Population Synthesis in the CT-RAMP ABM

Peter Vovsha, Jim Hicks, Binny Paul (Parsons Brinckerhoff)

Contents

[1 Essence of Population Synthesis 4](#_Toc390719934)

[1.1 Role of Population Synthesizer in ABM 4](#_Toc390719935)

[1.2 Input 4](#_Toc390719936)

[1.3 Output 5](#_Toc390719937)

[1.4 Quality Criteria for Population Synthesis 5](#_Toc390719938)

[1.5 Controlled and Uncontrolled Variables 6](#_Toc390719939)

[2 Existing Methods 9](#_Toc390719940)

[2.1 3-Step Procedures Based on IPF of Joint Household Distributions 9](#_Toc390719941)

[2.2 Combinatorial Optimization and Simulated Annealing 15](#_Toc390719942)

[2.3 First Examples of List Balancing 18](#_Toc390719943)

[3 Proposed Method 20](#_Toc390719944)

[3.1 Requirements for Population Synthesis 20](#_Toc390719945)

[3.2 Outline of Main Steps 20](#_Toc390719946)

[4 List Balancing for Single Zone 21](#_Toc390719947)

[4.1 Basic Formulation with Fixed Constraints 21](#_Toc390719948)

[4.2 Introducing Relaxations 26](#_Toc390719949)

[4.3 Newton-Raphson Method w/o Relaxations 27](#_Toc390719950)

[4.4 Newton-Raphson Method w/Relaxations 27](#_Toc390719951)

[4.5 Efficient Solution Algorithm and Pseudo-Code 27](#_Toc390719952)

[4.6 Particular Case of Household-Level Controls Only 29](#_Toc390719953)

[4.7 Expansion of Household Survey Sample by List Balancing 30](#_Toc390719954)

[5 Discretizing Methods 34](#_Toc390719955)

[5.1 Post-Discretizing 34](#_Toc390719956)

[5.1.1 Fixed Constraints 34](#_Toc390719957)

[5.1.2 Relaxed Constraints 35](#_Toc390719958)

[5.2 Direct Formulation of Integer Problem 35](#_Toc390719959)

[5.2.1 Fixed Constraints 35](#_Toc390719960)

[5.2.2 Relaxed Constraints 36](#_Toc390719961)

[5.3 Discretizing Residual Fractions 36](#_Toc390719962)

[5.3.1 Fixed Constraints 37](#_Toc390719963)

[5.3.2 Relaxed Constraints 37](#_Toc390719964)

[6 Consideration of Multiple Zones 38](#_Toc390719965)

[6.1 Meta-Controls 38](#_Toc390719966)

[6.1.1 General Formulation 38](#_Toc390719967)

[6.1.2 Particular Case with no Explicit Consideration of Uniformity 40](#_Toc390719968)

[6.1.3 Meta-Balancing Efficient Implementation Algorithm 42](#_Toc390719969)

[6.2 Household Allocation to Sub-Zones 43](#_Toc390719970)

[7 Validation of Population Synthesizer 45](#_Toc390719971)

[8 Complete Regional Population Synthesizer for MAG 46](#_Toc390719972)

[8.1 Structure of Population Synthesizer 46](#_Toc390719973)

[8.2 Zone System in the MAG-PAG Metropolitan Region 48](#_Toc390719974)

[8.3 Household Sample from PUMS 48](#_Toc390719975)

[8.4 Controls 48](#_Toc390719976)

[8.5 Summary of Results 48](#_Toc390719977)

[9 Complete Regional Population Synthesizer for JTMT 49](#_Toc390719978)

[9.1 Zone System in the Jerusalem Metropolitan Region 49](#_Toc390719979)

[9.2 Household Sample from the Israeli Population Census 49](#_Toc390719980)

[9.3 Controls 49](#_Toc390719981)

[9.4 Summary of Results 49](#_Toc390719982)

[10 References 49](#_Toc390719983)

[10.1 Conventional PopSyn Based on Joint Distribution of Households and IPF 49](#_Toc390719984)

[10.2 Combinatorial Optimization and Annealing 50](#_Toc390719985)

[10.3 List Balancing 51](#_Toc390719986)

[11 List Balancing Using the Newton-Raphson Method 51](#_Toc390719987)

[11.1 Strict Formulation without Relaxations 51](#_Toc390719988)

# Essence of Population Synthesis

## Role of Population Synthesizer in ABM

A key technical advantage of Activity-Based Models (ABMs) is the disaggregate treatment of travelers in a microsimulation framework. Population Synthesizer (PopSyn) is the necessary starting point of any ABM that creates a list of households and persons with detailed individual attributes in each zone.

Population synthesis, within the context of transportation modeling, land use modeling and similar domains, is the process of creating a representation of a complete, disaggregate population by combining a sample of disaggregate members of a population in a way as to match key distributions for the entire population [*Beckman et al, 1996*].

Quality and reasonableness of the PopSyn has a crucial impact on the subsequent model chain of the entire ABM. PopSyn can also improve an aggregate 4-step model by providing zonal distributions of households and persons as an input to trip generation models.

There are more than 20 ABMs in practice in the U.S. at different stages of development that provides useful examples of various PopSyn structures and features. The purpose of this document is to synthesize the best features of these developed PopSyns including innovative components that were introduced recently and included in ABMs developed for the Maricopa Association of Governments (MAG) and Jerusalem Transportation Master plan Team (JTMT). The innovative components address the requirements to match *both household-level and person-level attributes* in the model area as well as the possibility to set the corresponding *controls* at *different levels of geography*. An additional advanced feature of the proposed approach that has been somewhat obscured in many previous implementations is *discretizing* of the household weights that is implemented as a theoretically consistent approach.

## Input

PopSyn requires two major inputs:

* Aggregate *controls* that have to be matched at certain level of geography. Most frequently, controls are specified at the level of Traffic Analysis Zones (TAZs) to be consistent with the spatial level of the ABM with respect to location choices. Recently, smaller units like Micro-Analysis Zones (MAZs) or even parcels were used since the corresponding ABMs operated at the finer level of spatial resolution. In addition to that, certain controls can only be specified at a more aggregate level such as district, county, or entire region. For the base year, these controls are normally based on the most recent and updated Census data. For future years, they are either modeled by a land-use / demographic model or provided by the planning group at the MPO. The household or person distributions are frequently generated from the average zonal values by means of percentage curves as part of the PopSyn. (This latter method become redundant in the framework of an advanced PopSyn described in the paper since the average zone values can be incorporated directly).
* Source of disaggregate data as a *sample of households* from PUMS or ACS or (less frequently) from other Household Surveys. This sample plays a dual role in PopSyn. First, it is used to create seed joint multidimensional distributions of households and/or persons that are balanced with the controls. (This latter method become redundant in the framework of an advanced PopSyn described in the paper since the balancing is implemented directly on the list of individual households from the sample). Secondly, this sample is used as a “stock” from which the actual households are further drawn (or matched) to form the synthetics population. Processing the raw data (household and person records) from PUMS or ACS into the format needed for PopSyn is in itself a non-trivial procedure that requires a significant programming effort.

## Output

Output of the PopSyn is a *list of households* in each zone (TAZ or smaller unit) with all household and person attributes available in the sample of households. Attributes that directly correspond to the aggregate controls are called *controlled attributes*. Other attributes are called *uncontrolled attributes*. This list of households represents of a complete, disaggregate population by combining a sample of households in a way as to match key distributions for the entire population.

Strictly speaking, the households themselves are real and taken from the PUMS or other available survey. However, the way how these households are expanded and allocated to zones is analytical, thus the resulted population is called “synthetic”. In statistical terms, the synthetic population is the most plausible population that can be created from the given sample and under the given controls.

## Quality Criteria for Population Synthesis

In most publications the quality of PopSyn has been judged by *matching the controls* [*Bowman, 2004; Bowman & Rousseau, 2006; Abraham et al, 2012*]. While it is a very important consideration in any PopSyn it is not the only one and it has to be considered in combination with another important criterion that is *uniformity of use of the household sample*. There are many synthetic populations that can match the controls equally well. However, in statistical terms, not all of them are equally probable. The rigorous statistical formulation of the most probable state can be expressed as the maximum entropy function calculated over the *a priori* weights of the households in the sample. In this regard, a synthetic population with a less or more uniform expansion factor is most probable than synthetic population with highly differential expansion factors (which means that some households will be use multiple times but some other once will be used only a few times). The entropy function that we want to optimize can be written in the following way:

, Equation

where:

= households in the sample,

= initial (a priori) weights of the households in the sample,

= expansion factors calculated by PopSyn.

This function can be calculated at different levels of geography and these details will be considered in the subsequent sections. If the sample is taken from PUMA than it makes sense to calculate it within each PUMA. Thus, in statistical sense, a population that has a more uniform expansion factor (more exactly, a more uniform ratio ) but somewhat worse match to the controls can be preferred to a population that exhibits a better match to the controls but has extreme expansion factors.

It should be noted, that the controls themselves are not exact and can be inconsistent, especially if they come from different sources. For this reason, rather than having the closeness to controls as the main criterion we suggest a compromise between the sample uniformity and control relaxations following the main maximum entropy paradigm that all sources of information should be used concurrently and the most plausible solution is a compromise rather than an extreme use of one set of criteria. In this regard, the controls can be classified by importance according to the reliability and accuracy of the corresponding source.

In technical terms, this is a welcome feature since a PopSyn that is based on an exact match to controls may not converge in practice (if the controls are not perfectly consistent) and the results would depend on how the balancing procedure was terminated. A model with relaxations guarantees convergence and also allows for a natural combination of the control-matching and uniformity (maximum entropy) criteria in one consistent formulation.

## Controlled and Uncontrolled Variables

Setting of controls is the first important input to the population synthesis. For each control, an appropriate level of geography has to be identified (TAZ, district, entire region). The controls can be set separately for each year depending on data availability, socio-economic, and demographic scenarios.

