Population Synthesis in Activity-Based Models

Tabular Rounding in Iterative Proportional Fitting

Abdoul-Ahad Choupani and Amir Reza Mamdoohi

Activity-based travel demand modeling requires socioeconomic microdata of the population under study. Because the acquisition of such data for the entire population is infeasible or highly expensive, techniques such as iterative proportional fitting (IPF) have been applied extensively to estimate such data for the population, synthetically. Despite its many advantages, IPF results in noninteger values instead of integers: fractions of households or individuals are obtained for zones. Although methods have been proposed to produce integers for noninteger tables, seldom has this problem been viewed as tabular rounding. This paper proposes a binary linear programming model for tabular rounding in which the integer-converted table totals and marginal sums perfectly fit the input data obtained from the Census Bureau. Furthermore, the model minimizes distortion to the correlation structure of the household- and individual-level noninteger tables. The model does not bias the joint or marginal distributions of the socioeconomic attributes of population units or the sampling of rare demographic groups (at a significance level of $\alpha = .05$). The empirical comparison of the proposed method with eight existing methods demonstrates that the proposed model outperforms the tested methods. Sensitivity analysis demonstrates that the integer conversion of small values is as significant as it is for large values. In this study, deterministic methods outperform stochastic methods in accuracy and perfect fit to census data. Finally, a scoring and ranking tool is used to reflect concisely the advantages and disadvantages of these methods.

Activity-based travel demand forecasting has received increasing attention from researchers and practitioners. This modeling approach requires microdata of the study area population; these data are usually inaccessible because of confidentiality concerns or costs. The two most widely applied techniques to produce the microdata synthetically are iterative proportional fitting (IPF) and combinatorial optimization (*1, 2*). Other techniques, such as a sample-free method to synthesize the Belgian population (*3*), have also been proposed. Farooq et al. applied Markov chain Monte Carlo simulation to synthesize the population of Switzerland (*4*). IPF, however, has been the workhorse for the synthesis of population because of the approach's many advantages: the need for fewer census data, the computational ease and speed, and the guarantee of convergence (*1, 5–8*).

A disadvantage of IPF is that it produces noninteger values: fractions of households or individuals (*1, 6, 9*). These noninteger values are not a problem for many applications, such as the aggregate zonal-based modeling of travel demand; however, disaggregate agent-based modeling requires integer rather than fractional values. Combinatorial optimization produces integer estimates of households or individuals for a given zone, and therefore some researchers favor this technique (*10*). However, combinatorial optimization needs considerable time for convergence and does not necessarily produce better results at all iterations because of the random selection of agents in the solution algorithms (*6, 11*).

The usefulness of IPF in population synthesis can be enhanced if the integer conversion problem is solved. Lovelace and Ballas proposed four methods of integer conversion and compared those methods with the conventional rounding method (*6*). However, none of those methods, and very few of the existing ones, regarded the problem as a tabular rounding problem.

The aims of this paper are as follows:

1. Propose a tabular rounding method that maintains the aggregate totals and marginal sums, as well as a close similarity in the correlation structure of the estimated tables and reference tables (*8, 12, 13*);
2. Investigate the bias caused by integer conversion in general, and by deterministic methods in particular;
3. Compare stochastic and deterministic methods for integer conversion; and
4. Study the effects of integer conversion on the sampling of rare demographic groups.

The rest of this paper is organized as follows. The following section reviews the literature and familiarizes readers with IPF and the integer conversion problem. The proposed and existing methods are then described, followed by a discussion of the results of the integer conversion methods and a comparison and ranking of their performance. The paper ends with concluding remarks.

LITERATURE REVIEW

A three-way count table (S_{ijk}) cross classifies the variables X, Y, and Z of sampled agents from a large area (region) into I, J, and K categories, respectively. Each cell S_{ijk} is a count of observations classified into the categories i, j, and k of the first, second, and third variables, respectively. However, such counts (n_{ijk}) are not available for the population of small areas (zones) and only the marginal sums (n_{i++}, n_{+j+}, and n_{++k}) are available. Each marginal cell $n_{i++} = \sum_j \sum_k n_{ijk}$ contains the total number of observations in category i of variable X

Department of Civil and Environmental Engineering, Tarbiat Modares University, Gisha Bridge, Tehran, Iran. Corresponding author: A. R. Mamdoohi, armamdoohi@modares.ac.ir.

Transportation Research Record: Journal of the Transportation Research Board, No. 2493, Transportation Research Board, Washington, D.C., 2015, pp. 1–10.
DOI: 10.3141/2493-01

(similarly for n_{+j+} and n_{++k}). Population synthesizers generally use IPF, a well-known statistical technique similar to the Furness technique, for the estimation of n_{ijk}.

The joint distribution is determined, and each cell indicates the number of agents of a certain type who reside in a specific zone. For each cell, its number of agents is selected randomly from the corresponding set of records available in the sample and placed into the zone population. However, these tables are not in integers.

Therefore, a similarly structured table with integers (N_{ijk}) is desired, in which the marginal sums and totals of N_{ijk} correspond to those of n_{ijk} (e.g., $n_{+++} = N_{+++}$, $n_{i++} = N_{i++}$). Synthesizers devise two approaches to deal with noninteger tables. The first approach converts noninteger tables to integers and then accomplishes the selection (*12*). The second approach carries out the integer conversion indirectly. Fitted tables are treated as joint probability mass functions that display the probabilities of observing agents with specific demographics. Then, n (n = table total) Monte Carlo draws with replacements are used to select n agents from the sample records and construct the zone population (*6, 13–16*). The second approach regards all types of agents simultaneously for a selection that results in a longer list of selections and requires more time and memory. Theoretically, agents with higher probabilities are more likely to appear and agents with lower probabilities are less likely to appear in the synthesized population (*6*). However, there is a nonzero chance that an agent with a lower probability (e.g., .002) will be replicated more times than an agent with a higher probability (e.g., .004). However, a disadvantage of the first approach is that the integer-converted table is not the best solution in terms of information discrimination: a value of 0.501 is treated the same as a value of 0.999 (*11*). (See Table 1 for a further comparison of the two approaches.) This paper revisits the first approach (i.e., tables are converted to integer values and then agents are selected from the sample records), which arguably outperforms the second approach.

