population technique that iteratively estimates cell values of a cross-table in the source in order to match given target marginal distributions.

IPU (Iterative Proportional Updating): A synthetic population technique that iteratively updates household weights of each sample record to match seed data to target marginals.

Marginal: A row or column total in a cross-table, normally calculated along each of the row or column dimension.

Source (or Seed Data): Initial value for the matrix cells that has been derived from domain knowledge, e.g. census survey.

Sparse List-Based Data Structure: A data structure consisting of the microdata (sample) entries and weights attached to each entry. List based data structures are used in cases where only a small part of the possible attribute combinations occur in reality (sparseness).

Target: Aggregate data for which the marginal distribution is given in a study area of interest.

APPENDIX

Table 1. List of synthetic population techniques

Researcher	Model	Goal	Input	Fitting	Drawing	Validation
Beckman *et al*. (1996)	Original IPF	Tarrant County, Texas, 1990	Census STF-3A (margin) PUMS (seed)	Single-level (only household)	Household selection probability	
Arentze *et al*. (2007)	Relation matrix	Dutch population, 1995	1995 OVG (seed)	Multi-level (dealt in a preprocessing step)		CHAID
Guo & Bhat (2007)	Original IPF	Dalla/Fort-Worth Metropolitan Area, Texas, 2000	2000 US census SF1 (margin) 2000 PUMS (seed)	Multi-level	Advanced household selection probability	PD (percentage difference) APD (absolute PD) AAPD (average APD)
Ye *et al*. (2009)	IPU algorithm	Maricopa County Region, Arizona, 2000	2000 Census summary file (margin) 2000 PUMS (seed)	Multi-level (by updating household weights)	MC (Monte-Carlo) sampling	Chi-square statistic
Auld *et al*. (2010)	Advanced IPF	Chicago-land six-country region		Multi-level Zone-by-zone	MC sampling New household selection probability	WAAPD (weighted average absolute percentage difference) FT (Freeman-Tukey) statistic
Muller & Axhausen (2011)	Hierarchical IPF	Switzerland, 2000	2000 Swiss census	Multi-level (using an entropy-optimizing method)		G^2 SRMSE (standardized root mean square error)
Ptrichard & Miller (2012)	Advanced IPF	Toronto Census Metropolitan Area	1986 Toronto census	Multi-level Multi-zone	Conditioned MC sampling	SRMSE

