Pergamon

Transpn Res.-A, Vol. 30, No. 6, pp. 415–429, 1996
Published by Elsevier Science Ltd. Printed in Great Britain
0965-8564/96 $15.00 + 0.00

PII: S0965-8564(96)00004-3

CREATING SYNTHETIC BASELINE POPULATIONS

RICHARD J. BECKMAN, KEITH A. BAGGERLY and MICHAEL D. McKAY
Statistics Group, Los Alamos National Laboratory, Los Alamos, NM 87545, U.S.A.

(*Received 23 June 1995; in revised form 14 January 1996*)

Abstract—To develop activity-based travel models using microsimulation, individual travelers and households must be considered. Methods for creating baseline synthetic populations of households and persons using 1990 census data are given. Summary tables from the Census Bureau STF-3A are used in conjunction with the Public Use Microdata Sample (PUMS), and Iterative Proportional Fitting (IPF) is applied to estimate the proportion of households in a block group or census tract with a desired combination of demographics. Households are generated by selection of households from the associated PUMS according to these proportions. The tables of demographic proportions which are exploited here to make household selections from the PUMS may be used in traditional modeling. The procedures are validated by creating pseudo census tracts from PUMS samples and considering the joint distribution of the size of households and the number of vehicles in the households. It is shown that the joint distributions created by these methods do not differ substantially from the true values. Additionally the effects of small changes in the procedure, such as imputation of additional demographics and adding partial counts to the constructed demographic tables are discussed in the paper.

INTRODUCTION

Activity-based transportation models, such as those outlined and/or reviewed in Recker *et al.* (1986a,b), Kitamura (1988), Axhausen and Gärling (1991), Bhat and Koppelman, (1993), Gärling *et al.* (1994), and Recker (1994) may require that individual travelers and households rather than aggregates be considered. In particular microsimulations of travel models such as those outlined in Smith *et al.* (1995) are based on the travel behaviors of individual travelers. The purpose of this paper is to outline a methodology for the creation of a synthetic baseline population of individuals and households which can be used in such models. While the populations developed here are necessary for microsimulations in these activity-based models, aggregated demographic characteristics of these populations can also be used in the traditional four step process to estimate travel demand. Any technique for trip generation which is dependent on local demographic characteristics (see for example Papacostas & Prevedouros, 1993, pp. 310–324.) could make use of the methodology for generating demographic tables which is given here.

The technique for construction of the synthetic populations uses 1990 census data given in Census Standard Tape File 3A (STF-3A) (see Census, 1992a) and the Public Use Microdata Sample (PUMS) (see Census, 1992b). STF-3A is a collection of summary tables of demographics, such as the number of persons per household, for census tract or census block group sized areas. It is often used in transportation studies. Most tables in STF-3A summarize one demographic characteristic, but a few cross-classified summary tables of two or three demographics are also given.

The PUMS is beginning to be utilized in transportation studies (for example see Purvis, 1994). The PUMS file consists of a 5% representative sample of almost complete census records (addresses and other unique identifiers are missing) from those contained in a collection of census tracts or other small geographic census areas, which collectively is called a Public Use Micro Area (PUMA). A PUMA is constructed so that it contains approximately 100,000 individuals. Since essentially complete demographics are given for each individual and household in the PUMS, entire multiway summary tables which are not available in STF-3A can be created for the PUMA area.

The procedure for synthetic population generation requires both a collection of summary tables and a PUMS type sample. It does not require however that they be exactly that given for the 1990 census. Summary tables may be added to or deleted from those given in STF-3A, and as long as there are matching variables in the PUMS the procedure will not be invalidated.

The basic algorithm for the construction of a synthetic population is based on census data at either the census tract or block group level. Here, for ease of presentation, we consider only census tract data, but if small area geographic data is needed, block group construction of the population is recommended. Each census tract or part of a census tract which contributes to a given PUMA is considered. Summary tables for a selected set of demographics from STF-3A are assembled for each of the census tracts. The multiway table of these demographics is constructed from the PUMS of the corresponding PUMA. Then, a multiway table is estimated for each census tract where the marginal totals in the constructed tables match the marginal totals given by STF-3A, and the correlation structure in the multiway table constructed from the PUMS is maintained. In the last step of the algorithm, households to makeup the synthetic population are drawn from the PUMS in proportion to the estimated entries in the multiway tables for the census tracts.

The form of the data in STF-3A and the exact tables used for this procedure are given in the next section. Methodologies are illustrated for the creation of the estimated multiway table for each census tract in a PUMA, while techniques for the creation of the synthetic population are given in the following section. Methodologies for the validation of the technique are given and are followed by Discussion and Summary sections.

CENSUS DATA

Census tract summary tables in STF-3A used in the creation of synthetic households can be divided into three types: family households, nonfamily households, and group quarters. The methodology that follows can be applied to households in general without regard to household type. However, in travel activity models it may be important to segregate the households into family households, nonfamily households and group quarters, since activity patterns within these types of households may differ. This approach is presented.

Family households are those households with two or more related persons. Households with persons living alone or related persons living together are nonfamily households. Group quarters are dwellings such as prisons or college dormitories. Since the summary tables in STF-3A are different for each of the three types, the corresponding synthetic populations are generated separately. Family households are considered first. The summary tables in STF-3A which concern family households are:

1. P24: Age of the householder,
2. P107: Family income,
3. P112: Number of workers in the family,
4. P124A&B: Poverty status (which is not used here) × race × family type × presence and age of children.

Not all categories that are given for the above STF-3A tables are used in the procedure. For example, the summary table P107 of family income has 25 categories. By trial and error these were collapsed to 7 categories, as the properties of the resulting populations using all 25 categories are almost identical with those using 7 categories. This does not preclude the use of all 25 categories in practice.