In practice, there has been a great deal of variation with respect to the set of controls from region to region depending on the data available from the Metropolitan Planning Organization (MPO). MPOs that have a land-use model (SANDAG and MAG) take a full advantage of the data provided by the land-use model that includes, for example, household distribution by 4-5 income groups, worker distribution by occupation, and other control variables that are in general not available. When the base year is close to the year of Census, all available census tables (including some joint distributions) can be effectively used as controls (ARC and MTC experience). In other cases, the population synthesis can be based on a minimal number of control variables that are available, supported by MPO for future years, and frequently used for the 4-step model as well (examples of MORPC and NYMTC).

It should be understood that with any number of controls, the population synthesis procedure will eventually create a synthetic population with all variables needed for the most advanced ABM. The rest of variables will be added as *uncontrolled*. However, there is a principal advantage in using controlled variables over uncontrolled in two respects:

* Controlled variables ensure *reasonability of the synthetic population* while uncontrolled variables may exhibit unpredictable effects (this in part can be addressed in the validation process, but can never be fully eliminated). One of the known examples that manifested itself many times in the past is the inherent inability of PopSyn to generate and locate the population of university students living in rent apartments properly if no specific person-type controls are introduced. Standard controls like household distribution by size, number of workers, and income do not automatically ensure a right number of university students in TAZ. Since university students can represent an essential transit market in some corridors as was the case with large universities in Phoenix, AZ (ASU) and Columbus, OH (OSU) that had to be modeled as part of MAG and MORPC ABMs, it is important to make a special effort to address this particular segment. This is another example where adding person-level controls is essential.
* Only controlled variables can be use as policy levers to create *specific socio-economic and demographic scenarios* for future years. These scenarios may include different levels of income growth, labor participation, aging, diminishing household size, shifts in ethnicity or education levels, etc. In each case, where any socio-economic or demographic trend is involved, the corresponding control has to be introduced. It should be understood that PopSyn by itself is not a land-use model that can predict major trends and tendencies. PopSyn is a tool that creates the most statistically probable synthetic population under certain input constraints that have to be provided externally. This constraining is based on the controlled variables.

We can recommend a list of the most important variables to consider in . These variables are very important determinants of travel behavior and are used in many components of an ABM. If any of these variables can be provided as input we suggest including it as a control. Otherwise these variables will be synthesized as uncontrolled. Some variables can be introduces as household-level controls only. Some variables can be introduced as person-level controls only. Some other variables like household size and number of workers can be introduced both ways. Typically, these variables have to be provided for each TAZ in addition to the number of households.

Table : Main Important Variables to Consider as Controls

| **Variable** | **Household-level** | **Person-level** |
| --- | --- | --- |
| Household size | Household distribution by size category (1,2,3,4,5+) | Average household size by providing total population |
| Number of workers | Household distribution by number of workers (0,1,2,3+) | Average number of workers per household by providing total labor force by place of residence |
| Household income | Household distribution by income groups defined in terms of absolute thresholds or percentiles; alternatively average household income per TAZ can be used directly |  |
| Housing type | Household distribution by such categories as single-family detached house and apartment in multi-family house |  |
| Person age | Household distribution by age of the household head (only a proxy for age of the other household members) | Population brackets of which the most important for ABM are 0-5, 6-18, 19-35, 36-64, 64+) |
| Person type used in advanced ABMs (it is partially correlated with age; any of controls for some of these variables would be useful) |  | 1=Full time worker  2=Part time worker  3=University student  4=Adult non-worker under 65  5=retiree  6=driving age school child  7=pre-driving age school child  8=preschool child |
| Worker distribution by occupation (may improve ABM significantly but is only available if and advanced land-use model is in place) |  | Labor force distribution by occupation categories provided by the land-use model |

It should be mentioned that for the most commonly used household variables (household size, number of workers, and income group) the household distribution is frequently created by applying household distribution curves developed from the Census data. These curves represent percentage of households in each category (for example 1-person, 2-person, etc) as the function of the average value per household (for example, household size) in each zone. This step is frequently included in PopSyn as a routine part of data processing. This step however, is necessary only of the PopSyn is limited to household-level controls and matrix-based balancing. It will be shown below that with more flexible balancing procedures this step can be eliminated and the average zonal values can be incorporated directly.

The mathematical essence of our approach is completely generic and flexible with respect to the number of controls and their categorization.

# Existing Methods

## 3-Step Procedures Based on IPF of Joint Household Distributions

In recent years, several population synthesizers have been developed, primarily focusing on Iterative Proportional Fitting (IPF) techniques [*Muller & Axhausen, 2010*]. A number of advances have been made in the use of this class of algorithms, including enabling control at the household and person level [*Guo & Bhat, 2007*] and automatically dealing with the zero-cell problem [*Auld et al, 2010*].

Most population synthesizers currently in use are based in some way upon the method developed by Beckman et al. (1996) for use in the TRANSIMS model, with some variations in the types of input and how certain routines are carried out. The TRANSIMS procedure matches exact large area multi-dimensional distributions of select variables from the Census Public Use Microdata Sample (PUMS) files to small area marginal distributions from the Census Summary files to estimate the multi-dimensional distributions for the small areas (Beckman et al. 1996). The procedure in fact works for any situation where an exact multi-dimensional distribution and marginal totals are known, and is not limited to using only the Census data files. A process of this type is necessary for all population synthesizers due to confidentiality and data aggregation issues which preclude most agencies from releasing multi-dimensional distributions for the small areas typically needed for most models. Other models sometimes use different data sources or even estimated data as is the case in the MORPC model (PB Consult 2003). In general, the population synthesis is typically carried out in two stages: fitting the multi-dimensional distribution and selecting the households.

The first stage in the synthesizing procedure is developing the multi-dimensional distribution matrix (Beckman et al. 1996). This stage makes use of the technique of iterative proportional fitting or IPF which has long been used in many fields (Deming and Stephan, 1940). This procedure assumes that the correlation structure amongst the variables is similar between the large area and the smaller areas of which it is constituted. In the IPF procedure, an initial seed distribution is used and fit to known marginal totals. The difference between the current total and the marginal total for each category of the current variable is then calculated and the cells of that category updated accordingly. This process continues for each variable until the current totals and the known marginal total match to some level of tolerance, producing a distribution which matches the control marginal totals and is as similar to the initial distribution as possible. For a more detailed overview of the IPF procedure see the paper by Beckman et al. (1996) and technical report by Hobeika (2005).

In most population synthesizers, the multi-dimensional distribution from the sub-region is created then fit to the control variable marginal totals through the IPF procedure and used as a seed distribution for the zones where the marginal totals are known (Beckman et al. 1996). The IPF procedure is then used again to fit each zone distribution to the marginal totals, and often to fit all of the zone distributions to the initial sub-regional distribution. After this process, the required number of households of each type represented in the distribution matrix is known for all of the zones in the sub-region.

This first stage leads to one frequently observed problem in the development of population synthesis routines, which is the occurrence of false zero-cells in the distribution matrix (Guo and Bhat 2007). This occurs because the distribution matrix is typically created from a data source which is only a sample of the actual population, such as the PUMS data which represents a 5% sample. This means that for a control category for which households are known to actually exist in an area, but are infrequent, there is a likelihood of a household of that type not being selected in the 5% sample. This causes a zero value to be entered for the cell in the creation of the initial distribution for the region, which will never be changed during the IPF procedure. Several methods have been proposed for dealing with the false-zero cell issue, including “tweaking” the entries in the initial distribution to be an arbitrary small number rather than zero, setting a maximum iteration counter on the IPF procedure and ignoring the effects of non-convergence (Beckman et al. 1996), or adjusting the number of categories within the control variables to eliminate the highly infrequent categories and thereby reducing the likelihood of false zero-cells (Guo and Bhat 2007). Most current population synthesizers implement some combination of all of these strategies. However, each strategy has individual drawbacks. Tweaking the initial cell values was shown by Beckman (1996) to produce little benefit in model accuracy and significantly bias the resulting distribution. Tweaking also add to the complexity of the household selection stage by creating non-zero cell values for unobserved households in the micro-level data file. For these reasons, the tweaking procedure was not recommended (Beckman et al. 1996). Aggregating the control variable categories is another option, and is usually done implicitly by the modeler in the creation of the input data files. This approach can also reduce the accuracy of the model by reducing the amount of discriminating information available.

The second stage in the population synthesis creates the actual synthesized individuals or households. There are several ways in which this is accomplished depending on certain other model characteristics. The simplest method is to calculate a selection probability for choosing each households from the actual household microdata file for the large area, based on the need for households of that type determined from the final calculated multi-dimensional distribution table for the small area and the number of households fitting the demographic type in the household list for the area (Bowman 2004). The households are then selected randomly according to their selection probabilities for each zone as is done for the TRANSIMS model when the PUMS data is used as input (Beckman et al. 1996). However, it has been observed that selection procedures of this type do not force the generated households due to the non-integer values in the distribution matrix. So the generated marginal totals tend to vary around the known totals, which is not a desirable trait for the control variables of the model (Bowman 2004).

Variations on this basic population synthesis formulation by a variety of researchers have been attempted in response to this issue (Guo and Bhat 2007; PB Consult 2003). Usually, the fitted multi-dimensional distributions which are calculated in the first stage are converted to integer values for use in the selection process, after which the households are selected as above. This “integerization” procedure ensures that the control variable marginals can be matched exactly. Unfortunately, the integerization procedure usually results in a biased distribution or inconsistencies in the total number of households to be generated (Bowman 2004). Therefore, further research in the area of selection procedures is most likely warranted.

[*Auld et al, 2010*] describes development of a fully-customizable population synthesis program which has been designed to be used for microsimulation of travel demand in any geographic area. The design addresses several important issues that have been raised regarding other population synthesis methods. The false zero-cell problem, which been a major concern in the development of most population synthesis procedures, is addressed with a routine that allows for the aggregation of control variable categories during execution at the region-level based on a user-controlled aggregation threshold parameter. This reduces the areas where false zero-cells are most likely to occur and also greatly eases the task of data preparation. As part of the study, validation methods were used to compare the new procedure against other procedures and known test variable distributions. The use of these new routines and validation methods will allow for the development of more realistic synthetic populations.