Conventional rounding (which rounds fractions down to zero if they are smaller than 0.5 and up to one otherwise) minimizes the overall discrepancy between unrounded and rounded values; however, the approach fails to add to the primary total (*17*). The implicit assumption of conventional rounding is that the frequencies of fractional parts smaller and greater than 0.5 are equal and result in additive rounding. Since this assumption does not always hold true, non–additive rounding will occur in most cases.

The construction of additive rounding in multiway tables is much more profound, because rounding should retain not only the additivity of the table total but also the additivity of the marginal sums (*18*). Because for large tables the error in the marginal sums can be very large, sophisticated methods are needed when multiway tables are rounded.

IPF also results in sparse tables (i.e., tables composed of many cells, most of which have low values). Figure 1*a* shows (for the case study of the current paper) that household-level tables contain values that range from small values close to zero to values greater than 100, but 15% of values fall into the narrow interval of [0 1). The integer conversion of these small values is much more sensitive than for large values.

The aim of this paper is to propose a tabular rounding method and compare its performance with eight existing methods. Among the existing methods, an attempt is made to select the most competitive ones. However, to demonstrate that conventional methods (e.g., rounding) may not be appropriate, a few such methods are included in this study.

Because synthesizing population at both household and individual levels is becoming a common practice (*15, 19, 20*), the effects of integer conversion at both levels will be empirically studied and the case of Wyoming will be considered.

METHODOLOGY

Most integer conversion methods truncate noninteger numbers to separate integer parts from those numbers' fractional parts. Integer parts are usually considered as deterministic numbers, which determine how many times agents of a certain type should be cloned and placed into the synthesized population. Then, the remaining fractional parts are used to determine deterministically or stochastically whether an agent of a certain type should be placed into the population one more time or not. When the fractional parts are rounded, they add to the stored integer parts of the numbers to give the total times that an agent should be placed into the zone population.

A discussion of how the existing methods convert fractional parts of numbers into integers now follows. Then, the proposed method is described in detail.

TABLE 1 Comparison of Two Approaches for Integer Conversion of Noninteger Tables

	Integer Conversion Approach		
Criterion	Direct	Indirect	Reference
How the approach treats the fitted tables	Count table	Probability mass function	*1*
Shorter list of selections	✓	✘	*9*
Requiring less time and memory	✓	✘	*9*
Replicating an agent with higher replication count (or probability) more times than an agent with lower replication count (or probability)	✓	✘	*5*
Discrimination information	✘	✓	*6, 11*
Selection of agents with a specific order[a]	✓	✘	na
Possibility of investigating integer conversion and selection effects separately[a]	✓	✘	na

NOTE: ✓ = applies; ✘ = does not apply. na = not applicable.
[a]Criterion proposed by the current paper.

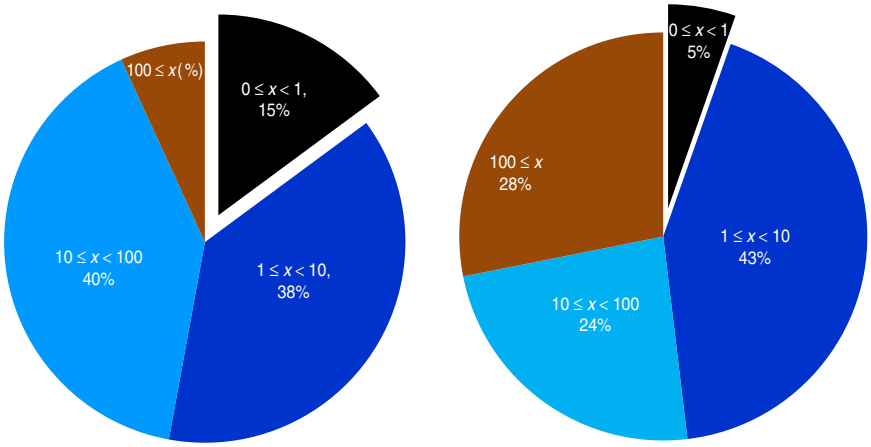

FIGURE 1 Relative frequency of values (*x*) of estimated (*a*) household-level (observations = 8,223) and (*b*) individual-level tables (observations = 6,213) for census tracts in Wyoming.

1. The first method, rounding, refers here to conventional rounding, in which the fractional parts are rounded down if they are smaller than 0.5 and rounded up otherwise.

2. POPGENRANDOM uses a random method of integer conversion (*20–22*):

$$F_{ijk} = \begin{cases} 1 & \text{with probability } \mathrm{pu}_{ijk} = f_{ijk} \\ 0 & \text{with probability } \mathrm{pd}_{ijk} = 1 - f_{ijk} \end{cases} \quad (1)$$

where

F_{ijk} = integer-converted value of f_{ijk},
pu_{ijk} = probability of rounding f_{ijk} up to one,
pd_{ijk} = probability of rounding f_{ijk} down to zero, and
f_{ijk} = fractional part of cell values.

Then, ±1 is added to the rounded cells to account for the differences between the sums of the rounded and unrounded values. If the integer conversion underestimates the table totals, it adds +1 to the rounded-down cells. If the integer conversion overestimates the table totals, it adds −1 to the rounded-up cells.

3. POPGENBUCKET applies bucket rounding to the cells of a specific table (*21*). This deterministic method keeps track of the accumulated rounding errors. The accumulated rounding error in the previous cell of a table is used to bias the rounding of the next cell in the same table. The procedure preserves the populations of agents (table totals) in the zone.

4. POPGENROUNDING first uses conventional rounding (*21*). The sum of the rounded values will not add up to the table total in most cases. Therefore, POPGENROUNDING adds ±1 to the rounded cells in the way described for POPGENRANDOM.

5. TRANSISM applies bucket rounding to specific cells of all tables (*23*). This process preserves the total population of each demographic group in the study area but may slightly change the total population of a given zone (*23*).

6. TRESIS uses a simple rounding procedure, except all nonzero values less than one are rounded up to one (*24*). This method may be biased for values less than one and can yield aggregations that differ substantially from the control values (*5*).

7. ARC and MORPC round up and begin with the demographic groups with the largest fractional part. Rounding up is avoided if it would cause a control value to be exceeded (*25, 26*).