Examples of the four summary tables for census tract 1216.04 in Tarrant County, Texas are given in Table 1. The family class demographic given in Table 1 is derived from the family type and age of children categories given in summary table P124A&B. The 12 family classes are:

1. Married couple: children under age 5 only,
2. Married couple: all children between 5 and 17,
3. Married couple: children under 5 and 5–17,
4. Married couple: no children under 18,
5. Male householder—no wife present: children under age 5 only,
6. Male householder—no wife present: all children between 5–17,
7. Male householder—no wife present: children under 5 and 5–17,
8. Male householder—no wife present: no children under 18,
9. Female householder—no husband present: children under age 5 only,
10. Female householder—no husband present: all children between 5–17,
11. Female householder—no husband present: children under 5 and 5–17,
12. Female householder—no husband present: no children under 18.

The poverty level in summary table P124A&B is not considered. Data in the categories of "below the poverty level" and "above the poverty level" were summed, yielding the resulting race by family class summary table given in Table 1.

Table 1. Family summary statistics from Census Bureau file STF-3A for census tract 1216.04 of Tarrant County, TX

Panel a.	P24: Age of householder						
Age	15–24	25–34	35–44	45–54	55–64	65–74	>74
n	100	445	382	283	164	78	39

Panel b.	P107: Family income						
Income	<$10K	$10–15K	$15–25K	$25–35K	$35–50K	$50–100K	>$100K
n	147	117	216	324	371	267	49

Panel c.	P112: Workers in family			
Workers	0	1	2	>2
n	89	489	792	121

Panel d.	P124A&B Family class × race				
	White	Black	A. Indian	Asian	Other
(1) Couple child <5	73	7	0	0	14
(2) Couple child 5–17	276	23	0	6	18
(3) Couple child'n <5 and 5–17	150	0	0	0	0
(4) Couple no child'n <18	533	0	0	0	0
(5) Male HH child <5	26	0	0	0	0
(6) Male HH child 5–17	0	15	0	0	0
(7) M. HH child'n <5 and 5–17	0	0	0	0	0
(8) Male HH no child'n <18	19	8	0	11	0
(9) Fem. HH child <5	28	13	0	0	0
(10) Fem. HH child 5–17	119	45	0	0	11
(11) F. HH child'n <5 and 5–17	11	0	0	0	0
(12) Fem. HH no child'n <18	85	0	0	0	0

Summary tables in STF-3A for nonfamily households are:

1. P17: Household type and gender,
2. P20: Race × household type × presence and age of children. (The race of nonfamily holders is derived from this table and is the only demographic used from this table.),
3. P24: Age of nonfamily householder,
4. P110: Nonfamily household income,
5. P127: Poverty status (not used here) × age of householder × household type.

These tables for census tract 1216.04 in Tarrant County, Texas are shown in Table 2. Once again some of the categories shown in Table 2 are collapsed from the full set of categories given in STF-3A. However, the three categories of age shown for the P127 summary table are exactly those given in STF-3A.

There are only two summary tables for group quarters in STF-3A. These are:

1. P40: Group quarters,
2. P41: Group quarters × age.

These two tables for census tract 1216.04 of Tarrant County, Texas are given in Table 3.

Table 2. Nonfamily summary statistics from Census Bureau file STF-3A for census tract 1216.04 of Tarrant County, TX

Panel a.	P17: Household type × gender	
Type/gender	Male	Female
Living alone	243	403
Not living alone	120	61

Panel b.	P20: Race of householder				
Race	White	Black	Am. Indian	Asian	Other
n	793	34	0	0	0

Panel c.	P24: Age of householder						
Age	15–24	25–34	35–44	45–54	55–64	65–74	>74
n	139	146	105	97	123	74	143

Panel d.	P110: Household income						
Income	<$10K	$10–15K	$15–25K	$25–35K	$35–50K	$50–100K	>$100K
n	250	69	184	140	75	101	8

Panel e.	P127: Age × household type		
Type/Age	15–64	65–74	>74
Living/alone	439	64	143
Not living alone	171	10	0

Table 3. Group quarters summary statistics from Census Bureau file STF-3A for census tract 1216.04 of Tarrant County, TX

Panel a.	P40: Persons in groups quarters	
Institutionalized:		
	Correctional institutions	0
	Nursing homes	211
	Mental hospitals	0
	Juvenile institutions	0
	Other	0
Other:		
	College dormitories	0
	Military quarters	0
	Emergency shelter	0
	Visible in street	0
	Other	0

Panel b.	P41: Group quarters × age		
Type/age	<18	18–64	>64
Institutionalized:	0	0	211
Other:	0	0	0

The second major source of census data used here is the PUMS. The samples given in the PUMS are a representative (not necessarily random) 5% sample of households and group quarters from the PUMA weights are assigned to each household and person in the sample so that weighted summary statistics can be formed.

In the technique presented here, weighted multiway summary tables corresponding to the demographic variables and categories in Tables 1, 2 and 3 are formed from the PUMS for family households, nonfamily households and group quarters for each PUMA in the metropolitan study area. The tables are constructed from the PUMS by adding the household weights for households in each category of the multiway table. For group quarters, the person weights are summed for each category.

ESTIMATING THE CROSS-CLASSIFIED TABLE

In this section a method for the construction of cross-classified tables of demographics for each census tract in the area of study is discussed. These multiway tables are constructed to satisfy the marginal summaries of STF-3A. Additionally, the estimate of the correlation structure of the census tract multiway table as given by the PUMS is maintained. For mathematical tractability purposes it is assumed that all of the census tracts and census places that contribute to a PUMA have the same correlation structure. Due to the relationships between the demographics in a local area this is probably not a serious restriction.

