New Features of Population Synthesis

Peter Vovsha
Parsons Brinckerhoff
1 Penn Plaza, 3rd Floor, New York, NY 10119,
212-465-5511, Vovsha@PBworld.com

James E. Hicks
Parsons Brinckerhoff,
6100 Uptown Blvd, Suite 700, Albuquerque, NM 87110
505-881-5357, Hicksji@PBworld.com

Binny M. Paul
Parsons Brinckerhoff
1 Penn Plaza, 3rd Floor, New York, NY 10119,
212-465-5528, Paulbm@PBworld.com

Vladimir Livshits
Maricopa Association of Governments
302 North First Avenue, Suite 300, Phoenix, AZ 85003
602-254-6300, vlivshits@mag.maricopa.gov

Petya Maneva
Maricopa Association of Governments
302 North First Avenue, Suite 300, Phoenix, AZ 85003
602-452-5075, PManeva@azmag.gov

Kyunghwi Jeon
Maricopa Association of Governments
302 North First Avenue, Suite 300, Phoenix, AZ 85003
602-254-6300, KJeon@azmag.gov

Paper size: 5,750 words (text) + 3*250 (tables) + 4*250 *(figures) = 7,500 words

Submitted for presentation at the 94th Annual Meeting of the Transportation Research Board and publication in the Transportation Research Records

August 1, 2014

Abstract

Population synthesis represents an important first step in the Activity-Based Modeling (ABM) chain. The current research belongs to the group of analytical methods that balance a “list” (or sample) of households to meet the controls imposed at some level of geography, normally, for each Traffic Analysis Zone (TAZ). The paper is based on the implemented and validated Population Synthesizer that constitutes a part of the ABM developed for the Maricopa Association of Governments (MAG). The focus of the paper is on several innovative features. The first one relates to a general formulation of convergence of the balancing procedure with imperfect (i.e. not fully consistent) controls. The second one relates to the optimized discretizing of the fractional outcomes of the balancing procedure to form a list of discrete households. The proposed method employs a Linear Programming (LP) approach in order to optimize the discretized weights and preserve the best possible match to the controls. The third one relates to multiple levels of geography where the controls can be set. The geographic flexibility is essential for two different reasons. First, some important demographic, socio-economic, and land-use development trends that affect population synthesis can only be translated into more aggregate controls than a TAZ-level control. Secondly, ABMs of the new generation operate with an enhanced level of spatial resolution where all location choices are modeled at the level of Micro-Analysis Zones (MAZs) nested within the TAZs. The paper provides validation results for the MAG region.

Keywords:

Population synthesis, entropy maximization, balancing procedure, household distribution

Motivation and Statement of Innovations

Substantial progress has been made recently in the Population Synthesis (PopSyn) methods applied in Activity-Based Models (ABMs). Some long-standing issues like addressing both household-level and person-level controls have been successfully addressed. Overall, the applied methods can be broken into two groups:

- Analytical methods that balance a "list" (or sample) of household weights to meet the controls imposed at some level of geography, normally, for each Traffic Analysis Zone (TAZ) (1,2,3,4,5).
- Combinatorial methods that are based on a random swapping of households between TAZs if the fit measures can be improved (6,7,8,9,10,11,12,13,14).

The current research belongs to the first group. The paper is based on the implemented and validated PopSyn that constitutes a part of the CT-RAMP (Coordinated Travel & Regional Activity Modeling Platform) ABM developed for the Maricopa Association of Governments (MAG). The paper describes the following innovative features of the MAG PopSyn:

- General formulation of convergence of the balancing procedure with imperfect (i.e. not fully consistent) controls. In addition to guaranteeing a unique repeatable solution this approach is also useful for screening inconsistencies and addressing a differential degree of confidence in different controls.
- Optimized discretizing of the fractional outcomes of the balancing procedure to form a list of discrete household weights. It should be noted that with the enhanced level of spatial resolution and growing number of controls to meet, rounding errors become substantial if a simple "bucket" rounding is applied. The proposed method employs a Linear Programming (LP) approach in order to optimize the discretized weights and preserve the best possible match to the controls.
- Multiple levels of geography where the controls can be set. This flexibility is essential for two different reasons. First, some important demographic, socio-economic, and land-use development trends that affect population synthesis can only be translated into more aggregate controls than a TAZ-level control. Secondly, the new generation of CT-RAMP ABMs operate with an enhanced level of spatial resolution where all location choices are modeled at the level of Micro-Analysis Zones (MAZs) nested within the TAZs. Thus, it is essential to extend the PopSyn procedure to handle MAZs in addition to TAZs. These considerations lead to a multi-level procedure where aggregate levels such as sub-area have to be effectively combined with TAZ and MAZ levels in a consistent way.

Basic Formulation of List Balancing

The core List Balancing method is formulated for a single zone (TAZ). We first formulate the method for a case where the controls are consistent and should be satisfied exactly. This method creates a list of households directly from the sample and eliminates the need for subsequent random drawing of households from the PUMS. Existing PopSyn procedures that are based on the multidimensional household distribution in each zone require balancing to be combined with an additional procedure of drawing individual households from the PUMS. The drawing procedure in itself can introduce uncontrolled randomness and violate the uniformity principle. While a certain effort has been made in the existing PopSyn procedures to ensure a uniform drawing, it is still largely empirical. Avoidance of the post-balancing drawing with an explicit analytical control for uniformity represents an advantage of the List Balancing method.

Introduce the following notation:

$i = 1,2...I$ = indices for household and person controls,

A_i = values of controls to be met for the given zone,

$n \in N$ = seed set of households in the zone's PUMA (or any other sample),

w_n = a priori household weights assigned in the PUMS,

$a_{ni} \geq 0$ = household attribute incidence, i.e. coefficients of contribution to each control.