8. Truncate–replicate–sample (TRS) (*6*) is formally defined as

$$F_{ijk} = \begin{cases} 1 & \text{with probability } \mathrm{pu}_{ijk} = \dfrac{f_{ijk}}{\sum_i \sum_j \sum_k f_{ijk}} \\ 0 & \text{with probability } \mathrm{pd}_{ijk} = 1 - \mathrm{pu}_{ijk} \end{cases} \quad (2)$$

The pu_{ijk} is proportional to the share of f_{ijk} in $\Sigma_i \Sigma_j \Sigma_k f_{ijk}$. This method guarantees that higher fractional parts have a higher chance of being rounded up. Also, rounding until $\Sigma_i \Sigma_j \Sigma_k f_{ijk} = \Sigma_i \Sigma_j \Sigma_k F_{ijk}$ ensures that the population size remains the same.

The existing methods are not tabular rounding methods, in which totals and marginal sums of integer-converted tables perfectly fit the input data obtained from the Census Bureau. Tabular rounding also minimizes distortion to the primary noninteger table.

One way of tabular rounding two-way tables is to apply the transportation problem (*18*). Transportation is a well-known problem in economics that aims, originally, to determine a minimal-cost shipping schedule between sources and destinations (*27*). The classic statement of this problem uses a matrix in which the rows represent sources and the columns represent destinations. The margins of the matrix show supplies and demands. The costs of shipping from sources to destinations are indicated by the entries in the matrix.

If the integer conversion of two-way tables that contain positive nonintegers is formulated similarly to the linear programming of the transportation problem (*18*), the integer conversion problem may not have primarily integer solutions. The triangular basis property of the transportation problem is then brought to prove that solutions do exist and that the optimal solutions are also integer if the marginal sums are integer (*28*). This property is still true when the integer values are restricted to sets of {0, 1} through the introduction of capacity constraints on the decision variables.

The above properties of the transportation problem help to find the solution of the tabular rounding of two-way tables without solving an integer-programming model. However, it has been proved that the three-dimensional linear program does not always have integer solutions (*22*).

Thus, an integer-programming model similar to the transportation problem construct is proposed with modifications for the integer conversion of three-way tables, given below in Program A:

$$\text{minimize} \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K} -f_{ijk} * F_{ijk}$$

subject to

$$\sum_{j=1}^{J}\sum_{k=1}^{K} F_{ijk} = \sum_{j=1}^{J}\sum_{k=1}^{K=1} f_{ijk} \qquad i = 1, 2, \ldots, I$$

$$\sum_{i=1}^{I}\sum_{k=1}^{K} F_{ijk} = \sum_{i=1}^{I}\sum_{k=1}^{K} f_{ijk} \qquad j = 1, 2, \ldots, J$$

$$\sum_{j=1}^{J}\sum_{i=1}^{I} F_{ijk} = \sum_{j=1}^{J}\sum_{i=1}^{I} f_{ijk} \qquad k = 1, 2, \ldots, K$$

$$F_{ijk} = \{0, 1\}$$

Constraints guarantee that the observed marginal sums do not change as a result of the integer conversion. If marginal sums remain the same (e.g., $f_{++k} = F_{++k}$), the totals of the tables will remain the same ($n_{+++} = N_{+++}$) through integer conversion. Program A minimizes the error introduced into the results of IPF. Like conventional rounding, this program gives a greater chance of being rounded up to higher f_{ijk}. If n_{ijk} is an integer, there is no need to consider it in the program, which is typically solved for each zone and generally has $I \times J \times K$ binary variables and $I + J + K$ equality constraints.

Program A is scalable to the many dimensions. If more dimensions (variables) are used, the rounding problem can still be converted to a binary linear programming model, in which the objective function and constraints are convex. Therefore, the model always has an integer solution, irrespective of the number of dimensions.

The characteristics of the existing methods and the proposed method are summarized in Table 2.

RESULTS AND DISCUSSION

This section compares integer conversion methods for census tracts as zones in the state of Wyoming and uses year 2010 census data. Wyoming had a population of 563,626 persons residing in 226,879 households in 132 census tracts (one census tract had zero households). The marginal distributions of control variables were extracted from Summary Files 1, and the 1% public use microdata sample, which included 5,665 persons residing in 2,475 households, served as the sample from which the reference tables were established.

The household-level control variables were

1. Household size (seven categories),
2. Householder age (seven categories), and
3. Householder race (four categories).

The individual-level control variables were

1. Gender (two categories),
2. Age (seven categories), and
3. Race (four categories).

The household- and individual-level tables contained 196 ($7 \times 7 \times 4$) and 56 ($2 \times 7 \times 4$) cells, respectively.

IPF and all the other methods were coded in MATLAB. To guarantee representative results for the stochastic methods, they were run 30 times, and the best-fit result with the minimum total absolute error (TAE) was selected.

After different methods were applied for integer conversion, the same method was applied for a selection of the households and individuals. Since there was no variation in the methods' selection processes, only the analysis of the integer conversion step is presented in this paper. For the selection of households or individuals, the method proposed by Auld et al. was used for all the methods; this method aims to synthesize population at both levels (households and persons) simultaneously (*16*).

Accuracy

The first statistic used for the comparison of methods was the absolute error. Figure 2 shows the cumulative percentage distributions of the absolute errors of the table totals and marginal sums. The absolute errors of the totals of the household- and individual-level tables can be seen in Figure 2, *a* and *c*; the 100% absolute errors were zero in the proposed, TRS, and POPGEN methods [note the single triangle in the Cartesian coordinate (0,100) of the figure]. ARC was the second-best method, since 90% and 100% of the absolute errors of the household- and individual-level tables, respectively, were equal to or less than one.

Marginal sums classify the population of agents into distinct groups. As Figure 2, *b* and *d*, shows for household- and individual-level tables, respectively, 100% of the absolute errors of marginal sums were zero in the proposed method. This result was attributable to keeping all the marginal sums the same during integer conversion through the use of a linear constraint in the optimization problem. It can also be seen in the figure that around 90% of the absolute errors of the household- or individual-level marginal sums were zero in ARC. This good performance was a result of rounding up, starting with the demographic groups with the largest fractional part, but avoiding rounding up if it would cause a control value to be exceeded.

The TAE was used to measure the aggregate errors of a table, as shown below (*30*):

$$\text{TAE} = \sum_{i}\sum_{j}\sum_{k} \left| N_{ijk} - n_{ijk} \right| \tag{3}$$

where N_{ijk} is the integer-converted value of cell *ijk* and n_{ijk} is the noninteger value of cell *ijk*.