Correlation in a multiway table is measured by odds ratios. For a 2×2 table with cell proportions $p_{i,j}$, $i = 1,2$, $j = 1,2$ the odds ratio is

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

Odds ratios for multiway tables or $n \times m$ two-way tables are a given by a simple extension of the formula given above for the 2×2 table. For a $n_1 \times n_2 \times \ldots \times n_m$ table the odds ratios have the general form

$$\phi = \frac{(p_{i_1,i_2,\ldots,i_m})\,(p_{i_1,\ldots,i_j+c_1,\ldots,i_k+c_2,\ldots,i_m})}{(p_{i_1,\ldots,i_j+c_1,\ldots,i_k,\ldots,i_m})\,(p_{i_1,\ldots,i_j,\ldots,i_k+c_2,\ldots,i_m})}$$

where

c_1 and c_2 are positive integers such that $i_j + c_1 \leq n_j$ and $i_k + c_2 \leq n_k$.

The primary tool used to complete the multiway table for each census tract is iterative proportional fitting (IPF) (Deming & Stephan, 1940). It can be shown that in situations where the marginal totals of a multiway table are known and a sample from the population which generated these totals is provided IPF gives a constrained maximum entropy estimate of the true proportions in the population multiway table (Ireland & Kullback, 1968). Additionally, IPF estimates maintain the same odds ratios as those in the sample table in the absence of any marginal information to the contrary (see, for example, Little & Wu, 1991).

IPF has been used in transportation modeling. In some circles it is known as the Fratar model (see for example Papacostas & Prevedouros, 1993). It was used by Dugway *et al.* (1976) to synthesize households employing a technique similar to the one presented here.

To describe IPF the following notation is used. Let the proportion of observations in a sample from a m-way marginal table from the PUMS be denoted by

$$p_{i_1,i_2,\ldots,i_m} = n_{i_1,i_2,\ldots,i_m}/n$$

where

$i_j = 1,2,\ldots,n_j$ represents the observed value of the jth demographic with n_j categories (for example age of the householder in Table 1 has $n_j = 7$ categories), n is the total number of observations in the table and $n_{i_1,i_2,\ldots,i_m}$ is the number of counts in cell $(i_1,i_2,\ldots,i_m)$. Also, let $T_k^{(j)}$ be the known marginal totals for the kth category of the jth demographic. The total number n is

$$n = \sum_{k=1}^{n_j} T_k^{(j)} \text{ for all } j.$$

Let $p^{(t)}_{i_1,i_2,\ldots,i_m}$ denote the estimated proportions in cell $(i_1,i_2,\ldots,i_m)$ at iteration t of the IPF procedure and let

$$p^{(t)}_{\ldots,i_j=k,\ldots} = \sum_{i_1=1}^{n_1} \cdots \sum_{i_m=1}^{n_m} p^{(t)}_{i_1,i_2,\ldots,i_j=k,\ldots,i_m}$$

where the above sums are not over the index corresponding to the fixed category k.

Iterative proportional fitting begins by letting

$$p^{(0)}_{i_1,i_2,\ldots,i_m} = p_{i_1,i_2,\ldots,i_m}$$

At iteration t the estimated proportions $p^{(t)}_{i_1,i_2,\ldots,i_m}$ are derived using the following procedure. For each margin in turn, update the estimated proportions (for all values of i_1, i_2 etc., and the kth category of the jth marginal) by

$$p^{(\text{new})}_{i_1,i_2,\ldots,i_j=k,\ldots,i_m} = p^{(\text{old})}_{i_1,i_2,\ldots,i_j=k,\ldots,i_m} (T_k^{(j)}/n)/p^{(\text{old})}_{\ldots,i_j=k,\ldots}$$

where for the first marginal for $p^{(\text{old})}$ corresponds to $p^{(t-1)}$, the resulting estimates from the last iteration. For the second and later marginals $p^{(\text{old})}$ is set equal to the $p^{(\text{new})}$ estimated for the previous marginal. Finally, $p^{(t)}$ is set equal to $p^{(\text{new})}$ resulting from the last marginal. The iterations continue until the relative change between iterations in each estimated $p^{(t)}_{i_1,i_2,\ldots,i_m}$ is small. In practice we find this procedure converges in 10–20 iterations.

Minor adjustments must be made to the IPF routine in order to handle marginal tables in the forms given in Tables 1, 2 and 3. Two way marginals such as the race × family class table in Table 1 present no problem to the IPF routine. Such marginal tables are converted to a single demographic whose categories are all the combinations of two demographics involved and IPF proceeds as usual. For example the 5 × 12 table for race × family class in Table 1 is considered as a table of one demographic variable with 60 categories. If two marginal tables contain a common demographic variable (e.g. the alone–not alone demographic in tables P17 and P127 in Table 2) the procedure is not altered. The fitting proceeds as above treating each marginal separately. For the case where one demographic variable is in two summary tables and has fewer categories in one than the other (e.g. summary tables P24 and P127 shown in Table 2) an additional step is required. The procedure uses only one marginal table at a time. When the table with the "collapsed"

marginal is considered, the procedure updates the cells as above where all of the cells that contribute to the individual "collapsed" categories are updated by the same proportion.