An example is shown in **Table 1** below for controls specified by household size categories and person age brackets.

Table 1: Controls and Household Contribution Coefficients

HH ID	HH size				Person age				HH initial weight
	1	2	3	4+	0-15	16-35	36-64	65+	
	$i=1$	$i=2$	$i=3$	$i=4$	$i=5$	$i=6$	$i=7$	$i=8$	w_n
$n=1$	1							1	20
$n=2$		1			1	1			20
$n=3$			1			1	2		20
$n=4$				1		2	2		20
$n=5$				1	1	3	2		20
Initial weighted values	20	20	20	40	40	140	120	20	
Control (A_i)	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 six persons: one person of age 0-15, three persons of age 16-35, and two persons of age 36-64. The essence of List Balancing is to find household weights $\{x_n\}$ that would make current values in each column equal to the controls for the column.

The List Balancing procedure can be written as a convex entropy-maximization problem where household weights $\{x_n\}$ are optimized in the following way:

$$\min_{\{x_n\}} \sum_n x_n \ln(\frac{x_n}{w_n}), \qquad \text{Equation 1}$$

Subject to constraints:

$$\sum_n a_{ni} \times x_n = A_i, (\alpha_i), \qquad \text{Equation 2}$$

$$x_n \geq 0 \, , \qquad \text{Equation 3}$$

where α_i represents dual variables that give rise to the balancing factors.

The objective function expresses the principle of using all households as uniformly as possible, i.e. proportionally to the assigned a priori weight. The constraints ensure matching the controls. This is a convex objective function 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:

$$L(\{x_n\}) = \left(\sum_n x_n \ln \frac{x_n}{w_n}\right) - \sum_i \alpha_i \left[\left(\sum_n a_{ni} x_n\right) - A_i\right]. \qquad \text{Equation 4}$$

We calculate partial derivatives and equate them to zero:

$$\frac{\partial L(\{x_n\})}{\partial x_n} = \ln \frac{x_n}{w_n} + 1 - \sum_i \alpha_i (a_{ni}) = 0 \, . \qquad \text{Equation 5}$$

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

$$x_n = k \times w_n \times exp(\textstyle\sum_i a_{ni}\, \alpha_i) = w_n \times \prod_i [exp(\alpha_i)]^{a_{ni}} = w_n \times \prod_i (\hat{\alpha}_i)^{a_{ni}}, \qquad \text{Equation 6}$$

where $\hat{\alpha}_i$ represents balancing factors that have to be calculated iteratively.

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

The basic formulation may be generalized to account for relaxation of constraints that avoid non-convergence if the controls are not consistent within themselves. In practice, there is a strong reason why relaxations of the controls can be necessary. The controls are not always perfectly consistent and may result in an infeasible system of constraints. This is especially frequent when household-level and person-level constraints are considered. Also, the controls represent some approximation of reality and frequently come from different land-use or socio-demographic models. This is especially relevant for future years. A simple fixed-constraint formulation cannot converge if the controls are not perfectly consistent and the result of balancing would be arbitrary dependent on the order of processing of controls and the step where the balancing procedure has been terminated.

Without a systematic way to resolve inconsistencies in different controls, adjustments made to enforce consistency are often based on arbitrary or inequitable decisions. It is therefore desirable to generalize the List Balancing procedure to address this issue and support operating with imperfect controls in a systematic, unbiased way. In this case, rather than trying to match all controls exactly the procedure should search for a compromise solution where the controls as specified are met to the extent possible. This can be achieved by the following formulation with relaxations that also has a unique solution:

$$\min_{\{x_n,z_i\}} \sum_n x_n \ln \frac{x_n}{w_n} + \sum_i \mu_i z_i \ln z_i, \qquad \text{Equation 7}$$

Subject to constraints:

$$\sum_n a_{ni} \times x_n = A_i \times z_i, (\alpha_i), \qquad \text{Equation 8}$$

$$x_n \geq 0, z_i \geq 0, \qquad \text{Equation 9}$$

Where z_i represent relaxation factors and μ_i represent importance weights for the controls. The greater is the weight the more the corresponding relaxation factor is penalized for being different from 1.0.

This program also represents a convex optimization problem with linear constraints that can be solved by simple balancing as was the previously described fixed-constraint problem (balancing is now implemented over household weights and control relaxations). However, a much more efficient Newton-Raphson method can also be employed. In general, all $x_n > 0$ if all $A_i > 0$. However, for some zones some controls can take a zero value: $A_i = 0$. Solution with relaxation has the following form:

$$x_n = w_n \times \prod_i (\alpha_i)^{a_{ni}},$$

$$z_i = (1/\alpha_i)^{(A_i/\mu_i)},$$

This case is treated as special in the Newton-Raphson method (all households 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). The corresponding algorithm is described below. This algorithm also allows for a natural incorporation of another constraint that is useful in practice – boundaries on the upper and lower values for the household weights. This is important to preserve a desired uniformity in household weights and prevent the solution of going into extremes when some households are expanded heavily while some other ones obtain a relatively small weight. For this reason, each household weight is bound to be between the lower bound $\underline{w}_n$ and upper bound $\overline{w}_n$ that was most frequently set as one fifth of the initial household weight and five times of the initial household weight correspondingly.