Figure 3, *a* and *b*, shows that the lowest TAEs belong to methods that apply conventional rounding (including rounding and POPGENROUNDING), since they choose the closest integers to the fractional values. However, methods of stochastic rounding (including TRS and POPGENRANDOM) or bucket rounding (including TRANSIMS and POPGENBUCKET) have the largest TAEs. The errors of the remaining methods (including the proposed method, TRESIS, and ARC) are moderate.

TABLE 2 Existing and Proposed Integer Conversion Methods and Their Characteristics

Method	Synthesizer	Country	Integer Conversion Method	Perfect Fit to Table Totals		Perfect Fit to Marginal Values?	Reference	Scoring and Ranking with TOPSIS	
				Obtained?	How?			Weight	Rank
Rounding	TRANSISM 3.0 and SIMBRITAIN	U.S. and UK	Conventional rounding	✘	✘	✘	*6, 29*	0.5272	7
POPGENRANDOM	POPGEN	U.S.	Random rounding of each cell separately	✓	Adding ±1 to integer-converted cells	✘	*21*	0.5407	6
POPGENBUCKET	POPGEN	U.S.	Applying bucket rounding to cells of a specific table	✓	Inherently results in perfect fit	✘	*21*	0.6894	3
POPGENROUNDING	POPGEN	U.S.	Conventional rounding	✓	Adding ±1 to integer-converted cells	✘	*21*	0.6778	4
TRANSIMS	TRANSISM 4.0.10	U.S.	Applying bucket rounding to specific cell of all tables	✘	✘	✘	*23*	0.5272	7
TRESIS	TRESIS	Australia	Rounding up values lower than 1 and rounding other cells	✘	✘	✘	*24*	0.5272	7
ARC	ARC and MORPC	U.S.	Rounding up cells with the largest fractional parts if they do not exceed marginals	✘	✘	✘	*25, 26*	0.6933	2
TRS	SIMBRITAIN	UK	Random rounding of all cells simultaneously	✓	Inherently results in perfect fit	✘	*6*	0.6056	5
The proposed optimization	na	na	Binary linear programming	✓	Inherently results in perfect fit	✓	na	0.8437	1

NOTE: UK = United Kingdom.

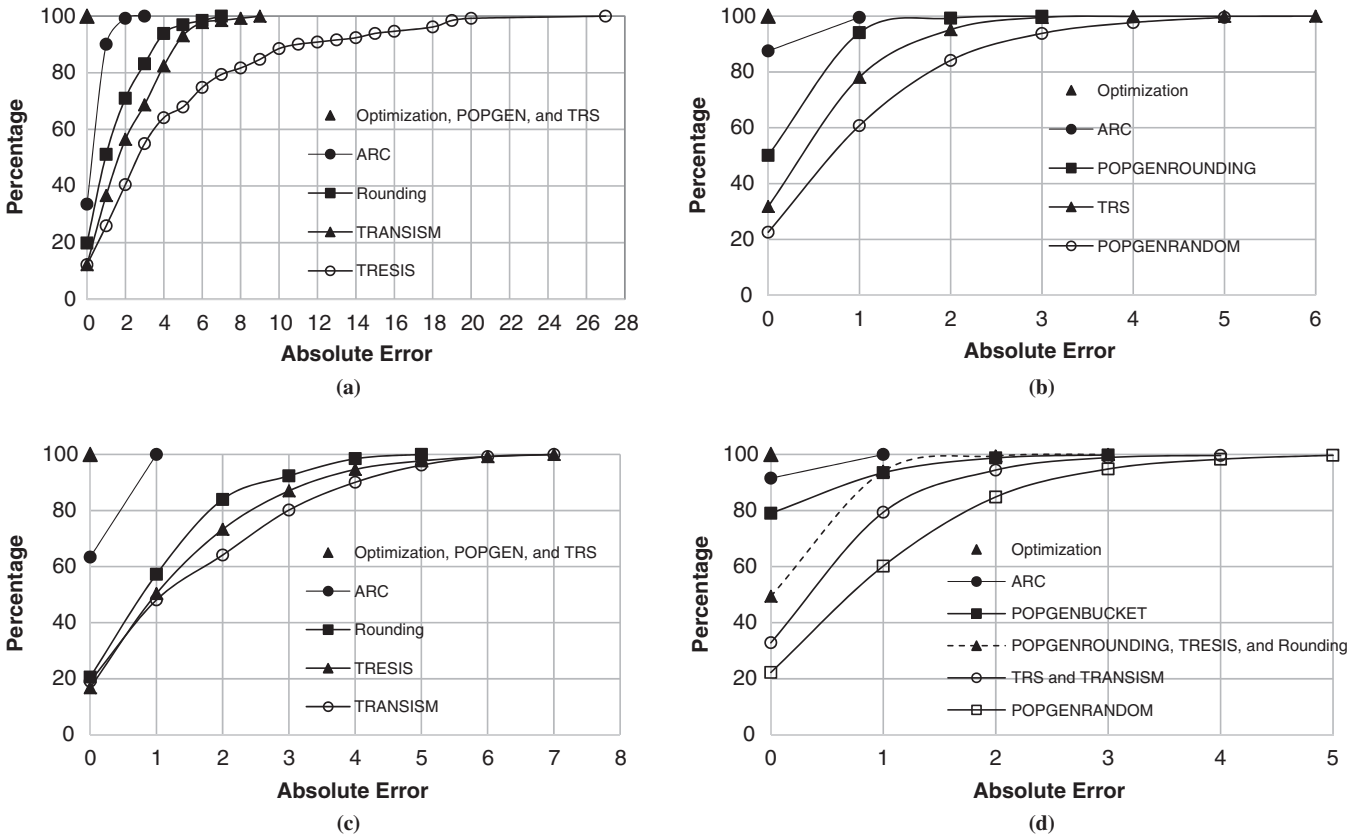

FIGURE 2 Cumulative percentage distributions of absolute errors of household-level (*a*) table totals (observations = 131) and (*b*) marginal sums (observations = 2,358), as well as individual-level (*c*) table totals (observations = 131) and (*d*) marginal sums (observations = 1,703).