As stated before it is assumed that each of the census tracts contained in a PUMA has the same set of odds ratios. Without this or similar assumptions the estimation of odds ratios for individual tables is mathematically intractable. Even with this assumption, however, the PUMS sample does not reflect the correlation structure of the individual tables. This is easily seen by considering two 2 × 2 tables with sample sizes n_1 and n_2 and the same odds ratio ϕ. Let the proportions in the cells of the two tables be $p_{i,j}^{(1)}$ and $p_{i,j}^{(2)}$ for $i = 1,2$ and $j = 1,2$

With the assumption of equal odds ratios in the tables we have

$$\phi = \frac{p_{1,1}^{(1)} p_{2,2}^{(1)}}{p_{1,2}^{(1)} p_{2,1}^{(1)}} = \frac{p_{1,1}^{(2)} p_{2,2}^{(2)}}{p_{1,2}^{(2)} p_{2,1}^{(2)}}$$

Now, the odds ratio for the combined tables and thus for the sample is

$$\phi' = \frac{p_{1,1}\, p_{2,2}}{p_{1,2}\, p_{2,1}}$$

where

$$p_{i,j} = (n_1 p_{i,j}^{(1)} + n_2 p_{i,j}^{(2)})/(n_1 + n_2).$$

From the above equations it is easy to see that in general $\phi' \neq \phi$.

The fact that $\phi' \neq \phi$ implies that correct statistical techniques for the estimation of ϕ must simultaneously take into account the data from all of the census tracts or parts of census tracts that contribute to the PUMS.

The following two step IPF procedure considers all census tracts and parts of tracts that contribute to the PUMS. It results in estimated multiway tables for which the odds ratios in each table are identical, and when the tables are combined the odds ratios are the odds ratios of the PUMS. First, the marginal tables for all n census tracts in the PUMA are added. Then an m-dimensional multiway table denoted by TO is obtained by IPF of the PUMS against the summed marginals. The second step may be viewed as the construction of an $(m+1)$-dimensional table. The first m dimensions of the table are the m marginals. The last dimension is created by "stacking" the n tables for the census tracts. IPF is used a second time where the known marginal tables are the combined marginals for the n tables and TO is taken as an additional marginal table for the "stacked" tables. The final estimated tables are obtained by IPF of a $(m+1)$-dimensional table with entries of 1 against these marginal tables. The following example of two 2×2 tables illustrates this procedure.

Suppose that a PUMA consists of two "census tracts" and that there are two 2×2 tables with known marginals but unknown cell proportions as shown in Tables 4(a) and (b). These two tables correspond to census summary tables from STF-3A. The total marginal table formed by adding the entries in Tables 4(a) and (b) is in Table 4(c). The PUMS for the two "census tracts" is given in Table 4(d). IPF of the "PUMS" [Table 4(d)] against the marginal totals table [Table 4(c)] gives table TO shown in Table 4(e).

TO is used as one of the marginal tables for a 2×2×2 table constructed by "stacking" the two original tables on top of one another. The three variables of the 2×2×2 table are: v_1, v_2 and the table number. Two additional 2×2 marginal tables are those given by variable v_1 × table number and v_2 × table number. These are shown in Tables 4(f) and (g).

To obtain the final estimated tables, the three 2×2 marginal tables are used as marginals in the IPF of a 2×2×2 table with cell entries equal to 1. The resulting IPF estimates are given in Tables 4(h) and (i).

Each of the above estimated tables has an odds ratio of 0.21; the odds ratio of these two tables combined is 0.24, which is also the odds ratio of the PUMS. If the two tables are fitted individually to the PUMS by IPF, the odds ratio of each table is the same as the PUMS odds ratio, 0.24. The odds ratio of the combined tables is then 0.28 which is not the odds ratio of the PUMS.

Table 4. Subtables for the IPF example

Panel a.	Table 1		
	$v2=1$	$v2=2$	Total
$v1=1$	?	?	1700
$v1=2$	?	?	1050
Total	1505	1245	2750

Panel b.	Table 2		
	$v2=1$	$v2=2$	Total
$v1=1$	?	?	1405
$v1=2$	?	?	905
Total	700	1610	2310

Panel c.	Table-1 + table-2		
	$v2=1$	$v2=2$	Total
$v1=1$	?	?	3105
$v1=2$	?	?	1955
Total	2205	2855	5060

Panel d.	PUMS		
	$v2=1$	$v2=2$	Total
$v1=1$	45	108	153
$v1=2$	63	37	100
Total	108	145	253

Panel e.	TO		
	$v2=1$	$v2=1$	Total
$v1=1$	949	2156	3105
$v1=2$	1256	699	1955
Total	2205	2855	5060

Panel f.	Marginal: $v1$ × tables		
	table-1	table-2	Total
$v1=1$	1700	1405	3105
$v1=2$	1050	905	1955
Total	2750	2310	5060

Panel g.	Marginal: $v2$ × tables		
	table-1	table-2	Total
$v2=1$	1505	700	2205
$v2=2$	1245	610	1855
Total	2750	1310	5060

Panel h.	Table-1: final estimate		
	$v2=1$	$v2=2$	Total
$v1=1$	701	999	1700
$v1=2$	804	246	1050
Total	1505	1245	2750

Panel i.	Table-2: final estimate		
	$v2=1$	$v2=2$	Total
$v1=1$	248	1157	1405
$v1=2$	452	453	905
Total	700	1610	2310

Panels (a), (b) and (d) represent two "census tracts" and the corresponding PUMS. Panel (c) adds the marginal entries from Panels (a) and (b). Panel (e) is the result of an IPF of Panel (c) to Panel (d). Panels (f) and (g) are variables by table number marginal tables. Panels (h) and (i) are the final fits generated by an IPF with Panels (e), (f) and (g) as marginal tables against a 2×2×2 table of ones.