$\frac{w_n}{5} = \underline{w}_n \leq x_n \leq \overline{w}_n = 5 \times w_n,$ Equation 10

The iterative algorithm uses two stopping criteria (whichever applies first): max allowable gap ($MaxGap$) and max number of iterations (T). The algorithm goes through the following five steps:

```
Step 1: Set initial conditions:
$z_i = 1.0$ // initial relaxation factors
$x_n(0) = w_n$ // initial household weights
For each iteration $t = 1,2, \dots T$
    For each control $i = 1,2, \dots I$
        //Step 2: Calculate balancing factors by 1 step of Newton-Raphson method:
        $X = \sum_n [x_{n,t-1} \times a_{ni}]$ ; $Y = \sum_n [x_{n,t-1} \times (a_{ni})^2]$ ;      Equation 11
        If $X$>0
            If $A_i$>0
                $\alpha_i = 1 - \frac{X - A_i \times z_i}{Y + A_i \times z_i \times (1/\mu_i)}$,      Equation 12
            Else // $X$>0 and $A_i = 0$:
                $\alpha_i$=0.01,
        Else // $X = 0$
            $\alpha_i$=1,
        Endif
        // Step 3: Update HH weights:
        If $a_{ni}$>0
            $x_{n,t} = x_{n,t-1} \times (\alpha_i)^{a_{ni}}$,      Equation 13
            $x_{n,t} = \max(x_{n,t}, \underline{w}_n)$; $x_{n,t} = \min(x_{n,t}, \overline{w}_n)$      Equation 14
        Endif
        // Step 4: Update relaxation factors:
        $z_i = z_i \times (1/\alpha_i)^{(1/\mu_i)}$,      Equation 15
    End of loop over controls
    // Step 5: Check for convergence:
    If $[\sum_n |x_{n,t} - x_{n,t-1}|]/N \leq MaxGap$, end      Equation 16
End of loop over iterations
```

If the controls are consistent, this algorithm performs exactly as the balancing algorithm with fixed constraints and delivers the same solution if all importance factors are set to large values, for example $\mu_i \geq 1{,}000$. However, if the controls are internally inconsistent and no solution exists that would satisfy them exactly, this algorithm has a clear advantage over the balancing procedure with fixed constraints. While the basic balancing procedure does not converge at all, this algorithm still produces a unique convergent solution where controls would be satisfied to the extent possible. The degree of the necessary relaxation of each control would be inversely related to the assigned importance weight. Controls with relatively higher weights would be satisfied to a higher degree while the less important controls, or those with less confidence, would be sacrificed more.

Discretizing Fractional Outcomes of Balancing

An additional enhancement that was added to the MAG PopSyn is an innovative discretizing method applied for the household weights and integrated with the balancing procedure. Discretizing is not a trivial problem when the population is synthesized at a finer level of spatial resolution (30,000-40,000 MAZs) and the balancing results in many small fractional numbers. Simple rounding may cause substantial deviations from the controls that can subsequently be accumulated across multiple MAZs. If simple rounding is forced to match controls exactly it may cause significant deviation from the distribution of initial weights.

The discretizing problem can be formulated as replacing the fractional household weights x_n with integer weights y_n that would preserve the matched controls as well as possible *and* would achieve uniformity of household weights to the maximum extent.

We assume that all $x_n > 0$, since it makes sense for all $x_n = 0$ to set the corresponding $y_n = 0$ and exclude them from the optimization. We make another empirical step to preserve all integer parts of x_n and adjust the constraints correspondingly where $\underline{A}_i$ represents residual controls net of the values taken by the integer parts of x_n:

$$\underline{A}_i = A_i - \sum_n a_{ni} \times int\ (x_n), \qquad \text{Equation 17}$$

This would lead to residual fractional values $0 < \underline{x}_n < 1$ where $\underline{x}_n = x_n - int\ (x_n)$ and a reasonable assumption that we can limit $y_n = 0{,}1$ (i.e.to be Boolean).

Under these assumptions, the discretizing problem can be written as a maximum entropy problem that can be linearized (note that $\underline{x}_n$ are fixed inputs and y_n are the variables to be optimized):

$$\min \sum_n y_n \times \begin{bmatrix} \ln\left(\frac{y_n}{x_n}\right), & if & y_n = 1 \\ 0 & if & y_n = 0 \end{bmatrix} \quad \Rightarrow \quad \max \sum_n y_n \times \ln x_n\ , \qquad \text{Equation 18}$$

Subject to constraints:

$$\sum_n a_{ni} \times y_n = \underline{A}_i\ , \qquad \text{Equation 19}$$

$$y_n = 0{,}1 \qquad \text{Equation 20}$$

This LP problem also needs relaxations since there is no guarantee that its constraints are always feasible. This can be achieved by introducing linear relaxation factors with large penalties and replacing the main equality constraint with two inequalities:

$$\max\{\sum_n y_n \times \ln \underline{x}_n - 999 \times \sum_i U_i - 999 \times \sum_i V_i\}, \qquad \text{Equation 21}$$

Subject to constraints:

$$\sum_n a_{ni} \times y_n \leq \underline{A}_i + U_i, \qquad \text{Equation 22}$$

$$\sum_n a_{ni} \times y_n \geq \underline{A}_i - V_i, \qquad \text{Equation 23}$$

$$y_n = 0{,}1, \qquad \text{Equation 24}$$

$$U_i \geq 0, V_i \geq 0. \qquad \text{Equation 25}$$

This is a LP problem with discrete and continuous variables that is solved efficiently by using a standard method (LP Solver). If there is a feasible solution where relaxations are equal to zero this solution will be found. If there is no feasible solution, some minimal relaxations will be introduced. The essence of this program is to find discrete variables y_n that represent the original fractional variables $\underline{x}_n$ in the best statistical way (bigger $\underline{x}_n$ most probably converted to 1 while smaller $\underline{x}_n$ most probably converted to zero) but also satisfies the constraints as close as possible.