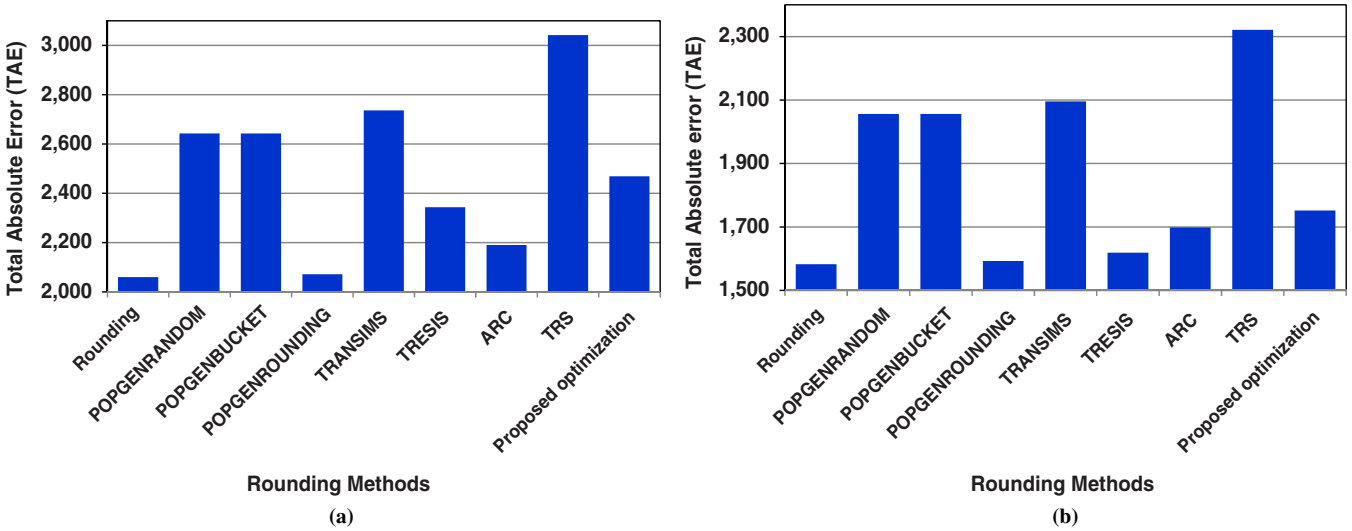

FIGURE 3 Sum of TAEs of (*a*) household-level tables (observations = 131) and (*b*) individual-level tables (observations = 131).

To shed more light on the performance of deterministic and stochastic methods, POPGENROUNDING is compared with the two stochastic methods, POPGENRANDOM and TRS.

Both POPGENRANDOM and POPGENROUNDING convert integers for each cell independently from other cells. The former method randomly decides to round each cell to the closest integer; the latter method rounds the cells deterministically to the closest integer. To account for the difference between the sum of the rounded values and the sum of the unrounded values, both methods add ±1 to rounded cells. TRS randomly rounds all cells simultaneously through Equation 2 and does not need further adjustments to preserve the table totals.

Figure 2, *a* and *c*, shows that for the three methods, 100% of the absolute errors of household- and individual-level totals are zero. However, the errors of the marginal sums are not similar. Figure 2, *b* and *d*, shows that around 95% of the absolute errors of household- and individual-level marginal sums are equal to or less than unity in POPGENROUNDING. In TRS and POPGENRANDOM, however, 85% and 60%, respectively, of the absolute errors of the marginal sums are equal to or less than unity.

Furthermore, POPGENROUNDING results in lower TAEs compared with POPGENRANDOM and TRS. The sums of the TAEs of household- and individual-level tables are approximately 2,050 and 1,550, respectively, in POPGENROUNDING (Figure 3, *a* and *b*). These values are respectively 2,650 and 2,050 in POPGENRANDOM. In TRS, the sums of the TAEs of household- and individual-level tables are approximately 3,050 and 2,300, respectively (Figure 3, *a* and *b*). Figure 3, *a* and *b,* shows that the lowest and highest TAEs belong to POPGENROUNDING and TRS, respectively. POPGENROUNDING outperforms TRS and POPGENRANDOM in this case study. Collectively, the deterministic methods are better than the stochastic methods.

The Freeman–Tukey (FT) test investigates whether the joint (or marginal) distributions of the attributes of agents remain the same through integer conversion (as the null hypothesis) or not (as the alternative hypothesis) (*30*). This test uses the χ^2-type FT statistic, as defined in Equation 4:

$$\mathrm{FT} = 4\sum_i\sum_j\sum_k\left(\sqrt{N_{ijk}} - \sqrt{n_{ijk}}\right)^2 \tag{4}$$

This test may perform poorly when dealing with cell frequencies under five. However, this condition can be relaxed after investigations show that the χ^2 can still be a good approximation for the FT statistic, even when some cell values are lower than one (*31*). In addition, this test might perform poorly for sparse tables that contain both very small and moderately large values (*32*). However, as can be seen in Figure 1, *a* and *b*, the household- and individual-level tables contain values that range from close to zero to greater than 100, and the table cells include a wide range of small, moderate, and large values. Therefore, χ^2 can be a good approximation for the FT statistic, even when some cell values are smaller than one.

The results of the FT test (Tables 3 and 4) show that the null hypothesis is not rejected for the proposed method. This means that the proposed integer conversion method does not change the joint (or marginal) distributions of agents' attributes and does not enter bias into the synthesis process. The FT test (Tables 3 and 4) demonstrates that only two of the existing integer conversion methods change the joint (or marginal) distributions of attributes (at significance level of $\alpha = .05$). TRESIS changes (at significance level of $\alpha = .05$) the marginal distributions of householder age and race in two and seven census tracts, respectively (collectively, nine census tracts). POPGENRANDOM also changes the joint distributions of individuals' attributes in three census tracts (at a significance level of $\alpha = .05$). These 12 census tracts are less-populated zones whose populations are close to the minimum population or below the average.

It is desired to ensure that the fit is good overall, with no problems at any point. A cell-by-cell *Z*-test determines whether the probability of observing agents with specific sociodemographics in a specific zone changes as a result of integer conversion (*30*). This test statistic is given by

$$z_{ijk} = \frac{\dot{n}_{ijk} - \dot{N}_{ijk}}{\sqrt{\frac{\dot{N}_{ijk}\left(1 - \dot{N}_{ijk}\right)}{N}}} \tag{5}$$

TABLE 3 Number of Census Tracts in Which Null Hypothesis Is Rejected: Household-Level Tables

	FT Test[a] for Marginal Sums				
Method	Household Size[b]	Householder Age[b]	Householder Race[c]	FT Test[a] for Entire Table[d]	*Z*-Test[a] for Cells[e]
Rounding	0	0	0	0	0
POPGENRANDOM	0	0	0	0	0
POPGENBUCKET	0	0	0	0	0
POPGENROUNDING	0	0	0	0	0
TRANSIMS	0	0	0	0	0
TRESIS	0	2	7	0	0
ARC	0	0	0	0	0
TRS	0	0	0	0	0
Proposed optimization	0	0	0	0	0

NOTE: df = degrees of freedom.
[a] $\alpha = .05$.
[b] $\mathrm{FT}_{\mathrm{critical}} = 12.5916$; df = 6.
[c] $\mathrm{FT}_{\mathrm{critical}} = 7.8147$; df = 3.
[d] $\mathrm{FT}_{\mathrm{critical}} = 133.2569$; df = 108 (6 ∗ 6 ∗ 3).
[e] $Z_{\mathrm{critical}} = \pm 1.960$.