Figure 4. Working process of synthetic population technique

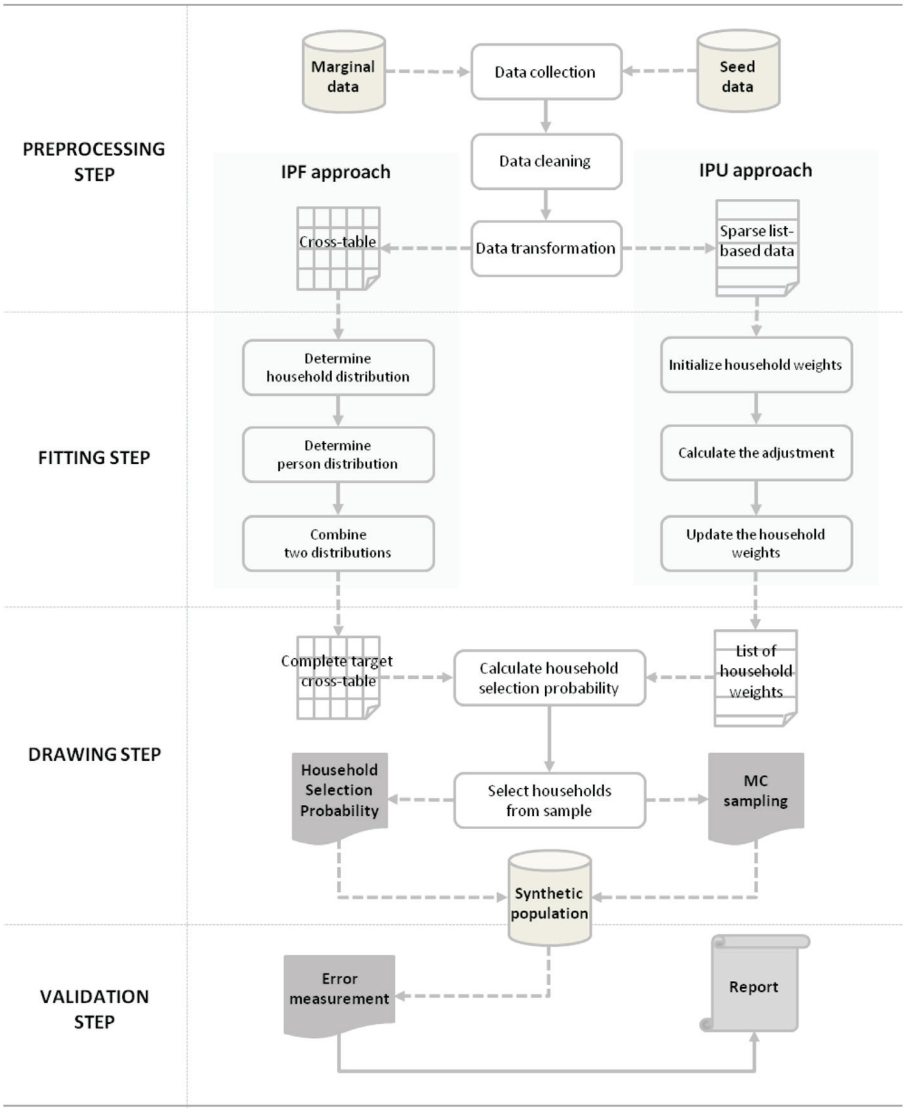

Marginal data
Data collection
Seed data
PREPROCESSING STEP
IPF approach
Data cleaning
IPU approach
Cross-table
Data transformation
Sparse list-based data
FITTING STEP
Determine household distribution
Determine person distribution
Combine two distributions
Initialize household weights
Calculate the adjustment
Update the household weights
DRAWING STEP
Complete target cross-table
Calculate household selection probability
List of household weights
Household Selection Probability
Select households from sample
MC sampling
Synthetic population
VALIDATION STEP
Error measurement
Report

Table 2. List of household and person attributes

Attribute	Category
Residence	**Subzone ID**
Income	‘1’ = 0 – 1249 ‘2’ = 1250 – 2249 ‘3’ = 2250 – 2249 ‘4’ = 3250+ Unit: euro
Number of members	‘1’ = 1 ‘2’ = 2
Age	‘1’ = 18 – 34 ‘2’ = 35 –54 ‘3’ = 55 – 64 ‘4’ = 65 – 74 ‘5’ = 75+
Job	‘0’ = No work ‘2’ = Work
Gender	‘1’ = Male ‘2’ = Female
Driver license	‘0’ = No ‘1’ = Yes

Table 3. Example of source data: household data

Household ID	Residence	Income	Size
1	3	1	1
2	15	1	1
3	20	4	2
4	12	1	1
5	19	5	2
6	25	4	2

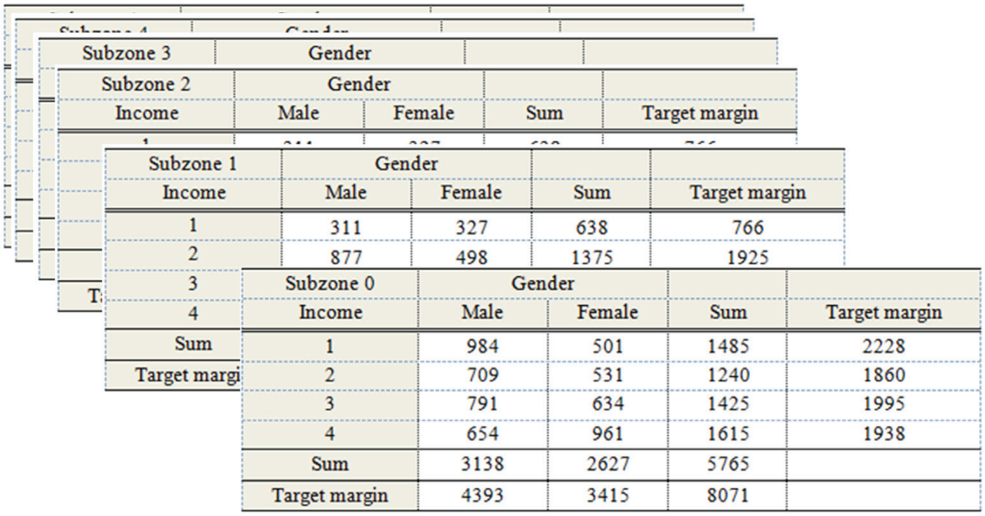
Table 4. Example of source data: person data