Multiple Levels of Geography

So far, we have described the core balancing and discretizing procedures under an example of a single zone (TAZ). However, in general, synthetic population for each TAZ cannot be generated independently. Certain demographic tendencies can only be predicted 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 and consistent with the land-use forecast in terms of number of jobs by industry. 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.

Another important related detail of the new generation of ABMs is that all location choices are modeled with a finer spatial resolution (MAZs nested within TAZs as the main approach adopted for CT-RAMP ABMs; it can be even a smaller unit like parcel as was adopted in some other ABMs). However, 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 PopSyn to be flexible across multiple possible geographic levels for setting controls. It is expected that in most cases, controls will be set at the TAZ level while some additional controls can be added at a more aggregate level or disaggregate level (MAZ). It should be understood that any control set at a certain level can be propagated up the spatial hierarchy. For example, any control specified at the MAZ level can be summed up to the TAZ-level control. Similarly, any TAZ-level control can be summed up to the district-level or regional-level control. However, the controls, in general cannot be easily disaggregated. For example, a regional-level control like total number of workers in certain industry cannot be immediately disaggregated down to TAZs. This control is inherently of meta-type and its disaggregation by TAZs should be endogenous in the PopSyn.

We outline below two additional PopSyn features that relate to meta-balancing and disaggregation. These procedures together with the core balancing and discretizing procedures constitute the complete PopSyn methodology.

Meta-Balancing

Imposing meta-level controls can be written rigorously as an extension of the core List Balancing procedure described for a single zone above. This would result in an entropy-maximizing problem where the household weights would be optimized simultaneously for all TAZs in the region with accounting for controls at the TAZ level as well as all upper levels of geography. While this formulation is useful for theoretical analysis it is impractical due to a huge dimensionality. Thus, the problem has to be decomposed. The method that allows for such decomposition is outlined below for a case where meta-controls are set at the entire-region level.

Introduce additional notation:

$t = 1,2, \dots T$ = TAZs in the region,

A_{it} = values of TAZ-level controls,

$k = 1,2, \dots K$ = meta-controls,

B_k = values for meta-controls,

b_{nk} = coefficients for each household contribution with respect to meta-controls,

The procedure includes the following five steps:

Step1: Balancing for each TAZ without meta-controls as described above with the TAZ-level set of controls $\{A_{it}\}$ and contribution coefficients $\{a_{ni}\}$ to obtain a fractional solution for the household weights in each TAZ $\{x_{nt}\}$.

Step 2: Calculate resulted values for meta-control variables at the TAZ level:

$$C_{kt} = \sum_n x_{nt} \times b_{nk}, \qquad \text{Equation 26}$$

Step 3: Calculate current regional totals for meta-controls:

$$C_k = \sum_t C_{kt}, \qquad \text{Equation 27}$$

Step 4: Distribute regional meta-controls by TAZs:

$$B_{kt} = C_{kt} \times \frac{B_k}{C_k}, \qquad \text{Equation 28}$$

Step 5: Implement new balancing and discretizing for each TAZ with the extended set of controls $\{A_{it}, B_{kt}\}$ and contribution coefficients $\{a_{ni}, b_{nk}\}$ to obtain a final integer solution $\{y_{nt}\}$.

Allocation of Households from TAZ to MAZs

The essence of this procedure is to allocate household weights generated at the TAZ level to MAZs (or in a general case from any upper level of geography to the lower level of geography). It is essentially a balancing-and-discretizing procedure that is applied to each MAZ using the household weights generated for the TAZ as initial weights. However, this basic procedure is slightly adjusted to address the specifics of the allocation process compared to the general balancing. Additional details are added in order to ensure that the total weight summed across the MAZs is matched for each household after the allocation compared to the original TAZ-level weight. Consider again a single TAZ with household weights obtained after the TAZ-level balancing and discretizing and excluding households that obtained a zero weight $\{y_n = 1,2,3,4, \dots\}$.

Introduce the following additional notation:

$m = 1,2, \dots, M$ = MAZ index within TAZ,

$l = 1,2, \dots, L < I$ = subset of controls applied at MAZ level,

$D_{lm} = 1,2, \dots, L < I$ = values for MAZ-level controls.

All MAZ-level controls are consistent with the corresponding TAZ-level controls:

$$\sum_m D_{lm} = A_{l.} \qquad \text{Equation 29}$$

The proposed allocation algorithm loops over the MAZs and balances the TAZ sample of households with the MAZ controls for each MAZ in a sequence to obtain household weights in each MAZ $\{f_{mn}\}$. The TAZ-level weights of the households are not allowed to be exceeded for each MAZ: $f_{mn} \leq y_n$. The total weight for each household across MAZs should be equal to the TAZ weight: $\sum_m f_{mn} = y_n$. After processing of each MAZ the TAZ-level weights are adjusted to ensure that the TAZ stock of households is used without replacement. The last MAZ is automatically processed as a set of residual household weights ensuring that the total weight for each household is preserved at the TAZ level. In the process of empirical analysis, it was found that the results are slightly better in terms of uniformity of household weights across MAZs if the MAZs are ordered by size from the smallest to the largest rather than processed randomly. Empty MAZs are skipped. The control on total number of households was used for ordering the MAZs. The allocation algorithm can be outlined as follows:

```
For    $m = 1$ to $M - 1$
```

```
        Step 1: Balance and integerize household weights for MAZ using $y_n$ as the sample weights and
        constraining the solution $f_{mn} \leq y_n$. This constraint is introduced in both balancing and
        discretizing steps.

        Step 2: Update the sample weights to ensure allocation without replacement $y_n$ = $y_n - f_{mn}$.

End

Step 3: Set the last (biggest) MAZ weights to the residual $f_{Mn} = y_n$.
```