TABLE 4 Number of Census Tracts in Which Null Hypothesis Is Rejected: Individual-Level Tables

Method	FT Test[a] for Marginal Sums			FT Test[a] for Entire Table[e]	Z-Test[a] for Cells[f]
	Gender[b]	Age[c]	Race[d]		
Rounding	0	0	0	0	0
POPGENRANDOM	0	0	0	3	0
POPGENBUCKET	0	0	0	0	0
POPGENROUNDING	0	0	0	0	0
TRANSIMS	0	0	0	0	0
TRESIS	0	0	0	0	0
ARC	0	0	0	0	0
TRS	0	0	0	0	0
Proposed optimization	0	0	0	0	0

[a] $\alpha = .05$.
[b] $FT_{critical} = 3.8416$; df = 1.
[c] $FT_{critical} = 12.5916$; df = 6.
[d] $FT_{critical} = 7.8147$; df = 3.
[e] $FT_{critical} = 28.8693$; df = 18 (1 * 6 * 3).
[f] $Z_{critical} = \pm 1.960$.

where

Z_{ijk} = statistic of Z-test for cell ijk,
$\dot{n}_{ijk}$ = probability of observing agents with specific sociodemographics (ijk) in noninteger table (n_{ijk}/n),
$\dot{N}_{ijk}$ = probability of observing agents with specific sociodemographics (ijk) in integer table (N_{ijk}/N), and
$N = n$= table total.

The Z-test shows that the proposed method does not change the probabilities significantly at the $\alpha = .05$ level. This finding holds true for all the existing methods studied.

Pritchard and Miller remarked that deterministic rounding might bias the estimates, particularly for cells that represent rare characteristics, with a count under 0.5 (*11*). The cell-by-cell Z-test demonstrates that there is no strong evidence that cell proportions change as a result of integer conversion. Since the tables contain values lower than 0.5 (Figure 1), it can be concluded that deterministic methods do not bias the estimates for values lower than 0.5.

Muller and Axhausen stated that any integer conversion method might bias the population synthesis (*9*). However, when the FT and Z-statistics were considered, there was no significant evidence that deterministic or stochastic integer conversion methods biased the results. Since integer conversion is part of the synthesis process, it could be concluded that, generally, integer conversion did not introduce bias into the synthesis process. Only a few methods (including POPGENRANDOM and TRESIS) resulted in limited bias.

Computational Effort

The time spent on the integer conversion of IPF tables as a criterion of computational effort was measured on an Intel Core i3 (2.10-GHz) machine with 4 GB of random-access memory; the machine was running Windows 7.0. All the methods' run times were 1 s, except for the proposed and TRS methods. The proposed method spent, insignificantly, 5 s longer on integer conversion. The single run and multiple runs of the stochastic TRS method (*30*) took 6 and 37 s. The slower time of the proposed method was the result of solving one optimization problem with a relatively large number of binary variables for each zone ($I \times J \times K$ variables). TRS having the slowest time was a result of the random rounding of all cells simultaneously. It seems that the run time is not a concerning issue.

Sensitivity Analysis

The integer conversion of values smaller than one, which represent rare demographic groups, may be much more sensitive than the integer conversion of large values. Therefore, M (a coefficient defined for conducting sensitivity analysis) is introduced into Program A to conduct sensitivity tests to see how the proportions of rare demographic groups may vary as a result of integer conversion:

$$a_{ijk} = \begin{cases} f_{ijk} & \text{if } n_{ijk} > 1 \\ M \times f_{ijk} & \text{if } n_{ijk} < 1 \end{cases} \quad (6)$$

where a_{ijk} is the coefficients of the decision variables in the objective function of Program A.

If $M = 1$, then $a_{ijk} = f_{ijk}$ (as it appears in Program A). $M = 1$ implies that Program A treats the integer conversion of small cell values the same as that of large ones.

However, there is a concern that small values ($n_{ijk} < 1$) are rounded down to zero and that the corresponding demographic groups disappear from the zone population. This situation may cause the bias of not sampling enough population units of the rare groups. Thus, the proportions of rare demographic groups to the zone population are assessed. The assessments show that the proportions remain the same in the study area (at a significance level of $\alpha = .05$). This finding implies that the proposed method, which treated all values similarly, does not result in biased sampling.

As M increases, Equation 6 places higher priority on rounding up values smaller than one, and this process may lead to the biased

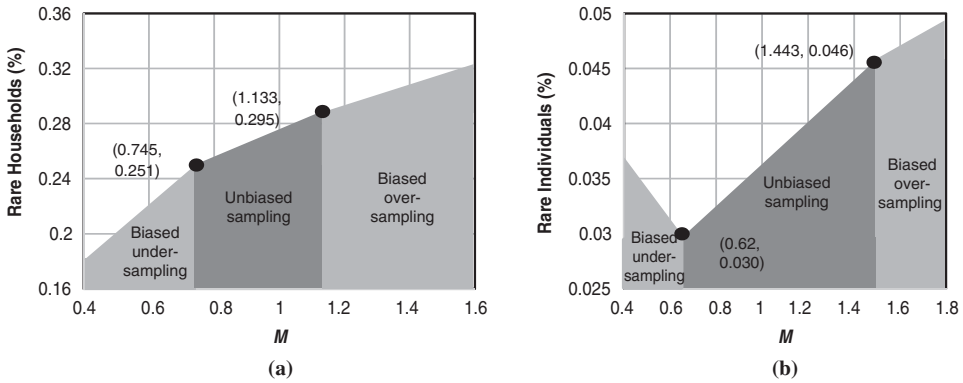

FIGURE 4 Sensitivity of proportions of rare demographic groups to changes in *M* for (*a*) households (observations = 131) and (*b*) individuals (observations = 131).

oversampling of rare types. Decreasing *M* may result in the biased undersampling of rare groups.