CREATING THE SYNTHETIC POPULATION

The synthetic population of households is constructed by selecting entire households from the PUMS in proportion to the estimated probabilities given in the multiway table obtained by the IPF technique of the previous section. It is not required that the IPF procedure be used to estimate the multiway table; any scheme to estimate the proportions in the tables may be used. For example, one could assume that each of the marginals is independent of the others and create the multiway table by multiplication of the marginal proportions. Other techniques for estimation of the tables such as maximum likelihood or minimum Chi-square could also be attempted. Some comparisons between the independent assumption and the IPF procedure are given in the Validation section.

The number of households to be generated of each demographic type (having a specific set of demographics) is determined for each census tract. These numbers can be obtained either, by multiplying the total number of households by the probabilities in the estimated multiway table, or by drawing the numbers at random according to these probabilities. The first case requires the addition or deletion of a few households due to rounding.

Once the number of households of each demographic type to be selected is determined, households with different demographics are considered separately. For a combination of demographic characteristics a set of probabilities is assigned to each household in the PUMS (after it has been split into family and nonfamily households and group quarters) where PUMS samples "close" to the combination of desired demographic characteristics are assigned higher probabilities. These probabilities are computed by considering the "distance" between a PUMS household, indicated by p, and the households characterized by a cell, c, in the multiway table for census tract. The following function is used to calculate the probabilities.

$$D(p,c) = w_p \prod_{i \in J} \left(1 - | (d_i^p - d_i^c)/r_i |^k \right) \cdot \prod_{i \notin J} (1 - \Delta (d_i^p, d_i^c))$$

where, for family households

1. J is the set of ordinal variables such as {income, age, workers} for family households and $\sim J$ is the set of categorical variables such as {family class, race} for the family households.
2. d_i^p is the value of the ith demographic for household p from the PUMS,
3. d_i^c is the value of the ith demographic of the household of cell type c from the census tract,
4. r_i is the range of demographic i in the PUMS,
5. w_p is the weight associated with household p from the PUMS,
6. $\Delta (d_i^p, d_i^c) = \begin{cases} \alpha & d_i^p = d_i^c \\ 1 - \alpha & d_i^p \neq d_i^c \end{cases}$
7. k is an arbitrary positive constant. Values of k near zero give most weight to those demographics, d_i^p, which are close to the constructed table demographics, d_i^c.

When $\alpha = 0$ and $k \to 0$ samples from the PUMS are considered only if there is an exact match in the demographics d_i^p and d_i^c. In this case we call $D(p,c)$ a 0–1 loss function.

To acquire households for the synthetic population, the probability of selecting household p for the synthetic population for a household with demographics c is given by

$$Pr\{\text{Selecting Household } p\} = D(p,c)/\sum_j D(j,c).$$

Using IPF with the true marginal tables will create a zero entry in a cell which has no samples from the PUMS. Hence, after IPF there is always at least one sample in the PUMS which exactly matches the table's demographics in cells with nonzero entries. Therefore, the 0–1 loss function of step 7 above may be used to select PUMS households for the synthetic population. Other procedures for estimating the multiway table, such as assuming independence between the marginals and multiplying marginal probabilities to create the table, will not necessarily have this property. In the Discussion section we report on various perturbations to the basic IPF procedure which require a non 0–1 loss function. In all cases the basic procedure with a 0–1 loss function is better.

Since there is randomness in the selection of the households by the above method, multiple populations are constructed for each study. These multiple populations may then be used to investigate the uncertainty in the results of the study which is due to the construction of the synthetic populations.

VALIDATION STUDIES

One method of validating the scheme for creating a synthetic population compares demographic characteristics of the synthetic populations with those of the true population

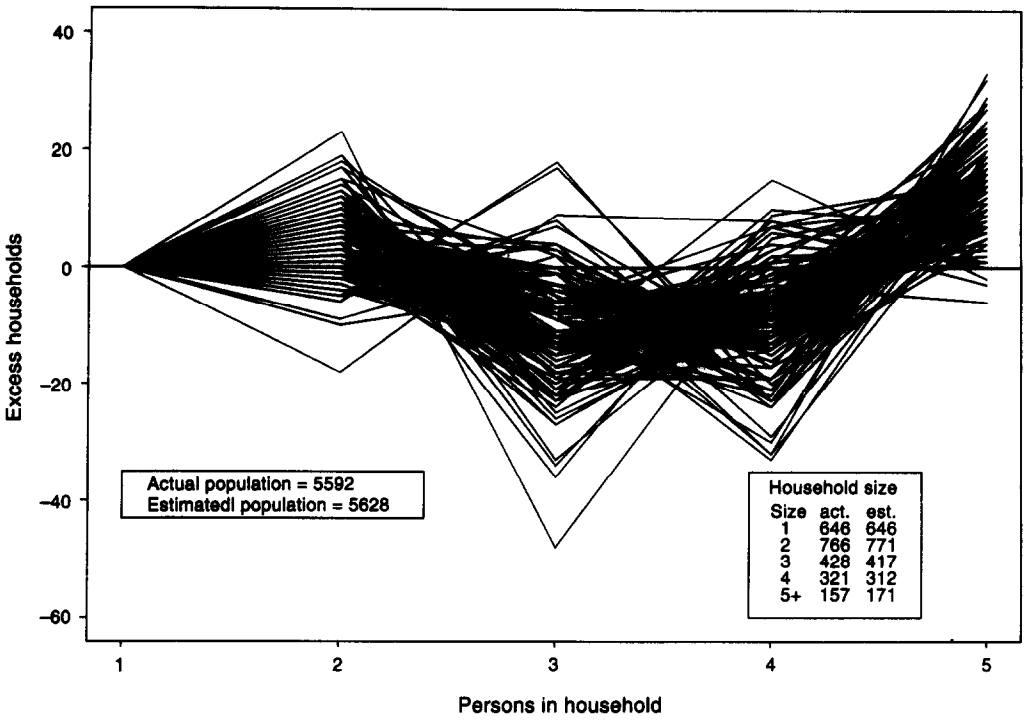


Fig. 1. Distribution of household size for 100 synthetic populations generated for census tract 1216.05 of Tarrant County, Texas. Each line represents one synthetic population and shows the difference between the number of households of a particular size and the true number.