The traditional population-synthesis procedure involves two steps (Beckman et al. 1996). First, a joint multi-way distribution of all attributes of interest is generated using the Iterative Proportional Fitting (IPF) procedure (Deming et al., 1940; Ireland and Kullback, 1968). Next, individual household records are drawn (with replacement) from the seed dataset using Monte Carlo simulation so as to satisfy the joint multi-way distributions. At the end of the second step, a list of synthesized households is generated.

Thus, the IPF is at the core of the traditional population-synthesis procedure. This procedure preserves the odds ratios among the different socio-economic attributes from from the seed data in generating the joint multi-way distribution. Further, when the generated multi-way distribution is appropriately aggregated, the results match the marginal distributions provided by the control tables (the extent of “matching” depends on the tolerance used). Thus, the IPF procedure fuses information from the seed data and the control tables to generate the synthetic population of interest. Ireland and Kullback (1968) also show that IPF minimizes the discrimination information (sometimes called negative entropy) between the two distributions subject to the controlled marginal distributions.

Since the work of Beckman et al. (1996), the IPF-based population synthesis methodology has been applied to support travel-demand modeling in several areas such as Portland Metro, San Francisco, New York, Columbus, Atlanta, Sacramento, Bay Area, and Denver [PB Consult, 2003; Bowman, 2004; Bowman & Rousseau, 2006, 2008].

Despite its theoretical basis and statistical properties, the IPF-based synthesis procedures have certain practical limitations. First, when the number of controlled attributes (or equivalently the total number of household types) is large, it is quite likely that several household types do not exist in the seed data. In this case, the odds ratio and discrimination information cannot be defined (they are infinite) and this phenomenon is called the “zero-cell problem” (Beckman et al. 1996).This limits the amount of information that can be controlled for in IPF methods as the zero-cell categories may have to be aggregated with adjacent non-zero value cells (see for example, Auld et al., 2008). Alternately, one may have to start with a small non-zero value for the zero cells (Beckman et al., 1996). From an implementation standpoint, the number of cells in the joint distribution grows drastically with increased number of control tables thereby impacting the computational resources (memory) needed.

A second downside of the IPF procedure is that it requires all the control tables to belong to the same “universe”. Therefore, it is not possible to apply this procedure directly to synthesize populations by simultaneously controlling for household- and person- level attributes. However, for achieving more accurate populations, it would be desirable to control for a wide range of attributes.

In the last few years, there have been several efforts to modify the traditional approach to deal with multi-level controls. Arentze et al. (2007) adopt a two-step procedure each involving IPFs. The first IPF aggregates person-level attributes (e.g., age and work status) to the household level. This procedure results in the creation of several household types that are defined based on both household- and person- attributes (for example, a household type is defined as a two person household with one male and one female). Now, these new household-level controls (which also include person-level characteristics) are used to synthesize the population using a second IPF procedure. The next three methods (Guo and Bhat, 2007; Ye et al., 2007; and Auld and Mohammadian, 2010) to be discussed all begin by generating multi-dimensional distributions for household- and person-level attributes independently using IPF techniques. In the Guo and Bhat (2007) procedure, households are then drawn from the seed data based on the household-level joint distribution. However, these are checked against the person-level distributions to ensure that person-level constraints are not violated, subject to certain tolerance prior to accepting the household into the synthetic population of the census tract. Ye et al. (2009) calculate “weights” (using an iterative procedure) for each household in the seed data such that the weighted population matches both the household- and person- joint-distributions. Similarly, Auld and Mohammadian (2010) also determine the “desirability” of a household drawn based on the extent to which the household satisfies both the joint distributions. In a more recent study by Lee and Fu (2011), a cross-entropy formulation is adopted in which the entropy is defined based on both household- and person- level characteristics. In summary, it is interesting to note that all methods developed thus far involve multiple applications of the IPF of IPF-like procedures (entropy models are similar to IPF) to deal with the issue of multi-level controls. Consequently, these methods still have to address the zero-cell problem. The reader is referred to Ma (2011) for a detailed discussion on population synthesis methods.

The discussion thus far has focused only on the generation of the joint multi-way tables. The cell values of these joint-distribution tables are non integers, and hence, they do not directly provide a list of households and persons (the fundamental output that is required). Past studies usually adopt the Monte Carol simulation for drawing individual households from the seed data based on the joint probability distribution to ultimately form the synthetic population. Since the cell values of the joint distribution effectively get rounded off after this simulation procedure, the accuracy of the joint-distribution table is now somewhat compromised. As an alternative, it is theoretically possible to model the synthesis process using integer programming / combinatorial optimization techniques (Ryan et al., 2010; Williamson et al., 1998; Voas and Williamson, 2000). While this procedure is guaranteed to generate integer number of households, solving IP problems with a very large number of decision variables (seed data) and constraints (control table cell values) is not straightforward.

The existing population synthesis procedures are largely empirical and include the following three general stages:

1. Build multidimensional household distribution in each TAZ based on the zonal controls and seed household distribution (by means of a multidimensional matrix balancing).
2. Create a list of households with controlled variables in each TAZ (by discretizing the multidimensional distribution).
3. Randomly draw households from the list (PUMS or household survey) to add uncontrolled variables.

Three major algorithmic steps implemented for each zone:

* *Balancing* of the seed distributions of households and/or persons for each zone to match the controls. This is generally implemented using some modification of the IPF (Iterative Proportional Fitting) method and results in a fractional number of households of each type.
* *Discretizing* of the balanced distributions to obtain an integer number of households of each type. This is not a trivial rounding procedure since simple uncontrolled rounding may result in significant discrepancies with the controls matched at the previous step.
* *Drawing* of households from the sample by controlled variables to form the synthetic population. In general, each synthetic household is matched to an actual household from the sample that has the same controlled variables. Then, the actual household provides all other (uncontrolled) variables.

Some drawbacks of the existing approaches:

* Lack of a single and coherent theoretical formulation; as a consequence two different synthetic populations (procedures) are difficult to compare.
* Difficulties in handling both household and person characteristics in the same procedure (stages 1 and 2).
* Difficulties in handling multiple geographies beyond the base zonal level (usually, TAZ). These difficulties relate to both imposing aggregate controls at the district, county, or region levels as well as allocation of households to smaller spatial units (MAZ or parcel).
* Ignoring discretizing problem and using simplified “bucket” rounding technique that can result in violation of controls.

Already the first SPGs developed for ABM in San Francisco (SFCTA), New York (NYMTC), Columbus (MORPC), and Denver (DRCOG) in the period 2000-2006 had all these features in place and in general produced acceptable results. However, several problems and limitations had become clear and required some modifications and additional advanced features introduced in later ABM development projects, many of them are still at different stages of development. The following recent advances are relevant for the BMC SPG:

* While the balancing procedures applied at Step 1 have a firm theoretical foundation in entropy-maximaztion principle (of which IPF is one of particular methods) the subsequent Steps 2 and 3 (Discretizing and Drawing) are largely *empirical*. As the result, the entire population synthesis procedure becomes empirical with no single over-arching theoretical framework. If it is strictly implemented in a standard way described above there is no guarantee of a unique solution as well no clear method to compare different solutions. In this regard, we propose several new features below that resolve many of the problems and bring the entire SPG structure closer to a single and consistent theoretical framework.
* All SPGs developed and applied in practice so far have been able to handle *either households-level controls or person level controls but not both*. This was a consequence of a limited *matrix-based* structure of the balancing procedure that requires only one unit (household or person) to be used to construct a full multidimensional distribution for balancing. However for an advanced ABM it is equally important to generate a reasonable and scenario-sensitive population with respect to household-level controls (like household distribution by size or income group) and person-level controls (like right proportions between workers, students, non-working adults, and children). Some simplified ways to obviate this limitation were proposed. For example, SPG developed for the Dallas-FW ABM (NCTCOG) by the University of Texas at Austin in 2006-2009 was based on a household-level balancing but additional person-level controls where introduced in the drawing procedure. This method cannot guarantee an exact match for person-level controls but ensures some approximate match. However, the real breakthrough approach in this regard was proposed and applied for Los-Angeles (SCAG) and Phoenix (MAG) ABMs in 2009-2010 that is based on a more general *list-balancing* procedure. This procedure is also proposed for the BMC SPG and discussed below in more detail.
* Practical experience has shown that the *controls themselves might be contradictory* (i.e. no solution exists that would satisfy all of them simultaneously). In fact, it is difficult to even check the controls for internal consistency especially if both household-level and person-level controls are applied. The inconsistency manifests itself in that the balancing procedure does not converge and produces different results depending on the arbitrary stoppage criteria. Some attempts to introduce meta-balancing procedures (i.e. balancing between the controls themselves) were made in San-Diego (SANDAG) ABM. However, they can only screen some obvious inconsistencies like different total number of households across different household distributions. The most general and theoretically solution for this problem, in our view, is to apply a relaxed balancing procedure proposed originally by the Arizona State University and subsequently incorporated in the Los-Angeles (SCAG) and Phoenix (MAG) ABMs. This procedure guarantees a unique solution and matches the controls exactly if they are consistent. If the controls are not consistent the procedure screens the inconsistencies and finds a compormize solution based on the relative importance of ach control. This procedure is proposed for the BMC SPG and discussed below in more detail.
* Drawing households from the sample has an inherent *Monte-Carlo variability*, i.e. can produce different results if implemented with exactly the same inputs by just changing the sequence of random numbers used. It may also result in a non-uniform use of the stock of households when some households are over-used and some other ones are under-used. Some improvements to the drawing procedure were made to the Atlanta (ARC) SPG in 2005 that were subsequently used for Sacramento (SACOG), Denver (DRCOG), San Francisco Bay Area (MTC), and San Diego (SANDAG) ABMs. In this regard, the innovative list-discretizing procedure integrated with the list-balancing procedure developed by PB holds promise since it eliminates the Monte-Carlo variability completely. It is currently being tested for the Phoenix (MAG) SPG. This innovative component is discussed below in more detail and can be consider as an optional feature for the BMC SPG.
* The conventional SPG procedure is designed to create a synthetic population for each zone separately and requires *all controls to be set for each zone* (TAZ or smaller unit). In practice, however, there might be situations where certain controls cannot be established for each zone exactly; however they can be introduced at more aggregate level of geography or at least at the entire-region level. This frequently happens when some general tendencies are needed to be accounted for long-term forecasting, for example diminishing household size or aging population. The *table balancing* procedure provides this option. This procedure was first introduced in the Atlanta (ARC) ABM in 2006 and subsequently adopted in Sacramento (SACOG), San-Francisco Bay Area (MTC), and San Diego (SANDAG) ABMs. The Table Balancing procedure is a particular case of List Balancing procedure and is also discussed below.