Figure 4 shows the changes in the proportions of rare groups with changes in *M*. Figure 4*a* shows for household-level tables that if *M* varies from 0.745 to 1.133, the sampling of rare households is not biased (see the darkened trapezoid in the middle of the figure). If $M > 1.133$, however, then the proportion of rare households increases significantly (at the significance level of $\alpha = .05$) and will result in the biased oversampling of rare households (see the lighter trapezoid on the right-hand side of Figure 4*a*). However, for $M < 0.745$, the proportions of rare households will decrease significantly (at the significance level of $\alpha = .05$), and this situation results in the biased undersampling of rare households (see the lighter trapezoid on the left-hand side of Figure 4*a*). Therefore, for an unbiased sampling, *M* should be between 0.745 and 1.133. Deviating *M* slightly from one results in a biased sampling. So the best value for *M* is one, for which the integer conversion of all values is treated the same.

Scoring and Ranking

Since several criteria are used to compare the integer conversion methods, a multiple-attribute decision-making tool should be applied to assess the overall performance of the methods and rank them. Multiple-attribute decision making is a subdiscipline of operations research and is the study of evaluating and choosing alternatives on the basis of the values (and preferences) of the decision maker. Multiple-attribute decision-making techniques are divided into two classes: compensatory and noncompensatory (*33*). The compensatory techniques allow trade-offs between criteria, in which a poor result in one criterion can be negated by a good result in another criterion. The noncompensatory techniques do not allow trade-offs between criteria. Given that all methods show strengths and weaknesses according to the criteria taken into account, a compensatory approach should be deployed for the assessments and ranking (*34*). There are many compensatory decision-making techniques in the literature; these techniques include the technique for order preference by similarity to ideal solution (TOPSIS), the analytic hierarchy process (AHP), *elimination et choix traduisant la realité* (ELECTRE), and simple additive weighting (SAW) (*33*). For brevity, only TOPSIS is explained here. Interested readers are referred to Tzeng and Huang for details of other multiple-attribute decision-making techniques (*33*). Of the above methods, TOPSIS is selected for the following two reasons:

1. TOPSIS, unlike AHP and SAW, does not require expert judgment for comparison or weighting, and thus the decision-making process cannot be biased by subjective judgments.
2. TOPSIS is more famous than methods such as ELECTRE and SAW.

TOPSIS is based on the concept that the chosen alternative should have the shortest Euclidean distance from the positive ideal solution (which has the best score in each criterion) and should have the farthest Euclidean distance from the negative ideal solution (which has the worst score in each criterion) (*33*).

Alternatives that are closer to the positive ideal solution are more similar to the positive ideal solution and receive higher ranks. The TAEs of table totals, marginal sums, and cells were the three uncorrelated criteria applied for comparison. Inferential statistics were not considered for ranking because all the integer conversion methods were nearly similar when these statistics were taken into account for comparison.

The results of TOPSIS (summarized in Table 2) show that the proposed deterministic method has the highest rank, and ARC takes the second place. The stochastic methods TRS and POPGENRANDOM are ranked fifth and sixth, respectively.

Although the best-performing method is the proposed deterministic method, and other deterministic methods (e.g., POPGEN BUCKET) outperform the stochastic methods, this paper proposes that stochasticity be addressed in the selection stage. When the integer frequency of each cell is obtained, corresponding agents can be drawn randomly from the available sample records. The randomness has been applied in the selection phase, since the exact locations of agents' residences are not known (in a given region). The second reason for random selection is that the agents who are similar in controlled attributes (the attributes used to build tables) may be completely dissimilar in other attributes (uncontrolled variables). Moreover, if desired, stochastic integer conversion can also be applied to produce multiple answers and quantitatively appraise the uncertainty introduced through integer conversion.

CONCLUSION

The proposed binary linear programming model outperforms the other methods of integer conversion tested in the paper and has the highest rank. This method as a tabular rounding method is inspired

by the transportation problem, which stems from economics and was used previously to produce integers for two-way tables. The proposed model decides whether to round fractional parts of cell values up or down while minimizing deviation from IPF noninteger tables, and keeps table totals and marginal sums the same through integer conversion.

Fortunately, hypothesis testing demonstrates that integer conversion methods do not bias the joint or marginal distributions of the attributes of agents in zones. Also, integer conversion methods do not alter the proportions of demographic groups in all zones of a study area significantly or bias the joint or marginal distributions of agents' attributes or the sampling of infrequent types of agent (at a significance level of $\alpha = .05$). Sensitivity tests reveal that the integer conversion of small and large values can be treated similarly in the proposed method.

Furthermore, deterministic and stochastic methods of rounding are compared. The assessments demonstrate that deterministic methods produce better results in accuracy and fitting perfectly to the census data.

Although the proposed method is deterministic and the studied deterministic methods outperform stochastic methods, randomness can be addressed in the selection phase when agents are drawn from the sample and replicated in the zone population. Agents are drawn randomly since their exact place of residence in a given region is not known. Also, agents that are similar in controlled variables may be completely different in uncontrolled variables.

The results provide insight into the advantages and disadvantages of nine integer conversion methods and offer guidance to researchers who aim to use IPF and produce integer values. Existing methods do not consider the integer conversion of tables as the tabular rounding problem. The current paper proposes a tabular rounding method in the context of IPF to be applied in population synthesis for activity-based models.