using variables not involved in the generation of the population. For example, to judge the bias and the variance of the generation technique, the distribution of the number of people per household (combining both family and nonfamily households) as summarized in STF-3A can be compared with its distribution in the synthetic population. This variable is available in STF-3A for all households. It is not available for family and nonfamily households individually and hence is not used in the generation of the synthetic family and nonfamily households.

Utilizing a 0–1 loss function, 100 synthetic populations of family and nonfamily households for census tract 1216.04 of Tarrant County, Texas were created. Figure 1 shows the differences in the true distribution of household size for this census tract and those distributions in the synthetic populations. Each line in the figure represents the difference between one of the synthetic populations and the known values for the census tract. The box in the lower right hand of the figure shows the actual distribution of households and the average number of synthetic households for each household size.

The largest discrepancy between the synthetic populations and the actual population is in the number of households with five or more persons where on the average 14 (with a base of 157) too many households of this size are generated. In each population of synthetic households exactly 464 households with exactly one person were produced. The number of generated one person households is always exactly correct with a 0–1 loss function due to the fact that the marginal totals in Table 2 for P17 and P127 are exactly satisfied. The average total population for the synthetic population is 5628 persons, while the true number of persons residing in the census tract is 5592 persons. The difference of only 36 persons is remarkable considering that the total number of persons is not controlled by the procedure to create the synthetic population.

In a second validation scheme a validation population is constructed from the existing PUMS data. The samples in 20–30 PUMS data sets are each considered as the full set of demographics of a "census tract". A "super PUMS" is constructed as a 5% random sample from this collection. Synthetic populations are generated for each of the validation

"census tracts". Since the entire set of demographics is known for the validation population of "census tracts", comparisons of the synthetic population with these demographics can be made.

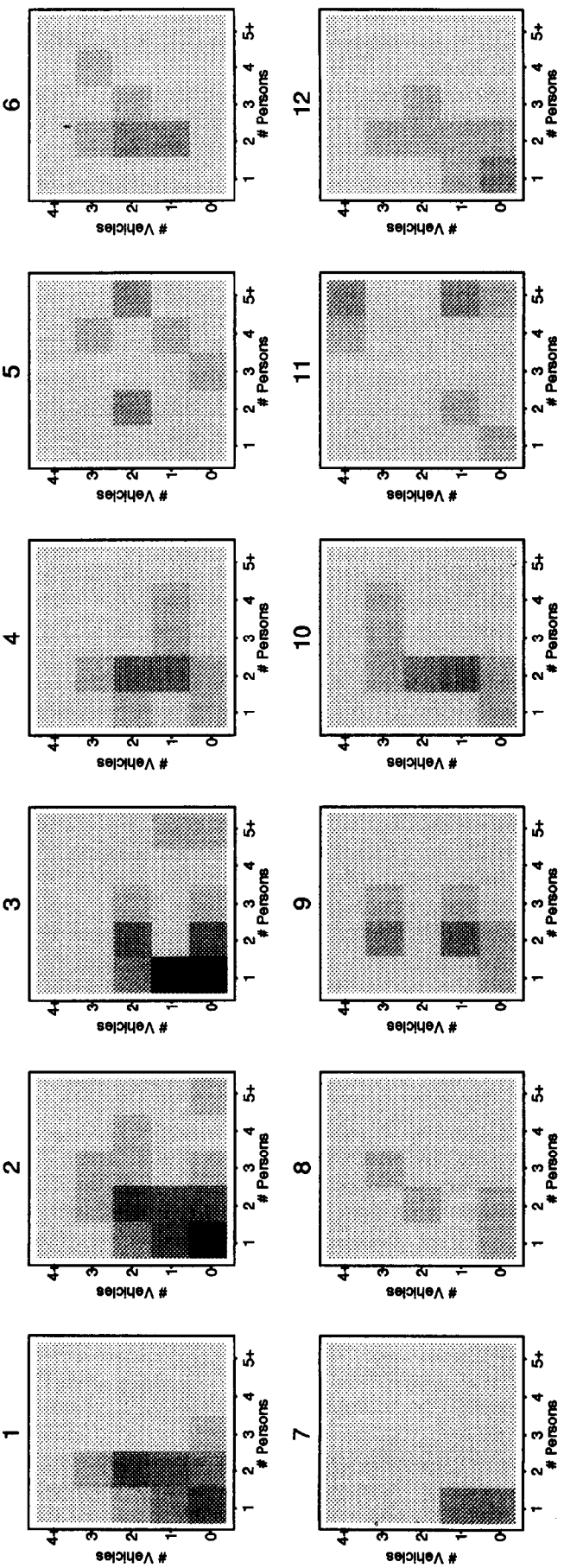
In the study here 22 PUMS were selected from the San Francisco Bay area to serve as a collection of pseudo census tracts. A "super PUMS" was created by taking a 5% sample from the collection of 22 PUMS data sets. The average number of persons and vehicles per household was computed for each synthetic population generated with a 0–1 loss function. The results are given in Fig. 2. For each of the 22 pseudo census tracts, the absolute differences between the proportion of synthetic households of sizes 1–5+ and with 0–3+ vehicles and the true proportions in the population are highlighted. The darker areas on the figures show the most discordant regions. It is seen in this figure that the procedure does well in the prediction of this joint distribution. This most discrepant results are in the first three "census tracts" where the number of vehicles for small household sizes are overestimated. In particular, the proportion of single person households with no vehicles is greatly underestimated. These three "census tracts" were known to be different from the rest as they are the PUMAs from the City of San Francisco, where due to parking difficulties and the availability of a good mass transit system the number of vehicles is lower than in the general population. They were added to the validation set to see if a few heterogeneous census tracts in a PUMA would invalidate the procedure. The results shown in Fig. 2 demonstrate the robustness of the procedure for the majority (and most homogeneous) "census tracts". In practice, however, the true census tracts from a PUMA being more geographically homogeneous are probably more demographically homogeneous, at least in correlation structure, than those constructed in the validation study by combining PUMAs. We expect in practice that the resulting estimated populations will be closer to the true values for all census tracts than shown in the validation study.