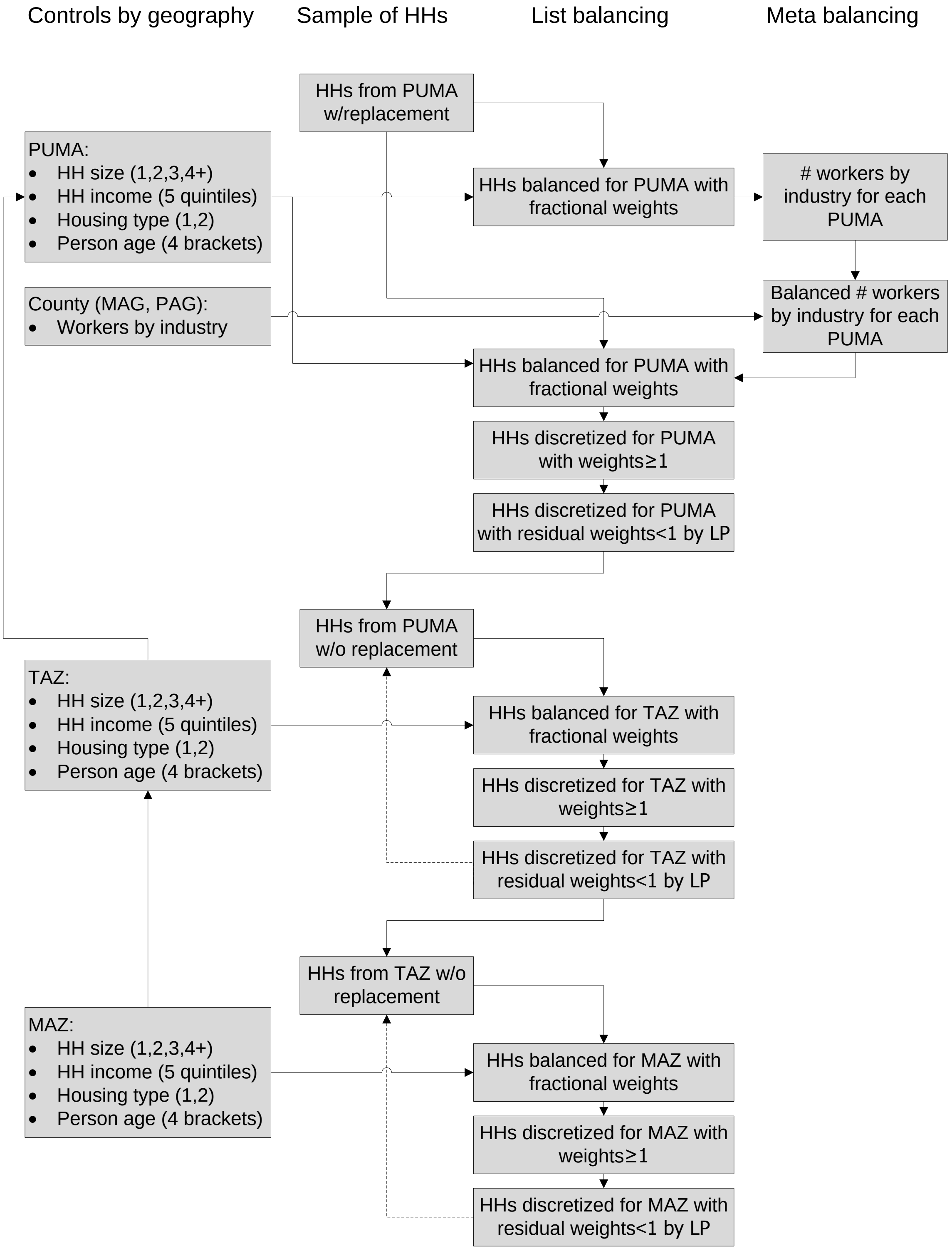
Entire Population Synthesis Algorithm

Overall structure of the PopSyn is presented in **Figure 1**. The input data to the population synthesis can be summarized as follows:

- MAG modeling region included 4 counties with a total population over 4 million.
- Modeled region included 24 PUMAs, 3,009 TAZs, and 26,321 MAZs
- Seed sample of households was based on *PUMS (ACS 2006-10, 5% sample)*
- The following controls were used:
    - **Total no of HHs**, very high importance, (set for each) MAZ
    - **HH size categories** (1,2,3,4+), medium importance, MAZ
    - **Income quintiles**, medium importance, MAZ
    - **Housing type** (single/multi-family), medium importance, MAZ
    - **Person age categories** (0-18,19-35,36-65,66+), medium importance, MAZ
    - **Number of workers by industry type**, medium importance, Meta (entire region)

The procedure goes through four consecutive steps: 1) Initial balancing at the PUMA level and meta-balancing, 2) Final balancing and discretizing at the PUMA level with meta-controls, 3) Allocation from PUMAs to TAZs, 4) Allocation from TAZs to MAZs. These four steps are constructed of the core balancing procedure, discretizing procedure, meta-balancing procedure, and allocation procedures described above. The first step includes initial balancing at the PUMA level and incorporation of meta-controls. The meta-controls were defined by worker industry to ensure that the generated synthetic population is consistent with and sensitive to the land-use projections for the number of jobs by industry. Balancing at the PUMA level proved to be beneficial for having the distribution of household weights as uniform as possible compared to an alternative approach where balancing is applied directly at the TAZ level. The second step finalizes discrete household weights for each PUMA. These weights are further allocated to TAZs within each PUMA (step 3) and to MAZs within each TAZ (step 4).