REFERENCES

1. Beckman, R. J., K. A. Baggerly, and M. D. McKay. Creating Synthetic Baseline Populations. *Transportation Research Part A,* Vol. 30, No. 6, 1996, pp. 415–429.
2. Ryan, J., H. Maoh, and P. Kanaroglou. Population Synthesis: Comparing the Major Techniques Using a Small, Complete Population of Firms. *Geographical Analysis,* Vol. 41, No. 2, 2009, pp. 181–203.
3. Barthelemy, J., and P. L. Toint. Synthetic Population Generation Without a Sample. *Transportation Science,* Vol. 47, No. 2, 2013, pp. 266–279.
4. Farooq, B., M. Bierlaire, R. Hurtubia, and G. Flötteröd. Simulation-Based Population Synthesis. *Transportation Research Part B,* Vol. 58, 2013, pp. 243–263.
5. Bowman, J. L. A Comparison of Population Synthesizers Used in Microsimulation Models of Activity and Travel Demand. http://jbowman.net/papers/. Accessed June 24, 2014.
6. Lovelace, R., and D. Ballas. Truncate, Replicate, Sample: A Method for Creating Integer Weights for Spatial Microsimulation. *Computers, Environment and Urban Systems,* Vol. 39, 2013, pp. 172–181.
7. Fienberg, S. E. An Iterative Procedure for Estimation in Contingency Tables. *Annals of Mathematical Statistics,* Vol. 41, 1970, pp. 907–917.
8. Pukelsheim, F. Biproportional Matrix Scaling and the Iterative Proportional Fitting Procedure. *Annals of Operations Research,* Vol. 215, No. 1, 2014, pp. 269–283.
9. Müller, K., and K. W. Axhausen. Population Synthesis for Microsimulation: State of the Art. Presented at 90th Annual Meeting of the Transportation Research Board, Washington, D.C., 2011.
10. Williamson, P., M. Birkin, and H. Rees. The Estimation of Population Microdata by Using Data from Small Area Statistics and Samples of Anonymized Records. *Environment and Planning A,* Vol. 30, No. 5, 1998, pp. 785–816.
11. Pritchard, R. D., and E. J. Miller. Advances in Population Synthesis: Fitting Many Attributes per Agent and Fitting to Household and Person Margins Simultaneously. *Transportation,* Vol. 39, No. 3, 2012, pp. 685–704.
12. Bowman, J. L. Population Synthesizers. *Traffic Engineering and Control,* Vol. 49, No. 9, 2009, p. 342.
13. Parsons Brinckerhoff, HBA Specto, Inc., and EcoNorthwest. *Oregon2 Model Development: HA Module Description at Finalization.* Oregon Department of Transportation, Salem, 2003.
14. Auld, J., and A. K. Mohammadian. *PopSyn-Win V 4.1 Methodology and Program Documentation.* Chicago Metropolitan Agency for Planning, Ill., 2007. http://www.travelbehavior.com/PopSynWINVersion 4.1 Users Guide.html. Accessed May 26, 2014.
15. Auld, J., and A. Mohammadian. Efficient Methodology for Generating Synthetic Populations with Multiple Control Levels. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2175,* Transportation Research Board of the National Academies, Washington, D.C., 2010, pp. 138–147.
16. Auld, J., A. K. Mohammadian, and K. Wies. Population Synthesis with Sub-Region Level Control Variable Aggregation. *Journal of Transportation Engineering,* 2009, pp. 632–639.
17. Causey, B. D., L. H. Cox, and L. R. Ernst. Applications of Transportation Theory to Statistical Problems. *Journal of the American Statistical Association,* Vol. 80, No. 392, 1985, pp. 903–909.
18. Šíma, J. Aunt's Problem: Table Rounding. *Computers and Artificial Intelligence,* Vol. 18, 1999, pp. 175–189.
19. Guo, J. Y., and C. R. Bhat. Population Synthesis for Microsimulating Travel Behavior. In *Transportation Research Record: Journal of the Transportation Research Board, No. 2014,* Transportation Research Board of the National Academies, Washington, D.C., 2007, pp. 92–101.
20. Ye, X., K. C. Konduri, R. M. Pendyala, B. Sana, and P. Waddell. Methodology to Match Distributions of Both Household and Person Attributes in Generation of Synthetic Populations. Presented at 88th Annual Meeting of the Transportation Research Board, Washington, D.C., 2009.
21. *Synthetic Population Generation for Travel Demand Forecasting.* SIMTRAVEL Research Initiative, Arizona State University, Tempe, 2010. http://urbanmodel.asu.edu/popgen/trainingmaterials.html. Accessed June 24, 2014.
22. Salazar-González, J. J. Controlled Rounding and Cell Perturbation: Statistical Disclosure Limitation Methods for Tabular Data. *Mathematical Programming,* Vol. 105, No. 2–3, 2006, pp. 583–603.
23. *Population Synthesizer: TRANSIMS 4.0.10 User Manual.* Los Alamos National Laboratory, Los Alamos, N. Mex., 2010. http://sourceforge.net/projects/transims/files/documentation/4.0.06/. Accessed June 24, 2014.
24. Hensher, D. A., and T. Ton. TRESIS: A Transportation, Land Use and Environmental Strategy Impact Simulator for Urban Areas. *Transportation,* Vol. 29, No. 4, 2002, pp. 439–457.
25. *Activity-Based Travel Model Specifications: Coordinated Travel–Regional Activity Based Modeling Platform (CT-RAMP) for the Atlanta Region.* Atlanta Regional Commission, Ga., 2012.
26. Parsons Brinckerhoff. *Task 2: Household and Population Synthesis.* Mid-Ohio Regional Planning Commission, Columbus, 2003.
27. Hitchcock, F. L. The Distribution of Product from Several Sources to Numerous Localities. *Journal of Mathematical Physics,* Vol. 20, No. 2, 1941, pp. 224–230.
28. Dantzig, G. B. *Linear Programming and Extensions.* Princeton University Press, Princeton, N.J., 1963.
29. Hobeika, A. *TRANSIMS Fundamentals: Population Synthesizer.* U.S. Department of Transportation, Washington, D.C., 2005.
30. Voas, D., and P. Williamson. Evaluating Goodness-of-Fit Measures for Synthetic Microdata. *Geographical and Environmental Modeling,* Vol. 5, No. 2, 2011, pp. 177–200.
31. Read, T. R., and N. A. Cressie. *Goodness-of-Fit Statistics for Discrete Multivariate Data.* Springer, Berlin, 1988.
32. Koehler, K. J. Goodness-of-Fit Tests for Log-Linear Models in Sparse Contingency Tables. *Journal of the American Statistical Association,* Vol. 81, No. 394, 1986, pp. 483–493.
33. Tzeng, G. H., and J. J. Huang. *Multiple Attribute Decision Making: Methods and Applications.* CRC Press, Boca Raton, Fla., 2011.
34. Jeffreys, I. The Use of Compensatory and Non-Compensatory Multi-Criteria Analysis for Small-Scale Forestry. *Small-Scale Forest Economics, Management and Policy,* Vol. 3, No. 1, 2004, pp. 99–117.

The Standing Committee on Transportation Demand Forecasting peer-reviewed this paper.