One potential methodology for the construction of a synthetic population is to forgo the fitting of the multiway table with IPF and draw the population directly from the PUMS according to the number of households given in STF-3A for block groups or census tracts. The improvement by using IPF for both family and nonfamily households is shown in Fig. 3. In this figure the mean absolute deviation between the proportions of household sizes by the number of vehicles estimated using IPF for the 22 "census tracts" is plotted against the mean absolute deviation of the proportions in the "census tracts" and the true proportions in the "super PUMA". The latter is equivalent to the selection of households directly from the PUMS without IPF. Most points on the two plots in Fig. 3 are above the line. These show the "census tracts" where the IPF routine does a better job in the prediction of the joint distribution of household size and the number of vehicles.

DISCUSSION

We now briefly discuss two additional facets of the procedure. First, the multiway table generated from the PUMS is sparse. There are 11,760 cells in the multiway table for families (obtained by multiplication of the number of categories for each of the marginals) while there are only 609 cells with nonzero entries in the multiway table constructed from the PUMS. The IPF algorithm estimates a zero proportion for all cells that are zero in the sample. Since the PUMS is a sample, many of the approximately 11,000 empty cells might not be empty in the population. Therefore, one may wish to "tweak" each of the empty cells with a partial count of 0.1 or 0.01 before using the IPF routine.

Second, given the estimated multiway table from the IPF routine, the potential exists for the imputation of an additional demographic from the PUMs. For example, one demographic that could be imputed is the number of people in the household which is not one of the demographic characteristics used to construct the multiway tables for either family or nonfamily households. The purpose for imputation is its potential to give more control over a possible important variable in activity prediction such as the household size.


1
2
3
4
5
6
7
8
9
10
11
12
Vehicles
Persons
0
1
2
3
4
5+

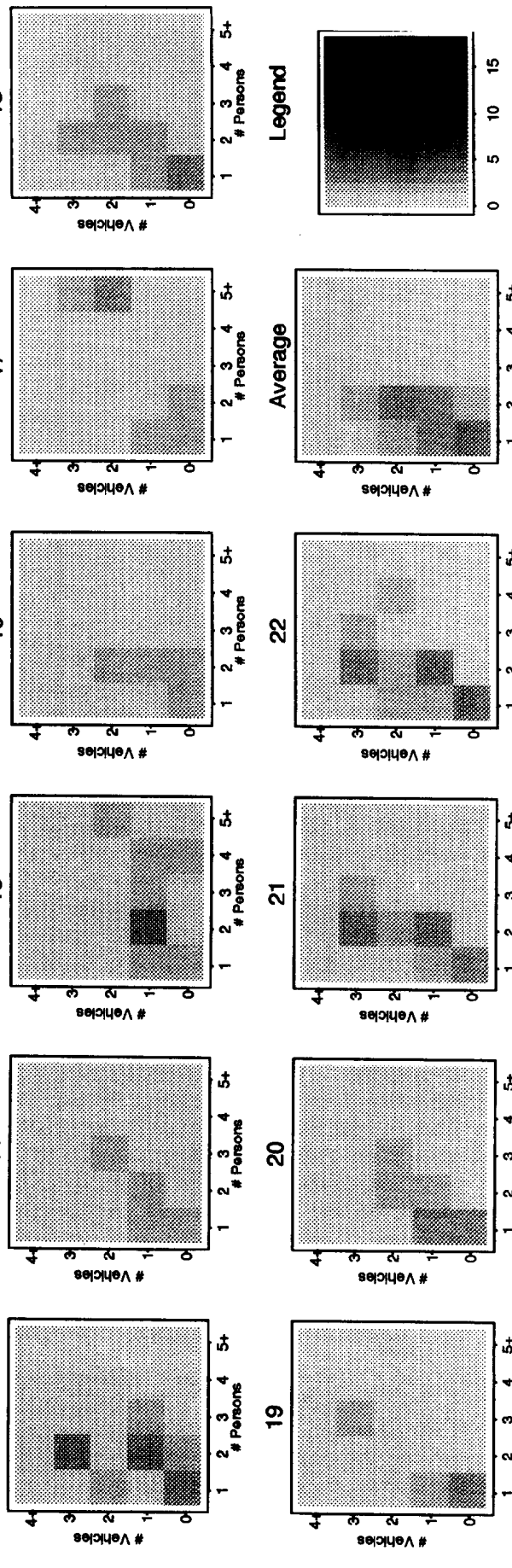


Fig. 2. The difference from the truth in the proportions of household size and number of vehicles for an average synthetic population. Each box on the plot represents 1 of 22 "census tracts" which were constructed from San Francisco Bay area PUMS. The darker the shading in the box the larger the difference between the synthetic population and the truth.