Figure 1: PopSyn Structure


Controls by geography
Sample of HHs
List balancing
Meta balancing
HHs from PUMA w/replacement
PUMA:
HH size (1,2,3,4+)
HH income (5 quintiles)
Housing type (1,2)
Person age (4 brackets)
HHs balanced for PUMA with fractional weights
# workers by industry for each PUMA
County (MAG, PAG):
Workers by industry
Balanced # workers by industry for each PUMA
HHs balanced for PUMA with fractional weights
HHs discretized for PUMA with weights≥1
HHs discretized for PUMA with residual weights<1 by LP
HHs from PUMA w/o replacement
TAZ:
HH size (1,2,3,4+)
HH income (5 quintiles)
Housing type (1,2)
Person age (4 brackets)
HHs balanced for TAZ with fractional weights
HHs discretized for TAZ with weights≥1
HHs discretized for TAZ with residual weights<1 by LP
HHs from TAZ w/o replacement
MAZ:
HH size (1,2,3,4+)
HH income (5 quintiles)
Housing type (1,2)
Person age (4 brackets)
HHs balanced for MAZ with fractional weights
HHs discretized for MAZ with weights≥1
HHs discretized for MAZ with residual weights<1 by LP

Analysis of the Results

PopSyn has been intensively tested and validated for the MAG region across the following particular dimensions:

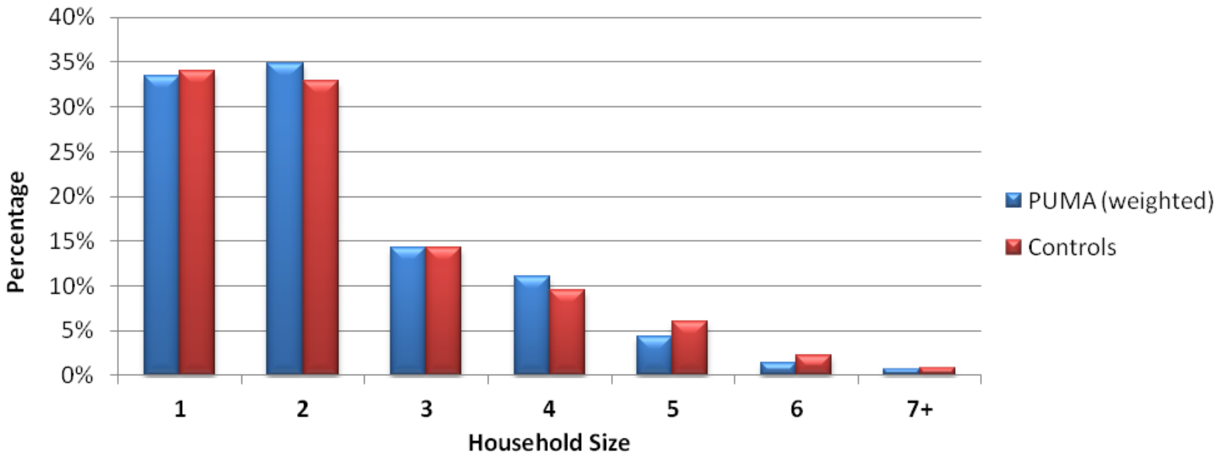
- Consistency of the PopSyn inputs in terms of the controls versus the PUMS sample that are analyzed at the PUMA level.
- PopSyn output with respect to matching all types of controls at different levels of geography (MAZ, TAZ, PUMA, Meta).
- PopSyn output with respect to uniformity of household expansion factors.
- PopSyn output with respect to uncontrolled variables and multi-dimensional distributions.

Below are examples and corresponding discussion for each dimension.

Consistency of the PopSyn Inputs

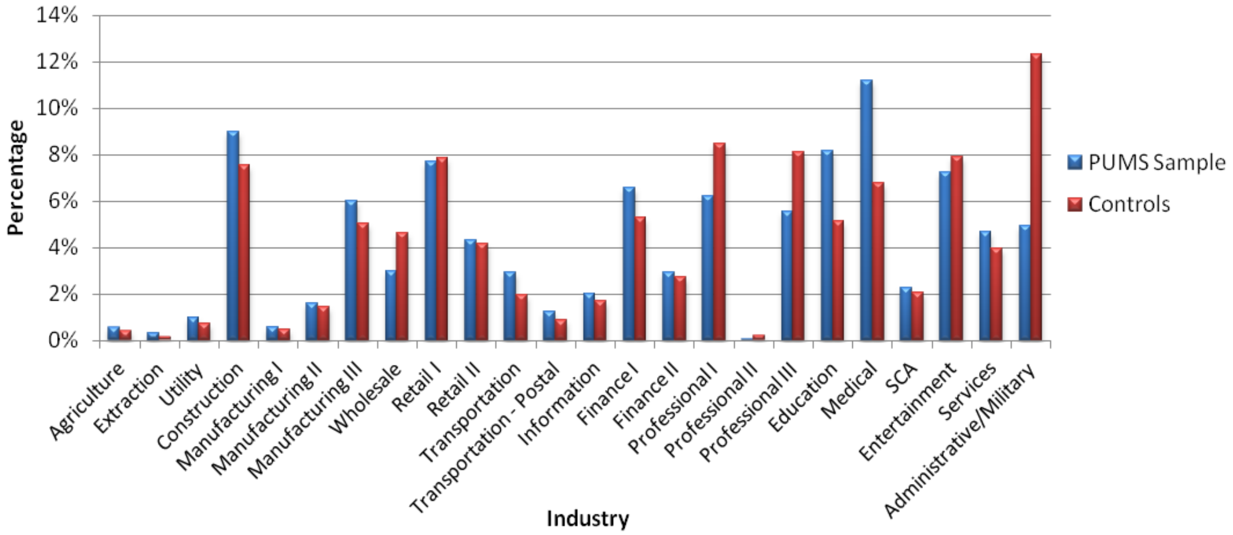
Consistency of the PopSyn inputs in terms of the controls versus the PUMS sample can be analyzed at the PUMA level. It is usual to have certain discrepancies in the household distribution (for example, household distribution by size) defined by the controls versus the distribution obtained from the PUMS sample. However, if discrepancies are substantial for the base year it should be specifically scrutinized and explained; it should be understood that large discrepancies would result in extreme expansion factors. Most of the controls were set very consistently with the proportions observed in the PUMS sample as shown in **Figure 2** for an example of household distribution by size for PUMA 103.