In the proposed design of the BMC SPG, we plan to include all advanced features described above. In particular, we will apply a generalized List Balancing procedure with relaxed constraints that allows for an effective handling of any number of household-level and person-level controls simultaneously.

On the **implementation side**, all PopSyns so far have been coded in advanced software languages such as Java or C to make use of their functionality in manipulation large data arrays. The major reason for using custom software for PopSyn was the fact that this component is not directly related to the transportation networks, trip tables, or level-of-service skim matrices where transportation planning package is essential. PopSyn from the programming perspective represents a set of intensive data processing and calculation steps with large arrays of data items that include multi-dimensional household and person distributions and lists of households. For example, it is usual for a balancing procedure with 20 controls (that is a common case) to take 1-2 seconds to converge for a single TAZ. This means that for a regional system with 2,000 TAZs the balancing step alone can take up to 1 hour of runtime. This runtime, however, can be reduced significantly by parallel processing (multi-threading and distributing) if a powerful computer cluster is applied. The PopSyn structure lends itself for parallel processing because most of the calculations are implemented separately for each TAZ.

## Combinatorial Optimization and Simulated Annealing

A second class of approaches, combinatorial optimization, has also been examined in the literature. Combinatorial optimization is a simpler and perhaps more direct technique than IPF. Broadly, a trial population is created from the disaggregate sample data, and then the overall level of fit is measured across all marginal targets. Units from the trial population are swapped with units chosen from the disaggregate samples, and if the measure of fit improves, the swap is made. Approaches such as simulated annealing [*Williamson et al, 1998*] can be used to modify this procedure to permit swaps that are suboptimal locally, but may help avoid becoming trapped in a local minimum. Combinatorial optimization techniques have been compared with the IPF-based 64 sample reconstruction techniques twice; [*Huang and Williamson, 2001*] described combinatorial optimization as “greatly superior”, and more recently, [*Ryan et al, 2007*] found it to produce more accurate results given similar input data.

[*Abraham et al, 2012*] describes a population synthesizer that uses a combinatorial optimization approach, with a variety of features that make it useful in working with multiple sources of data on difficult real-world problems. Two large-scale practical applications are then described, to illustrate some of the benefits and issues with this synthesizer. This paper describes an approach using a combinatorial optimization algorithm; a versatile technique capable of simultaneously matching targets at multiple agent levels, such as properties of households as well as for individuals within the households. The software also supports simultaneous targets defined for multiple geographical levels (such as zones, counties and states). The use of the software is demonstrated in two applications; the synthesis of the 2000 population of California (comprising some 33.9 million 39 individuals in 11.5 million households), and the synthesis of the 2008 employment in Oregon and surrounding areas (comprising 3.5 million workers). The algorithm is acceptably fast and matches the targets with a high degree of accuracy.

In linked publications [*Srinivasan et al, 2008, and Ma, 2011*] explain the Iterative Proportional Fitting (IPF) procedure is not conducive for multi-level controls. This article presents an alternate methodology called the Fitness Based Synthesis (FBS) approach that directly generates a list of households to match several multi-level controls without the need for determining a joint multi-way distribution. The application and validation results demonstrate both the feasibility of the approach and the accuracy gains relative to the IPF and methods using fewer control tables. This article also presents a comprehensive validation of the synthetic populations against the true populations and thereby demonstrates the ability of the FBS method to generate the multi-dimensional correlations among the attributes. The number of iterations to terminations is found to be between 1 and 3 times the number of households to be synthesized. In sum, the FBS is a fast, efficient, and scalable methodology that is easy to implement and as such is a valuable tool for generating the detailed socio-economic characteristics need for applying disaggregate travel-demand forecasting models.

In light of the above discussions, the intent of this study is to present an alternate, single-step mechanism for synthesizing populations that can efficiently deal with multi-level controls. We call this the fitness-based synthesis (FBS) approach as households are generated to match the control targets using a fitness function. The process of first developing a joint multi-way distribution is eliminated and the synthesis procedure directly generates a list of households and persons that satisfy several control targets. Further, the proposed method is fast, memory efficient, and scalable (can be applied to any number of control tables).

The FBS procedure broadly involves selecting a set of households (with replacement) from the seed data (PUMS) in such a way that the tract-level controls are satisfied. This section presents an outline of this algorithm identifying its salient features. The reader is referred to Ma (2011) for a simple numerical illustration of this procedure.

The FBS procedure starts with an initial set of households. This can either be a NULL set (no households) or a random sample from the seed data. The population of the census tract is synthesized (i.e., the initial list adjusted) in an iterative fashion. In each iteration, one household (with the details of all its household members) is either added or removed from the current list. Count tables are used to track the number of households of each type (defined in terms of the control attributes) that have already been “selected” until that point. There are as many count tables as there are control tables. Together, the count- and control- tables enable us to assess whether the control targets have been achieved (i.e., if the value in a cell of the count table is less than the corresponding value in the control table, then the target has not been achieved for that cell).

At the core of this procedure is a systematic procedure to determine if a household is to be added or removed at any iteration. For this purpose, two “fitness” values ( and ) are calculated for each household (see equations (1) and (2)). Conceptually, the “type I” fitness value is the reduced sum of squared error in matching the control tables if the household being selected for adding at the iteration. The “type II” fitness value is the corresponding error if the household is removed at the iteration.

[*Harland et al, 2012*] analyze several established methodologies for generating synthetic populations in the context of application of Agent-Based Modeling Paradigm in geography. These include deterministic reweighting, conditional probability (Monte Carlo simulation) and simulated annealing. However, each of these approaches is limited by, for example, the level of geography to which it can be applied, or number of characteristics of the real population that can be replicated. The research examines and critiques the performance of each of these methods over varying spatial scales. Results show that the most consistent and accurate populations generated over all the spatial scales are produced from the simulated annealing algorithm. The relative merits and limitations of each method are evaluated in the discussion.

"Simulated annealing is stochastic computational technique derived from statistical mechanics for finding near globally minimum-cost solutions to large optimisation problems" (Davies 1987). In this application, we take a sample and assume that this is large relative to the population of a small area which is to be modelled. The problem is to extract a subset from the sample which provides the best possible match to the small area population. In other words, the weights on the individual members of the sample are either one (if the individual is selected) or zero (if the individual is excluded). The essence of the procedure is to start by creating a population as a random extract from the sample file, and by aggregating for the various constraints, the goodness of fit of the population to the constraining tables can be evaluated. From this population, we now select an individual member at random, and we consider replacing it with another individual which is also selected at random from the sample population. The aggregation and goodness of fit evaluation is repeated and if the fit is improved then the new individual replaces the old.

## First Examples of List Balancing

[*Pritchard & Miller, 2009*] suggested a list based data structure to effectively deal with such sparse matrices.

[*Sun et al, 2011*] presented an example of an advanced Table Balancing procedure that in essence is very close to the List Balancing. The presented PopSyn incorporates a large number of household-level and person-level controls. Household level controls include Household size (1,2,3,4+), Household income (5 categories), Number of workers per household (0, 1, 2, 3+), Number of children in household (0, 1+), Dwelling unit type (3 categories), and Group quarter status (4 categories). Person level controls include Age (7 categories), Gender (2 categories), and Race (8 categories). However, this PopSyn is limited in terms of setting these controls at different levels of geography. The presented example can only handle one particular level of geography – TAZ, i.e. the procedure is applied to each TAZ separately. Also, the results of Table Balancing have to be discretized and eventually households have to be drawn from PUMS or ACS samples. These two steps are implemented with a Table Balancer in a conventional way as it was with the Matrix Balancing IPF procedure.

[*Ye, et al, 2009*] presented a heuristic iterative approach, dubbed the Iterative Proportional Updating (IPU) algorithm, for generating a synthetic population while simultaneously matching both household-level and person-level joint distributions of control variables of interest. In traditional population synthesis procedures, only household-level joint distributions are matched using the standard iterative proportional fitting (IPF) procedure, and entire households are randomly drawn according to the IPF-generated household-level joint distribution. Little regard is given to matching person-level distributions (all of the persons in the drawn households become part of the synthetic population). This results in a synthetic population wherein the household distributions are matched, but the person-level distributions are not likely to be matched well simply because all households in a certain cell of the joint distribution receive the same weight. In contrast to that approach, the IPU algorithm iteratively adjusts and reallocates weights among households so that person-level distributions are matched as closely as possible without compromising on the fit to household-level distributions. The paper presents the algorithm in detail, illustrates the algorithm using a small example, and then offers real-world case study results for small geographies (blockgroups) in the Maricopa County region of Arizona. While this paper essentially presents a List Balancing procedures its implementation was based on heuristics and was not derived from a rigorous framework.

A good example of a List Balancing procedure was described by [*Bar-Gera et al, 2011*]. Although the paper does not deal with population synthesis per se but rather focuses on expansion of a Household Survey, the method itself is very close to the List Balancing adopted in the current research. A challenge often faced by transportation professionals is to accurately expand the survey households to represent the population. The problem becomes even more complicated when dealing with household travel surveys, because the goal is to find household weights such that the distributions of the characteristics in the weighted sample not only match given distributions of households but also those of persons. This paper presents an Entropy Maximization methodology to estimate household survey weights to match the exogenously given distributions of the population, including both households and persons. The paper also presents a Relaxed Formulation to deal with cases when constraints are not feasible and convergence is not achieved. The methodology is applied to a large geography - Maricopa County region, Arizona, and a small geography –blockgroup, and estimation results are presented. Estimation results show that the Strict Formulation can be used to estimate the weights when constraints imposed by distributions of population characteristics are feasible. Relaxed formulation can be used to estimate weights when the constraints are infeasible such that distributions of the population characteristics are satisfied to within reasonable limits.