Person ID	Household ID	Age	Job	Gender	Driver License
1	1	2	2	2	1
2	2	3	0	2	0
3	3	4	0	2	0
4	3	4	0	1	1
5	4	1	2	1	1
6	5	2	2	2	1
7	5	2	2	1	1
8	6	3	0	1	1
9	6	3	0	2	1

Table 5. Example of target data

Subzone	Households	Persons	Income 1	Income 2	Size 1	Size 2	Age1
0	4138	7207	734	711	1451	2686	782
1	767	1442	136	132	269	498	141
2	1496	2709	265	257	525	972	314
3	8073	9169	1896	1905	3591	4483	1367
Age2	**No Work**	**Work**	**Male**	**Female**	**No Driving License**	**Driving License**	
1117	3031	4177	4052	3155	1688	5519	
202	679	763	751	691	416	1026	
448	1043	1666	1578	1131	566	2144	
1952	4267	4902	5074	4094	2533	6635	

Figure 5. Example of (multi-zone) cross-table



Subzone 3	Gender		

Subzone 2	Gender			
Income	Male	Female	Sum	Target margin

Subzone 1	Gender			
Income	Male	Female	Sum	Target margin
1	311	327	638	766
2	877	498	1375	1925
3				
4				
Sum				
Target margi				

Subzone 0	Gender			
Income	Male	Female	Sum	Target margin
1	984	501	1485	2228
2	709	531	1240	1860
3	791	634	1425	1995
4	654	961	1615	1938
Sum	3138	2627	5765	
Target margin	4393	3415	8071	

Table 6. Example of sparse list-based data structure

Index	Residence	Income	Size	Age	Job	Gender	Driver License	Weight
1	3	1	1	2	2	2	1	39.3
2	15	1	1	3	0	2	0	41.2
3	20	4	2	4	0	2	0	33.10
4	20	4	2	4	0	1	1	50.10
5	12	1	1	1	2	1	1	47.8
6	19	5	2	2	2	2	1	47.3
7	19	5	2	2	2	1	1	17.6
8	25	4	2	3	0	1	1	35.1
9	25	4	2	3	0	2	1	11.10