Figure 2: Consistency of Controls with the PUMS Sample (example for PUMA 103)



However, for the regional "meta" controls for number of workers by industry, there were several differences compared to the PUMS distribution as shown in **Figure 3**. These differences were introduced intentionally to account for the structural shifts in employment by industry that came from a different source. In this case, it is important to keep in mind that while PopSyn will try to match these controls it may result in non-uniform expansion factors and other derived effects that may result in less perfect match for some other controls as well as uncontrolled variables.

Figure 3: Consistency of Controls with the PUMS Sample (Regional Number of Workers by Industry)



Matching Controls

PopSyn output was first scrutinized with respect to matching all types of controls at different levels of geography (MAZ, TAZ, PUMA, Meta). In general, the developed PopSyn matches most of the controls very closely. The Rout Mean Square Error (RMSE) statistics calculated across all 3,009 TAZs and 26, 391 MAZs are summarized in **Table 2**.

Table 2: Root Mean Square Error (RMSE) – Control vs. Output

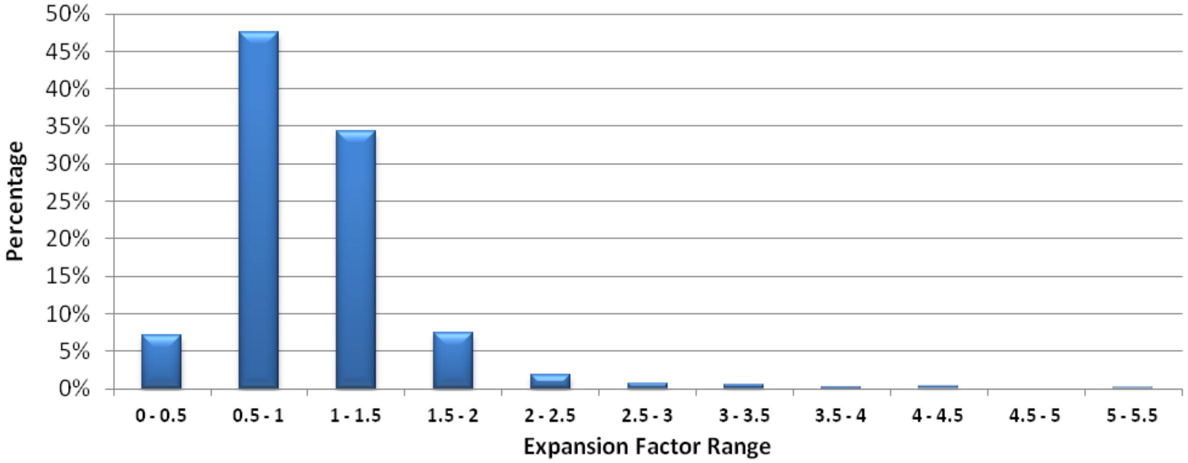
Control	**RMSE**	
	TAZ	**MAZ**
Total HH	0	0
HH Size=1	1.466361	6.02436
HH Size=2	1.814879	5.29567
HH Size=3	0.307219	3.83107
HH Size=4	2.978097	3.74151
HH Income: 0-$21K	0.843327	4.76162
HH Income: $21K-$42K	1.333416	3.89957
HH Income: $42K-$66K	1.088323	3.69059
HH Income: $66K-$102K	1.915115	3.33662
HH Income: >=$102K	3.225954	3.32541
Single Family	0.465135	4.23981
Multi Family	0.465135	4.23981
Age: 0-18 years	9.76039	10.16
Age: 19-35 years	8.585232	8.76668
Age: 36-65 years	6.265957	8.74013
Age: >=66 years	1.742572	6.30158

Logically, the controls with higher priority are satisfied better than controls with lower priority. Matching controls at the TAZ level is exceptional while it is more variation at the MAZ level where it is difficult to match all controls exactly. There were a few exceptional TAZs and MAZs level with relatively large discrepancies that required an individual analysis and explanation. If some controls cannot be matched, it means that they are internally inconsistent between themselves or inconsistent with the PUMS sample. Analysis of relaxation factors calculated by PopSyn allowed for screening and explanation of these cases.

Uniformity of Household Expansion Factors

PopSyn output was scrutinized with respect to uniformity of household expansion factors. This analysis is implemented at the PUMA level where in general we expect to preserve the initial household weights as much as possible. PopSyn is specifically designed to preserve uniformity to the maximum possible extent given the constraints. In general, the results for the MAG area were very good with some exceptions that were explained by the inputs. A typical example of the resulted distribution of expansion factors for PUMA 102 is shown in **Figure 4**. Ideally, if all controls were perfectly correlated with the sample, all expansion factors for the base year would have been equal to 1.0. However, some controls deviate from the sample proportions, sometimes intentionally like the number of workers by industry discussed above. As the result, certain households have to be over-expanded and some other ones have to be under-expanded. PopSyn implements this expansion with the minimal possible structural changes and, in particular, trying to avoid extreme weights (either very high or very low). This guarantees that the generated synthetic population is statistically representative.