# Proposed Method

## Requirements for Population Synthesis

The innovative components address the requirements to match *both household-level and person-level attributes* in the model area as well as the possibility to set the corresponding *controls at different levels of geography.* To the extent possible we tried to eliminate empirical components that result in ambiguity in the generated population and derive the PopSyn procedure from the rigorous mathematical framework. This framework accounts for the constraints but it also states explicitly what is the best synthetic population (most probable in statistical terms) out of many possible populations within these constraints.

## Outline of Main Steps

The current memo describes an alternative approach that has a firm theoretical basis and can handle any mix of person and household characteristics. This approach includes the following steps:

* Create a sample of households in each TAZ (all households from the correspondent PUMA can be used in a simplified case).
* Balance the individual household weights to ensure the controlled totals across all person and household dimensions.
* Create a list of households by discretizing the individual weights.

The advantage of working with the list of households compared to a multi-way distribution is that both person and household variables can be incorporated. It is shown that in case if only household or person attributes are controlled, the proposed procedure yields exactly the same multidimensional distribution as the conventional matrix balancing. Also, the elimination of the drawing procedure allows for a theoretically closed formulation with no unnecessary empirical components.

# List Balancing for Single Zone

## Basic Formulation with Fixed Constraints

In the heart of the advanced SPG methods are innovative balancing procedures that have been applied in tested in several recent ABMs. It is important to distinguish between three different types of balancing procedures as summarized in Table 2.

Table : Balancing Procedures

| **Type** | **Controls** | **A priori weights (seeds)** | **Contribution coefficients** |
| --- | --- | --- | --- |
| Multidimensional matrix balancing | Row/column totals (marginal one-dimensional distributions) | Initial matrix (multidimensional distribution) | Incidence of cells with rows and columns (0,1) |
| Balancing of table of categories | Column totals (marginal one-dimensional or multidimensional distributions; single category totals) | Initial weight for category (row) | Row/column incidence (0,1) |
| Balancing of list of individual records | Column totals (marginal one-dimensional or multi-dimensional distributions; single category totals) | Initial individual record weight (row) | Row/column coefficient (any positive value) |

All three balancing methods are derived from the general maximum entropy approach that is the most common statistical methods for adjusting multidimensional distribution to meet marginal controls. It can be shown that each subsequent method includes the previous one as a particular case and guarantees the same result. Thus any matrix balancing problem can be done through the table balancing and any table balancing can be done through the list balancing. However, the opposite is not true. *Matrix balancing* can only handle either household-level or person-level controls and additionally required that all controls to be set as complete marginal distributions over the entire set of multidimensional cells. Matrix balancing was the core procedure in early SPG designs that were applied in SFCTA, NYMTC, MORPC, and DRCOG ABMs.

*Table balancing* is more flexible in this regard. In the table, each cell is essentially considered as a row and there is no constraint on the nature and completeness of controls. Table balancing was first applied in the ARC ABM and since has become the most common method applied in other ABMs (MTC and SANDAG). However, table balancing still has a limitation of the row/column incidence coefficient to be equal to 0 (the category is excluded from the control) or 1 (the category is included in the control). This limitation is essential for the table balancing algorithm of IPF type. However, with this limitation, it is practically impossible to address both household-level and person-level controls. In order to “squeeze” a person-level control (say, age distribution by 5 brackets) into the general household-level balancing table one should explicitly consider all possible household compositions in terms of number of people of each age category. Even if we assume that not more than 4 persons of each age bracket can be in the same household, this would results in (4+1)5=3,125 household categories by age alone that is infeasible to handle taking into account the other household distributions that have to be combined with the age distribution in a Cartesian manner.

*List balancing* is a further generalization of the table balancing that resolves the limitations discussed previously. In particularly, it allows for incorporation of any number of household-level and person-level controls and its does not require transformation of the person-level controls to household-level controls in a Cartesian way. This procedure was first proposed an applied by ASU and PB for the MAG ABM and then was adopted for the SCAG ABM.

Since in most cases the procedure is applied for each TAZ separately we formulate the model for a single TAZ for simplicity of explanation. Introduce the following notation:

 = household and person controls,

 = seed set of households in the PUMA (or any other sample),

 = a priori weighs assigned in the PUMA (or any other sample),

= zonal controls,

= coefficients of contribution of household to each control.

The principal flexibility of the proposed procedure is that the contribution coefficients can take any non-negative value while in the conventional procedure the contribution coefficients are implied to be Boolean incidence indicators (belong or not belong). An example is shown in Table 3 below for controls specified by household size and person age brackets.

Table : Example of List Balancing

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| HH ID | HH size | | | | Person age | | | | HH initial weight |
| 1 | 2 | 3 | 4+ | 0-15 | 16-35 | 36-64 | 65+ |
|  |  |  |  |  |  |  |  |  |
|  | 1 |  |  |  |  |  |  | 1 | 20 |
|  |  | 1 |  |  | 1 | 1 |  |  | 20 |
|  |  |  | 1 |  |  | 1 | 2 |  | 20 |
|  |  |  |  | 1 |  | 2 | 2 |  | 20 |
|  |  |  |  | 1 | 1 | 3 | 2 |  | 20 |
| …. |  |  |  |  |  |  |  |  | … |
| Control | 100 | 200 | 250 | 300 | 400 | 400 | 650 | 250 |  |

The first household has one person of age 65+. The second household has two persons: one of age 0-15 and another one of age 16-35. The third household has three persons: one of age 16-35 and another two of age 36-64. The fourth household has four persons: two of age 16-35 and anther two of age 36-64. The fifth household has size persons: one person of age 0-15, three persons of age 16-35, and two persons of age 36-64. Household weights are all set to 20 that is a common case with 5% PUMS but in a general case the initial weight can be differentiated to take into account the sample structure.

The balancing problem can be written as a convex entropy-maximization problem in the following way:

, Equation

Subject to constraints:

, Equation

, Equation

where represents dual variables that give rise to balancing factors.

The objective function expresses the principle of using all households uniformly (proportionally to the assigned a priori weight). The constraints ensure matching the controls. This is a convex mathematical problem with linear constraints that can be solved by forming the Lagrangian and equating the partial derivatives to zero. The Lagrangian function can be written in the following way:

. Equation

We calculate partial derivatives and equate them to zero:

. Equation

By collect terms with constants on the right hand side and exponentiating both sides we obtain the following solution:

, Equation

where represents balancing factors that have to be calculated.

Note that the balancing factors correspond to the controls, not to households. For each household, the weight is calculated as a product of the initial weight by the relevant balancing factors exponentiated according to the participation coefficient. A zero participation coefficient automatically results in a balancing factor reset to 1 that does not affect the household weight.

The presented general formulation has been further generalized to account for relaxation of constraints that avoids a non-convergence problem is the controls are not consistent within themselves. An algorithm based on the gradient optimization of the objective function (Equation 1) has been implemented that is much more efficient that an IPF based on the solution (Equation 6).

Additional enhancement that has been recently added to the MAG SPG and is currently being tested is an innovative disretizing method that is applied on the household weights and integrated with the balancing procedure. This method creates a list of households directly from the sample and eliminates the need in subsequent drawing of households.

Since the procedure is applied for each TAZ separately we formulate the model for a single TAZ. Introduce the following notation:

 = household and person controls,

 = seed set of households in the PUMA (or any other sample),

 = a priori weighs assigned in the PUMA (or any other sample),

= zonal controls,

= coefficients of contribution of household to each control.

The principal flexibility of the proposed procedure is that the contribution coefficients can take any non-negative value while in the conventional procedure the contribution coefficients are implied to be Boolean incidence indicators (belong or not belong). An example is shown in Table 1 below for controls specified by household size and person age brackets.

Table : Controls and contribution coefficients

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| HH ID | HH size | | | | Person age | | | | HH initial weight |
| 1 | 2 | 3 | 4+ | 0-15 | 16-35 | 36-64 | 65+ |
|  |  |  |  |  |  |  |  |  |
|  | 1 |  |  |  |  |  |  | 1 | 20 |
|  |  | 1 |  |  | 1 | 1 |  |  | 20 |
|  |  |  | 1 |  |  | 1 | 2 |  | 20 |
|  |  |  |  | 1 |  | 2 | 2 |  | 20 |
|  |  |  |  | 1 | 1 | 3 | 2 |  | 20 |
| …. |  |  |  |  |  |  |  |  | … |
| Control | 100 | 200 | 250 | 300 | 400 | 400 | 650 | 250 |  |

The first household has one person of age 65+. The second household has two persons: one of age 0-15 and another one of age 16-35. The third household has three persons: one of age 16-35 and another two of age 36-64. The fourth household has four persons: two of age 16-35 and anther two of age 36-64. The fifth household has size persons: one person of age 0-15, three persons of age 16-35, and two persons of age 36-64.

The balancing problem can be written as a convex mathematical program of the entropy-maximization type in the following way:

, Equation

Subject to constraints:

, Equation

, Equation

where represents dual variables that give rise to balancing factors.

The objective function expresses the principle of using all households uniformly (proportionally to the assigned a priori weight). The constraints ensure matching the controls.

|  |  |
| --- | --- |
|  | Equation |
|  | Equation |
|  | Equation |

Can be solved either by IPF or (more efficiently) by the Newton-Raphson method.  if all  . However, for some zones and controls . This case is treated as special in the Newton-Raphson method (all HHs that contribute to a zero control with a non-zero coefficient have to be zeroed out). IPF does not require a special treatment for zero constraints.

By forming the Lagrangian and equating the derivatives to zero we obtain the following solution:

, Equation

where represent balancing factors that have to be calculated. Note that the balancing factors correspond to the controls, not to households. For each household, the weight is calculated as a product of the initial weight by the relevant balancing factors exponentiated according to the participation coefficient. A zero participation coefficient automatically results in a balancing factor reset to 1 that does not affect the household weight.