Figure 6. Example of zone-by-zone fitting process

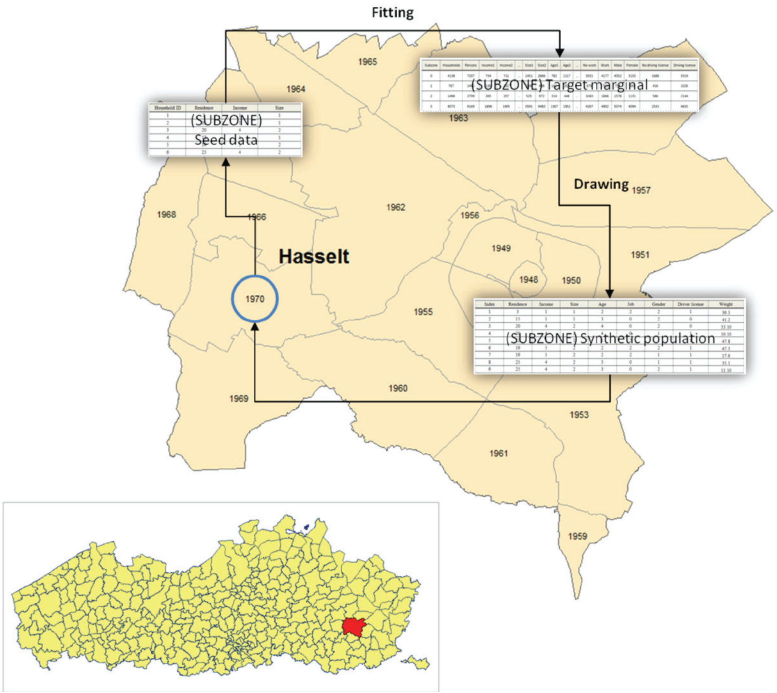


Figure 7. Example of multi-zones fitting process

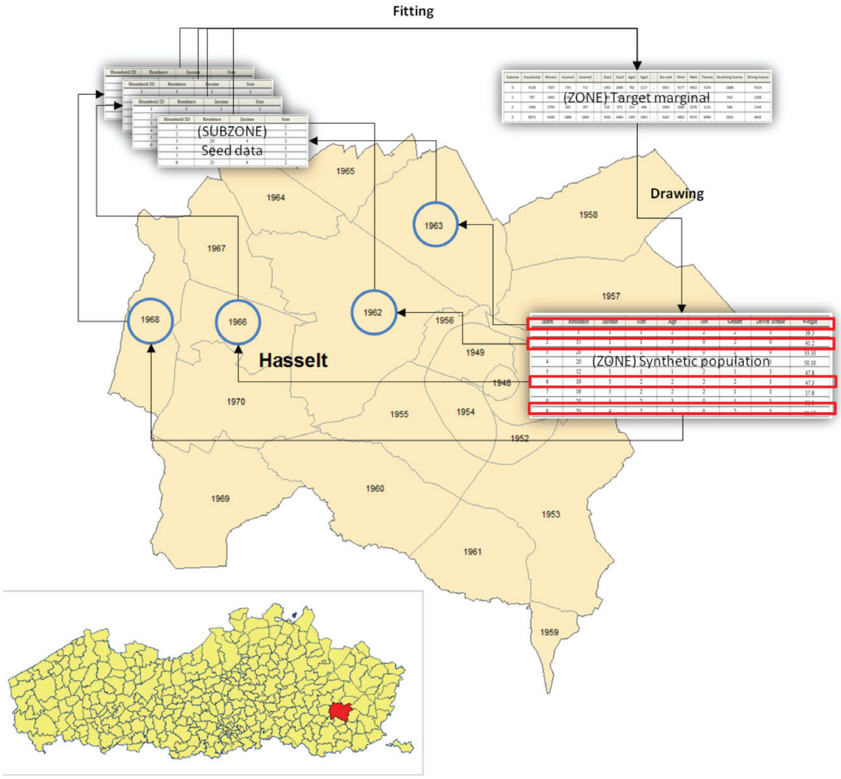


Table 7. List of error measurements (Hyndman and Koehler, 2006)

Category	Error Measurement	Limit
Scale-dependent error	RMSE (root mean square error)= $\sqrt{mean(N-\hat{N})^2}$ MAE (mean absolute error) = $\mathrm{mean}(\lvert\mathrm{N}-\hat{\mathrm{N}}\rvert)$	- sensitive to outliers
Percentage error	MAPE (mean absolute percentage error) = $\mathrm{mean}(\lvert\, 100\times\{(\mathrm{N}-\hat{N})\div \mathrm{N}\}\,\rvert)$ RMSPE (root mean square percentage error = $\sqrt{\mathrm{mean}\left\{100\times\left(\left(\mathrm{N}-\hat{\mathrm{N}}\right)\div\mathrm{N}\right)\right\}^2}$	- infinite or undefined if *N*=0 - skewed distribution if *N* is close to zero
Relative error	MRAE (mean relative absolute error) = $\mathrm{mean}(\lvert\,(\mathrm{N}-\hat{\mathrm{N}})\div(\mathrm{N}-\hat{\mathrm{N}})^*\,\rvert)$ GMRAE (geometric mean relative absolute error) = $gmean(\lvert\,(N-\hat{N})\div(N-\hat{N})^*\,\rvert)$ $\#(N-\hat{N})^*$ is an expected error by a benchmark method	- infinite variance if an expected error has positive probability density at 0

[1]Census Collection Districts (CDs) are designed for use in census years for the collection and dissemination of Population Census data (http://www.abs.gov.au)

[2]Basic Community Profile (BCP) is the primary profile. It consists of 46 tables containing key Census characteristics on persons, families and dwellings (http://www.abs.gov.au).