Figure 4: Distribution of Household Expansion Factors (Example for PUMA 102)



Replication of Uncontrolled Variables

PopSyn output was also analyzed with respect to uncontrolled variables and multi-dimensional distributions. This analysis in general is probably the most difficult test for a population synthesizer where a good match to the controls might not be helpful. This analysis is limited to the published

Census tables that provide data on the observed household or population distributions not used directly as an input in the population synthesis. An example of such analysis for a two-dimensional distribution of households by size and number of workers at the TAZ level (TAZs correspond to the Census Tracts where the observed distribution is available), PUMA level, and regional level was implemented. The results showed a reasonable match as illustrated in **Table 3** for the regional level.

Table 3: Uncontrolled Joint Distribution by Household Size and Number of Workers

Household size	**Variable**	**Output**	**Census**	**Difference**
All	Total HHs	1,540,010	1,535,588	0.288%
	HH Workers = 0	396,440	381,978	3.786%
	HH Workers = 1	620,011	623,439	-0.550%
	HH Workers = 2	423,135	435,286	-2.791%
	HH Workers >= 3	99,941	94,885	5.329%
HH Size = 1	HH Workers = 0	177,942	172,196	3.337%
	HH Workers = 1	220,626	229,828	-4.004%
HH Size = 2	HH Workers = 0	164,172	164,468	-0.180%
	HH Workers = 1	156,122	171,034	-8.719%
	HH Workers = 2	170,992	192,442	-11.146%
HH Size = 3	HH Workers = 0	24,693	20,395	21.074%
	HH Workers = 1	88,894	83,316	6.695%
	HH Workers = 2	98,303	91,591	7.328%
	HH Workers >= 3	29,892	29,249	2.198%
HH Size >= 4	HH Workers = 0	29,497	24,919	18.372%
	HH Workers = 1	154,041	139,261	10.613%
	HH Workers = 2	153,626	151,253	1.569%
	HH Workers >= 3	70,025	65,636	6.687%

It is important to analyze these results in the context of previous validation tests. On one hand, all else being equal it is desirable to replicate the uncontrolled variables as close as possible. However, an exact match is only realistic if all the previous dimensions are fully satisfied. As was explained above, some controls were intentionally set differently from the observed proportions in order to stir the procedure in a certain direction that was justified by the newer employment data compared to the census. Under these conditions, we should expect that the other dimensions will also be affected since there are many internal correlations embedded in the sample joint distribution and it is in general impossible to adjust one marginal distribution without affecting other marginal distributions. More specifically, it can be seen that changing the proportions of workers by industry and somewhat by household size results in certain structural shifts in the household distribution by number of workers and even bigger structural shifts in the joint distribution of households by size and number of workers. It is not a problem per se but rather an important consideration for analysis and evaluation of the synthetic population.

Conclusions

The paper presents how the population synthesis procedure can be extended to incorporate several innovative features. The first one relates to a general formulation of convergence of the balancing procedure with imperfect (i.e. not fully consistent) controls. The second one relates to the optimized discretizing of the fractional outcomes of the balancing procedure to form a list of discrete households. The third one relates to multiple levels of geography where the controls can be set. The developed PopSyn has been intensively tested and validated for the MAG region with overall a very good level of performance, and across the following particular dimensions:

- Consistency of the PopSyn inputs in terms of the controls versus the PUMS sample that are analyzed at the PUMA level. It is usual to have certain discrepancies in the household distribution (for example, household distribution by size) defined by the controls versus the distribution obtained from the PUMS sample. However, if discrepancies are substantial for the base year it should be specifically scrutinized and explained; it should be understood that large discrepancies would result in extreme expansion factors.
- PopSyn output with respect to matching all types of controls at different levels of geography (MAZ, TAZ, PUMA, Meta). In general, the developed PopSyn matches most of the controls very closely. However, there were exceptions that required an individual analysis and explanation. If some controls cannot be matched, it means that they are internally inconsistent between themselves or inconsistent with the PUMS sample. Analysis of relaxation factors calculated by PopSyn allowed for screening and explanation of these cases.
- PopSyn output with respect to uniformity of household expansion factors. This analysis is implemented at the PUMA level were in general we expect to preserve the initial household weights as much as possible. PopSyn is specifically designed to preserve uniformity as much as possible given the constraints. In general, the results for the MAG area were very good with some exceptions that were explained by the inputs.
- PopSyn output with respect to uncontrolled variables and multi-dimensional distributions. This analysis was implemented for a two-dimensional distribution of households by size and number of workers at the TAZ level (related to the Census Tract where the observed distribution is available). The results showed a reasonable match.

In the process of PopSyn validation, the authors came to certain general conclusions that might change the focus of subsequent validation efforts with this type of population synthesis. In general, this proved to be more debugging and/or analysis of controls and sample than PopSyn validation itself. The developed procedure is analytical and repeatable, there is no use of random numbers, and consequently there is no “mystery” associated with the outcome. Very good match to the controls or to the observed distributions from PUMS is comforting but it is not necessarily right. Sometimes the analyst may intentionally want to skew the distribution, especially, constructing future scenarios, and internal inconsistencies between the controls themselves might be inevitable. If the match is not good it is not necessarily wrong but the subsequent analysis is very important. In our experience two questions have to be asked each time if the controls cannot be matched at some level of geography. First, are the controls set consistently? Most frequently, this was the well-identified reason for discrepancy. Secondly, is there a structural discrepancy between the controls and sample and was it intentional? This was the second most frequent problem encountered that also might result in extreme expansion factors.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. 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.
7. 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.
8. Ma (2011) Generating Disaggregate Population Characteristics for Input to Travel Demand Models, Doctoral Dissertation, the University of Florida
9. 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.
10. 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.

11. 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.
12. 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
13. 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.
14. 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.