## Introducing Relaxations

In practice, there are two strong reasons why relaxations of the controls can be necessary. First, the controls themselves are not perfectly consistent and may result in an infeasible system of constraints. This is especially frequent when household-level and person-level constraints are considered. Secondly, the controls are not perfect and represent some approximation of reality (and frequently come from land-use or socio-demographic models). This is especially relevant for future years.

|  |  |
| --- | --- |
|  | Equation |
|  | Equation |
|  | Equation |

Can be solved either by IPF (over HH waits and control relaxations) or (more efficiently) by the Newton-Raphson method.  if all  . However, for some zones and controls . This case is treated as special in the Newton-Raphson method (all HHs that contribute to a zero control with a non-zero coefficient have to be zeroed out and the corresponding relaxation factor is set to 1). IPF does not require a special treatment for zero constraints.

## Newton-Raphson Method w/o Relaxations

The following constraint should be satisfied for each control:

, Equation

This is a polynomial expression with respect to the balancing factor and it has a unique solution because the left hand side is a monotonically increasing function from zero to infinity.

The Newton-Raphson method is based on the following iterative process to find a root of a function starting with the value of :

, Equation

Since the value of is updated at each iteration, only the first step is applied starting with the value of

## Newton-Raphson Method w/Relaxations

The following constraint should be satisfied for each control:

, Equation

This is a polynomial expression with respect to the balancing factor and it has a unique solution because the left hand side is a monotonically increasing function from zero to infinity.

The Newton-Raphson method is based on the following iterative process to find a root of a function starting with the value of :

, Equation

Since the value of is updated at each iteration, only the first step is applied starting with the value of

## Efficient Solution Algorithm and Pseudo-Code

N = households

I = controls

a[n][i] = incidence table

A[i] = control values

U[i] = control importance weights

W[n] = initial weights

LB[n] = Lower bounds on household weights

UP[n] = Upper bounds on household weights

**X[n]** = final weights

MaxGap= max allowable gap to stop

T = max #iterations

Set initial conditions:

R[i]=1.0 // initial relaxation factors

{X[n]=W[n]}

For each t=1….T

For each i=1….I

//Calculate balancing factors by 1 step of Newton-Raphson

XX=Sum by n {X[n]\*a[n][i]}

YY=Sum by n {X[n]\*(a[n][i]^2)}

If XX>0

If A[i]>0

Gamma[i] = 1 - ( XX-A[i]\*R[i] ) / (YY+A[i]\*R[i]\*( 1/U[i] ) )

// if U(i) tends to infinity it collapses to constraint model

// Gamma[i] = 1 - ( XX-A[i]) / YY

// if the control is HH (a[n][i]=0,1) YY=XX

// Gamma[i] = 1 - ( XX-A[i]) / XX = XX/XX - ( XX-A[i]) / XX = A[i]/XX

Else // XX>0.and.A[i]=0

Gamma[i]=0.01

Else // XX=0

Gamma[i]=1

Endif

// Update HH weights

If a[n][i]>0

X[n] = min { max { X[n] \* (Gamma[i]^a[n][i]), LB[n] }, UP[n] }

Endif

// Update relaxation factors

R[i] = R[i] \* (1/Gamma[i])^(1/U[i])

End of i-loop

// Check for convergence

Delta = (Sum over n abs{X[n]-Z[n]})/N

If Delta<=MaxGap END

Z[n]=X[n]

End of t-loop

Return **X[n]**

## Particular Case of Household-Level Controls Only

In this particular case, the contribution coefficients are all Boolean indicators . Consider an example of household controls by size and number of workers. The controls in this case can be broken into two groups:

= household size categories,

= household categories by number of workers.

The solution (4) for this case can be rewritten in the following way:

, Equation

Where:

= controls by number of workers (denoted separately),

= household size category in which the household belong,

= household category by number of workers in which the household belong.

If we sum household weights for each cell of the joint household distribution, we obtain the following formula:

, Equation

Where:

= household distribution by size and number of workers,

= total initial weight for each cell.

Formula (10) represents a conventional two-dimensional matrix balancing procedure to achieve the same control totals. Thus, the conventional balancing represents a particular case of the proposed procedure for balancing individual household weights. The table balancing yields exactly the same result as matrix balancing if the table controls and contribution coefficients are specified in a special way.

## Expansion of Household Survey Sample by List Balancing

In addition to its pivotal role in population synthesis, List Balancing proved to be a useful general method for expansion of Household Surveys. Calculation of household expansion factors is similar to generation of synthetic population in many respects. Instead of the PUMS, the sample of households is given by the survey. Sampling frame or other data on the entire regional population play the same role as controlled variables. Instead of setting these controls for each TAZ or smaller zones, they are normally specified at a district level (40-100 districts) to ensure that each district has a sub-sample of 100 or so households.

Notation:

 = sampled households

 = controls for synthetic population

 = control values

 = household attributes corresponding to controls

Problem formulation:

, **(1)**

subject to:

, **(2)**

**. (3)**

Solution:

 **(4)**

 **(5)**

** (6)**

. **(7)**

Thus, solution is reduced to a set of control-specific balancing factors  that should ensure that the controls (2) are met. The convex entropy function (1) guarantees that the solution is unique and also that the uniformity of the expansion factors would be the maximum possible. However, to reach the solution by iterative balancing without constraint (3) would be problematic since the constraint (3) makes the overall scaling parameter  in the expression (7) easily identifiable. Namely:

. **(8)**

Second description.

This step involves a generalized IPF “table balancing” procedure that is based on a maximum entropy principle and works directly with a list of households. This way both household and person characteristics can be taken into account in their relation to the corresponding controls. The model formulation and algorithm are described below for a simplified example where controls are considered for number of households, total population, and total residential labor force. We also consider a single district. Introduce the following notation:

 = micro-sample of households in the aggregate zone,

 = initial expansion factors,

 = household size,

 = number of workers in each household,

 = zonal control for number of households,

 = zonal control for total population,

 = zonal control for residential labor force,

 = (unknown) final expansion factors.

The model can be formulated as optimization of the following entropy function:

, Equation

subject to constraints:

, Equation

, Equation

. Equation

The solution for this problem has the following form:

. Equation

A simple algorithm (IPF) can be applied to practically find this solution:

1. Start with the initial expansion factors .
2. Set initial values for balancing factors .
3. Recalculate balancing factor for constraint (2): .
4. Recalculate balancing factor for constraint (3) by adding .
5. Recalculate balancing factor for constraint (4) by adding .
6. Check for convergence for all constraints (2-4) and go to step 3 if not achieved.

# Discretizing Methods

Discretizing is not a trivial problem when the population is synthesized at a finer level of spatial resolution (30,000-40,000 subzones) and the balancing results in many small fractional numbers. Simple rounding may cause substantial relative deviations from the controls that can subsequently accumulate across multiple subzones. Discretizing replaces drawing household from PUMS. Simple rounding or bucket rounding may violate constraints.

## Post-Discretizing

### Fixed Constraints

mmm

|  |  |
| --- | --- |
|  | Equation |
|  | Equation |
|  | Equation |

We assume that all , since it makes sense for all  to set the corresponding  and exclude them from the optimization (otherwise some of the constraints will be violated). This is, of course, empirical. If we make another empirical step to exclude all integer parts of  (and adjust the constraints correspondingly) this would lead to  and a reasonable assumption that we can limit  (i.e.to be Boolean). An alternative way to achieve  is to break all HHs with a large continuous weight into a set of identical HHs and discretize them separately (probably a cleaner solution although may not be as efficient as an immediate integerizing).

Under these assumptions, the objective function of the discretizing problem can be linearized:

|  |  |
| --- | --- |
|  | Equation |
|  | Equation |
|  | Equation |

This is a LP program with discrete variables that can be solved by using one of the standard methods (like B&B). However, it can be prohibitive in term of runtime since the number of variables can be as large as 5,000 (entire PUMA) or even greater if some HHs have to be borrow from other PUMAs.

### Relaxed Constraints

mmm

|  |  |
| --- | --- |
|  | Equation |
|  | Equation |
|  | Equation |

This problem is identical to discretizing a solution w/o relaxations.

## Direct Formulation of Integer Problem

Mmm

### Fixed Constraints

mmm

|  |  |
| --- | --- |
|  | Equation |
|  | Equation |
|  | Equation |

It is, of course, possible to write the problem directly and w/o breaking it into the continuous and discretizing parts. However, this way we cannot reduce it to an LP formulation since we cannot limit the variables to be Boolean. Thus, it is a discrete non-linear program of a large size that is probably impossible to solve in a reasonable amount of time by B&B or any other algorithm w/o heuristics.

### Relaxed Constraints

mmm

|  |  |
| --- | --- |
|  | Equation |
|  | Equation |
|  | Equation |

As for the case w/o relaxations, there is probably no efficient way to solve this problem directly.

## Discretizing Residual Fractions

Add boundaries for relaxations to make B&B more efficient.

Original LP:

Min Sum[n] C[n]\*X[n]

Sum[n] a[n][i]\*X[n] = A[i]

X[n]=0,1

Naïve relaxation:

Min Sum[n] C[n]\*X[n]

Sum[n] a[n][i]\*X[n] >= A[i]-R[i]

Sum[n] a[n][i]\*X[n] <= A[i]+R[i]

X[n]=0,1

Endogenous relaxation:

Min Sum[n] {C[n]\*X[n]} + Sum[i] {999\*Y[i]+999\*Z[i]}

Sum{[n] a[n][i]\*X[n]} + Y[i] >= A[i]

Sum{[n] a[n][i]\*X[n]} - Z[i] <= A[i]

X[n]=0,1

Y[i]>=0, Z[i]>=0

|  |  |  |  |
| --- | --- | --- | --- |
|  | X | Y | Z |
| Object. Func | C | 999 | 999 |
| 1 constr. | A | +1 | 0 |
| 2 cosntr. | A | 0 | -1 |
| Boundary | Binary | 0,inf | 0,inf |

### Fixed Constraints

mmm

The following steps are included:

1. Solve continuous problem.
2. Integerize variables that are greater or equal than 1.
3. Round down residual variables that are close to 0 (chose variables that are smaller than a predetermined threshold or just chose a predetermined number of variables with the smallest weights to be rounded). It is safer to round down variables that are close to 0 rather than round up variables that are close to 1 since the latter can result in exceeding the constraints.
4. Adjust constraints and calculate residual variables (by excluding the integer part). If residual variables are all zeros, stop.
5. Go to step 1.

### Relaxed Constraints

The following steps are included:

1. Solve continuous problem w/relaxations.
2. Round constraints
3. Solve continuous problem w/o relaxations
4. Integerize variables that are greater or equal than 1.
5. Round down residual variables that are close to 0 (chose variables that are smaller than a predetermined threshold or just chose a predetermined number of variables with the smallest weights to be rounded).
6. Adjust constraints and calculate residual variables (by excluding the integer part). If residual variables are all zeros, stop.
7. Go to step 3.

# Consideration of Multiple Zones

## Meta-Controls

In general, synthetic population for each TAZ cannot be generated independently. Certain demographic tendencies can only be forecasted at the regional level. In some cases, expected tendencies can be defined at the level of aggregate districts. It includes some general long-term demographic trends (for example, diminishing household size or population aging) as well as controls that relate to distribution of workers by industry and occupation. The industry-occupation controls are important to ensure that the generated synthetic population is sensitive to the projected employment forecast for the region. We refer to all controls set at the level of geography higher than TAZ as meta-controls since they can only be imposed on a group of TAZs.

On the other hand, even if MAZ is specified as the main geographic unit it can be too small for many controls to be reliably set. This results in necessity for Population Synthesizer to be flexible across multiple possible geographic levels for setting controls. It is expected that in most cases most of the controls will be set at the TAZ level while some additional controls can be added at a more aggregate level. A subset of TAZ-level controls can be specified at the MAZ level.

Consideration of multiple geographies and treatment of aggregate meta-controls are described in this sub-section (6.1). Allocation procedures for smaller zones are described in the next sub-section (6.2).

### General Formulation

Introduce the following additional notation:

 = TAZs (in the PUMA),

 = household size categories,

 = household categories by number of workers

 = zonal controls (1-dimensional household distributions)

 = seed set of households in the PUMA (survey)

 = a priori weighs assigned in the PUMA (survey)

 = Boolean indicators for a household (1 if belongs to the category, 0

otherwise)

Population synthesis is essentially a finding of the following variables:

 = household allocation to TAZs

where:

 = household expansion factors

We first consider a formulation with continuous variables that can be written as the following convex entropy-maximizing program:

, Equation

Subject to:

, Equation

, Equation

The objective function has two terms that express two intuitive criteria for quality of synthetic population:

 = uniform use of the household sample,

 = spatial allocation by probability,

where:

 = allocation probability estimated from the survey data.

 = relative weight of uniformity vs. spatial allocation probability.

The allocation probability is estimated by means of a disaggregate choice model (essentially, of household residential choice). This model can use any (controlled and uncontrolled) household variables that are available both in the survey and PUMS as well as TAZ characteristics.

By forming the Lagrangian and equating the derivatives to zero we obtain the following solution:

, Equation

where  and  corresponds to the actual household size and number of workers for which  and .

### Particular Case with no Explicit Consideration of Uniformity

If we assume  that is no uniformity consideration, the solution (4) takes the following particular form:

. Equation

This means that the balancing procedure starts with the household weights  and then proceeds through IPF by the zonal controls. In the absence of uniformity term, the IPF is implemented for each zone separately.

This formula preserves a full analogue with the conventional multidimensional balancing of household distribution to match the aggregate controls as applied at the first stage of many population synthesis procedures in practice.

The conventional multi-dimensional balancing of the household distribution can be written as the following convex entropy-maximizing problem:

, Equation

subject to the same constraints:

, Equation

 Equation

where:

 = unknown distribution of households in each TAZ,

 = seed distribution of households in the TAZ.

The problem is solved independently for each TAZ where the solution can be written in the following form:

. Equation

The seed distribution in normally calculated based on the observed set of households in the PUMA. It sometimes includes additional available information that makes it more specific to each TAZ. In general, the seed distribution is calculated by the following formula:

, Equation

where:

 = a priory probability of a household to be included in the TAZ sample.

In a simple case, the a priori probabilities are all set to 1 (i.e. a single and full set of households in the PUMA is used as a seed for all TAZs). By manipulating inclusion probabilities it is possible to create TAZ-specific seeds. For example, if the location of ethnic clusters in the PUMA is known, the households with the corresponding ethnicity could be given a high probability for TAZ located the cluster (compared to the other households).

By substituting (10) into (9) we obtain the following expression for a zonal household distribution:

, Equation

that is easy to compare to expression (5). If we aggregate the individual household allocation variable (5) into a zonal distribution by the following formula:

. Equation

It can be seen by comparing (11) and (12) that the disaggregate balancing of individual households (11) gives essentially the same aggregate household distribution as the conventional methods if the inclusion probabilities  used in the conventional method are identical to the allocation probabilities  used in the disaggregate balancing. In particular, the same results will be achieved if both inclusion probabilities and allocation probabilities are specified uniformly (i.e. no information was available to make the seed distribution or household allocation TAZ-specific).

The important advantage of the disaggregate formulation (1-3) is that it also explicitly considers the uniformity of household use. Introduction of the uniformity consideration into the aggregate framework (6-8) is problematic since it does not have index . Additional important technical advantage of the disaggregate formulation is that it is less sensitive to the number of dimensional constraints. In the disaggregate formulation (1-3), the number of variables is independent of number of controls. In the aggregate formulation (6-8), the number of variables is sensitive to the number of controls and is essentially proportional to the Cartesian product over all controlled dimensions.

Finally, the disaggregate approach can be better combined with the subsequent creation of the list of discrete individual households. With the disaggregate approach, the balancing outcome already has a form of a list of individual households with weights in each TAZ. Thus, the final step will only require a discretizing of the weight for each household. Contrary to that, with the aggregate approach that ends with the multidimensional household distribution in each zone, the discretizing should also be combined with an additional procedure of drawing individual households from the PUMA. The drawing procedure in itself can introduce uncontrolled randomness and violate the uniformity principle. While a certain effort has been made in the existing population synthesis procedures to ensure a uniform drawing, it is still largely empirical and a full avoidance of the post-balancing drawing (with an explicit analytical control for uniformity through the corresponding term in the objective function) in our view represents a clear advantage of the disaggregate approach.

### Meta-Balancing Efficient Implementation Algorithm

t = TAZs in county

n = unique index of PUMS HH across all PUMAs

a[n][k] = incidence for new controls (set at the county level)

A[k] = county level controls

After balancing for each TAZ:

X[n][t]>=0 = fractional balancing solution

Step 1: Calculate current values for county-level controls by TAZ:

z[k][t] = Sum by n { X[n][t] \* a[n][k]}

Step 2: Calculate current totals for county-level controls:

Z[k} = Sum by t { z[k][t] }

Step 3: Distribute county-level controls by TAZ:

w[k][t] = z[k][t] \* A[k] / Z[k}

… new round of balancing and LP discretizing for each TAZ:

X[n][t]>=0 = final integer solution

Performance against county-level controls:

Z[k] = Sum by n,t { X[n][t] \* a[n][k]}

Compare Z[k] to A[k]

## Household Allocation to Sub-Zones

Mmm

m = MAZ index within TAZ (ordered by #HHs from smallest to largest)

X[n]>0 = final integer solution for TAZ (excluding zero-weight HHs)

“i” = subset of controls applied at MAZ level

U[i] = control importance weights preserved

a[n][i] = incidence table (the same)

A[m][i] = MAZ-level controls consistent with the TAZ controls ({Sum by m A[m][i]}=A[i])

For m=1 to M-1

If #HH>0

Step 1: Balance and integerize for MAZ using X[n] as the sample weights

Solution is z[m][n]

Step 2: Update the sample to ensure allocation w/o replacement:

X[n] = max{X[n]-z[m][n],0}

End if

End for

Step 3: Set the last (biggest) MAZ weights to the residual:

z[M,n] = X[n]

Property:

{Sum by m z[m,n]} may not be equal to X[n].

# Validation of Population Synthesizer

Mmm

SPG has to be carefully validated and calibrated, i.e. some parameters that regulate balancing convergence and discretizing steps can be changed depending on the SPG performance. In general, SPG validation includes three steps:

* Validation of the *match achieved for the controlled variables*. Ideally, we would like to have a complete match across all controls for each TAZ. In reality, it is difficult to achieve for several reasons. First, the discretizing procedure can introduce some discrepancies even if the balancing has converged without a need in relaxation. Secondly, in many cases some structural problems with the controls themselves can result in relaxations. This step is also a very useful for a debugging of SPG. An example of a detailed graphical validation report developed by PB as a routine part of the San Diego SPG is shown in **Error! Reference source not found.**. In this report, for each controlled variable, the maximum, minimum, and mean percentage difference between the Census population and synthetic population across all TAZs is shown. A similar report will be developed for the BMC SPG.
* Validation of the *reasonability of the uncontrolled variables*. Uncontrolled variables are modeled through the drawing (or dicretizing) of households that is a less rigid procedure compared to balancing. As the result, more substantial discrepancies might be observed for uncontrolled variables compared to controlled. The validation with respect to uncontrolled variables can be implemented either for the base year or for one of the past years (so called “backcast” that is in general considered as a better test that the base year validation; backcast however, is much more time taking since it is require a setting of all inputs for one of the past years). In both cases, the calibration targets have to be developed from the Census data or PUMS and compared to the SPG output. A useful test of this type that we propose for the BMC SPG is comparison between the actual (Census or expanded PUMS) population by person type (1-8) listed in Table 1 above since these person types are the most important for the ABM. In general, uncontrolled variables are matched well if they are correlated with the controlled variables. In this sense, this step is also useful to assess the completeness and reasonability of the SPG input in terms of the controlled variables.
* Testing *future socio-economic and demographic scenarios*. It is useful to specify several (preferably extreme) scenarios with significant structural changes in the controlled variables, for example income growth and/or change in household size or average person age and test the SPG performance in terms of controlled variables as well as impact on uncontrolled variables. These tests will be specified jointly with the BMC staff as part of Task 3.3.

# Complete Regional Population Synthesizer for MAG

## Structure of Population Synthesizer

Overall structure of the MAG Population Synthesizer is presented in . The procedure goes through three consecutive steps: 1) Initial balancing at the TAZ level and meta-balancing, 2) Final balancing and discretizing at the TAZ level, 3) Allocation to MAZs.

Figure : MAG Population Synthesizer



## Zone System in the MAG-PAG Metropolitan Region

## Household Sample from PUMS

## Controls

## Summary of Results

# Complete Regional Population Synthesizer for JTMT

## Zone System in the Jerusalem Metropolitan Region

## Household Sample from the Israeli Population Census

## Controls

## Summary of Results

Mmm

# References

## Conventional PopSyn Based on Joint Distribution of Households and IPF

Arentze, T., H.J.P.Timmermans, and F. Hofman. (2007) Creating synthetic household populations: problems and approach. *Transportation Research Record* 2014, 85-91.

Auld, J., Mohammadian, A., and K. Wies (2008). Population synthesis with region-level control variable aggregation. *Presented at the 87th Annual Meeting of the Transportation Research Board*, Washington DC.

Auld, J. and A. K. Mohammadian. (2010) An Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels. Presented at the *89th Annual Meeting of the Transportation Research Board*, Washington, D.C.

Beckman, R. J., K. A. Baggerly and M. D. McKay. (1996) Creating synthetic baseline populations. *Transportation Research Part A: Policy and Practice*, Vol. 30 Issue 6, pp 415–429.

Bowman, J.L. (2004). A Comparison of Population Synthesizers Used in Microsimulation Models of Activity and Travel Demand. Working Paper, http://jbowman.net/papers/B04.pdf, accessed August 1, 2007.

Bowman, J.L and G. Rousseau (2006) Validation of the Atlanta (ARC) Population Synthesizer (PopSyn). Paper prepared for the TRB Conference on Innovations in Travel Modeling, May 21-23, 2006. Austin, TX.

Bowman, J.L., Rousseau, G. (2008) Validation of Atlanta, Georgia, regional commission population synthesizer. *Transportation Research Board Conference Proceedings* 2 (42), 54-62.

Frick, M. and K.W. Axhausen (2004). Generating Synthetic Populations using IPF and Monte Carlo Techniques: Some New Results. Paper presented at the 4th Swiss Transport Research Conference, March 25-26, 2004.

Guo, J. Y. and C. R. Bhat (2007) Population Synthesis for Microsimulating Travel Behavior, *Transportation Research Record*, No. 2014, pp. 92–101.

Hobeika, Antoine (2005). TRANSIMS Fundamentals: Chapter 3 Population Synthesizer, Technical report, U.S. Department of Transportation, Washington, D.C.

Müller K. and K. W. Axhausen. (2010) Population Synthesis for Microsimulation: State of the Art. Presented at Swiss Transport Research Conference, 2010.

PB Consult (2003). Task 2: Household and Population Synthesis Procedure, PB Consult/Parsons Brinckerhoff, report prepared for the Mid-Ohio Regional Planning Commission as part of the MORPC Model Improvement Project. March 12, 2003.

## Combinatorial Optimization and Annealing

Abraham, J.E., K.J. Stephan, and J.D. Hunt (2012) Population Synthesis Using Combinatorial Optimization at Multiple Levels. Presented at the *91st Annual Meeting of the Transportation Research Board*, Washington, DC.

Huang Z. and P. Williamson. (2001) *A Comparison of Synthetic Reconstruction and Combinatorial Optimization Approaches to the Creation of Small-area Microdata.* Working paper, University of Liverpool.

Ma (2011) Generating Disaggregate Population Characteristics for Input to Travel Demand Models, Doctoral Dissertation, the University of Florida

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

Ryan, J., H. Maoh and P. Kanaroglou. (2007) Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms. Working paper CSpA WP 026, McMaster University.

Ryan, J., H. Maoh, and P. Kanaroglou (2010) Population synthesis for microsimulating urban residential mobility. Paper presented at the 89th Transportation Research Board Annual Meeting, Washington, DC.

Srinivasan, S., Ma, L., Yathindra, K. (2008) Procedure for Forecasting Household Characteristics for Input to Travel-demand Models. Project Report of University of Florida, Gainesville; Florida Department of Transportation, TRC-FDOT-64011-2008

Voas, D., Williamson, P. (2000). An evaluation of the combinatorial optimization approach to the creation of synthetic microdata. International Journal of Population Geography 6 (5), 349-366.

Williamson, P., M. Birkin and P. H. Rees. (1998) The Estimation of Population Microdata by Using Data From Small Area Statistics and Samples of Anonymised Records. *Environment and Planning A*, Vol. 30, pp 785 – 816.

## List Balancing

Bar-Gera, H., K. Konduri, B. Sana, X. Ye, and R.M. Pendyala (2009) Estimating Survey Weights with Multiple Constraints Using Entropy Optimization Methods. Presented at the *88th* *Annual Meeting of the Transportation Research Board*, Washington, DC.

Ye, X., K. Konduri, R.M. Pendyala, B. Sana, and P. Waddell (2009) A Methodology to Match Distributions of Both Households and Person Attributes in the Generation of Synthetic Populations. Presented at the *88th* *Annual Meeting of the Transportation Research Board*, Washington, DC.

Lee, D., Fu, Y. (2011) A cross entropy optimization model for population synthesis used in activity-based micro-simulation models. Paper presented at the 90th Annual Meeting of the Transportation Research Board, Washington, DC.

Sun, W., C. Daniels, Z. Ouyang, P. Vovsha and J. Freedman (2011) Comparisons of Synthetic Populations Generated From Census 2000 and American Community Survey (ACS) Public Use Microdata Sample (PUMS). Presented at the 13th Transportation Planning Application Conference, Reno, NV.

Harland K., A. Heppenstall, D. Smith and M. Birkin. (2012) Creating Realistic Synthetic Populations at Varying Spatial Scales: A Comparative Critique of Population Synthesis Techniques, *Journal of Artificial Societies and Social Simulation*, **15 (1)**, 1-24.

# List Balancing Using the Newton-Raphson Method

This follows Bar-Gera, et al, 2009.

## Strict Formulation without Relaxations

|  |  |
| --- | --- |
|  | Equation |
|  | Equation 57 |
|  | Equation 58 |
| Subject to (Equation 56) | Equation 59 |
|  | Equation 60 |
| The Lagrangian: | Equation 61 |
| = 0 | Equation 62 |
|  | Equation 63 |
|  | Equation 64 |
|  | Equation 65 |
| Looks redundant | Equation 66 |
| Looks redundant | Equation 67 |
|  | Equation 68 |
| Coordinate-by-coordinate iterative search algorithm from to : |  |
|  | Equation 69 |
| From (Equation 64), (Equation 65): |  |
|  | Equation 70 |
| From (Equation 68), the optimal point with respect to : |  |
|  | Equation 71 |
|  | Equation 72 |
| (Equation 72) is polynomial in and has a unique solution because the left-hand side is monotonically increasing from 0 to when increases from zero to .. |  |
| Newton-Raphson method for the strict formulation w/o relaxations: |  |
| Step 1: Set iteration number . |  |
| Step 2: If , start with ; if , start with |  |
| Step 3: Loop over controls |  |
| Step 4: Calculate balancing factor correction by one step of Newton-Rapshon method using (Equation 72) with the latest weights and initial .  The function of for which we want to find a root is given by (Equation 72):    Using the Newton-Raphson method: |  |
| Step 5: Update weights using (Equation 70): |  |
| Step 6: If , go to Step 4; if , go to Step 7 |  |
| Step 7: If , stop; otherwise go to Step 2 |  |

## Generalized Formulation with Relaxations

|  |  |
| --- | --- |
| ,  Where represent relaxation factors | Equation |
|  | Equation |
| Subject to: |  |
|  | Equation |
|  | Equation |
|  | Equation |
|  | Equation |
| Entropy is used as a penalty function to obtain relaxation factors as close to 1 as possible. Coefficients reflect importance of the controls. |  |
| The Lagrangian: | Equation |
| The first-order condition for minimization is applied and the partial derivatives of Lagrangian with respect to will yield the same equations as for the strict formulation w/o relaxations: |  |
| = 0 (Equation 62) |  |
| (Equation 63) |  |
| First-order partial derivatives of Lagrangian with respect to relaxation factors will return the following constraint: | Equation 80 |
| From () relaxation factor can be expressed as | Equation 81 |
| First-order partial derivatives of dual Lagrangian with respect to will return the following constraint that is different from (Equation 68): | Equation 82 |
| Correction for balancing factor is defined as was done in (Equation 69)    and then from (Equation 81) | Equation 83 |
| By substituting (Equation 83) to (Equation 82 or identical Equation 75), the following non-linear equation in is obtained similar to (Equation 72):    This equation also has a unique solution like (Equation 72) because the left-hand side is monotonically increasing from zero to while the right-hand side is decreasing from to zero when increases from zero to . This can be solved as before by using the Newton-Raphson method. | Equation 84 |
| Newton-Raphson method for the generalized formulation with relaxations: |  |
| Step 1: Set iteration number . |  |
| Step 2: If , start with and  if , start with and |  |
| Step 3: Loop over controls |  |
| Step 4: Calculate balancing factor correction by one step of Newton-Rapshon method using (Equation 84) with the latest weights and initial .  The function of for which we want to find a root is given by (Equation 84):    Using the Newton-Raphson method: |  |
| Step 5: Update weights using (Equation 70): |  |
| Step 6: Update relaxation factors using (Equation 83): |  |
| Step 7: If , go to Step 4; if , go to Step 8 |  |
| Step 8: If , stop; otherwise go to Step 2 |  |