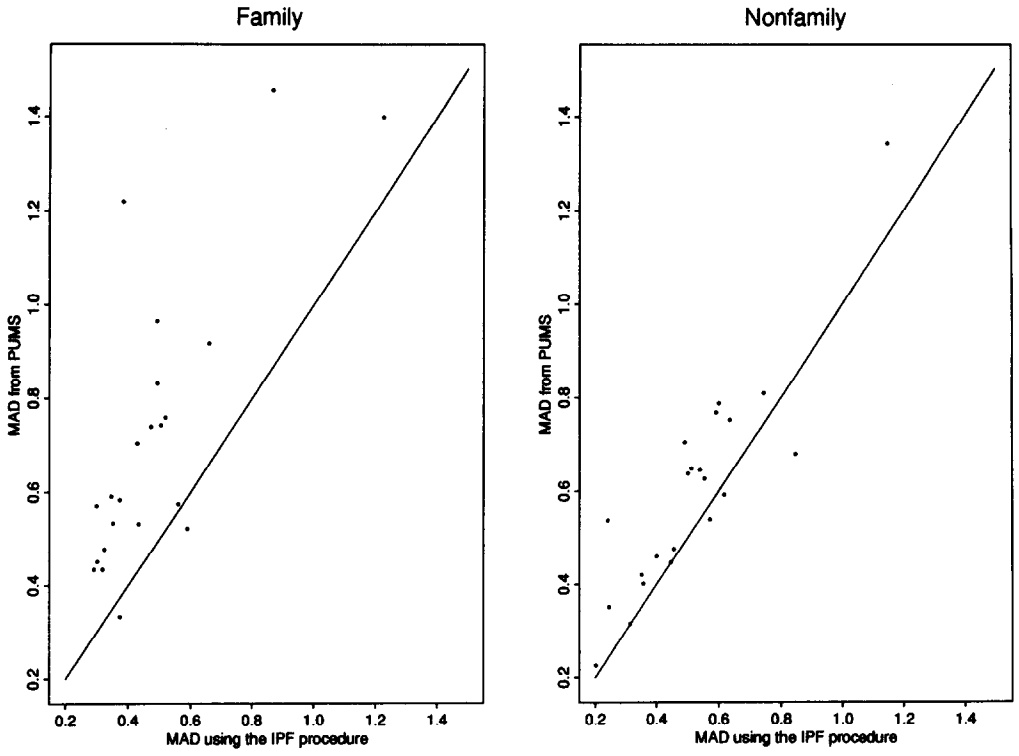


Fig. 3. The mean absolute deviation of the proportions of households with household sizes 1–5+ and 0–4+ vehicles in the synthetic population from the true proportions. Points on the plots represent these differences using only the PUMS and those differences using IPF. The differences are smaller for those populations generated with IPF.

Both imputation and tweaking were investigated and neither improved on the results obtained by using the basic algorithm. Synthetic family populations for census tract 1.07 of Bernalillo County, New Mexico were generated with and without both tweaking and imputing the persons in the household. Zero cells in the PUMS were replaced by 0.01 and imputation of the household size was accomplished using Classification And Regression Trees (CART) (Breiman *et al.*, 1984). Both of these investigations required the use of a non 0–1 loss function since in these cases there is no guarantee of PUMS samples matching exactly the nonzero cells of the multiway table generated by IPF. The values of k and α in $D(p,c)$ of section 4 were set to 0.1 and 0.05, respectively.

Tweaking badly biases the statistics for the marginal tables and is not recommended. Imputation does little more than add variability to the synthetic populations. It does not improve the results from the basic algorithm. These results are not surprising. Tweaking places an entry in every cell of the estimated multiway table. Some of these cells may be much different from the closest samples in the PUMS. Therefore, selection of even the closest PUMS samples using $D(p,c)$ will tend to bias the results. On the other hand the household size was imputed from the PUMS and the closest cells are not too distant from the nonzero cells in the multiway table. Hence, in this case no appreciable bias is added to the sample.

The use of a non 0–1 loss function was also investigated for populations generated with no imputation or tweaking. The results are similar to the imputation results where increased variability but no bias was noticed.

To maintain the proper odds ratio structure, the statistically correct procedure for IPF is to create an $(m+1)$ dimensional table, with the table number as the additional variable is the table number and to fit the $(m+1)$ tables simultaneously. Investigations show that this procedure only marginally outperforms the simpler procedure of fitting the multiway tables individually. Consequently, the simpler procedure can probably be used without much harm.

One could assume that the variables that make up the multiway table are independent. The marginal proportions for these variables could be multiplied and the synthetic households could be drawn from the PUMS using $D(p,c)$. However, this procedure has the same problem as "tweaking" in that it badly biases the marginal distributions of the individual variables, not to mention the joint distributions of variables such as household size and the number of vehicles. Assuming independence between the marginal variables is not recommended.

SUMMARY

A method has been given for the generation of synthetic populations on a census tract or block group basis. The technique uses only census data and it reproduces the existing population in a reasonable way. There are two steps in the methodology. First a multiway demographic table of proportions is estimated. Here, it is estimated using iterative proportional fitting. However, any reasonable statistical method for this estimation would be acceptable. Secondly, a synthetic population of households is drawn from the PUMS so as to match the proportions in the estimated table.

We have shown by validation that synthetic populations generated by this procedure have desirable characteristics. In the synthetic population the marginal distribution of variables used in the construction of the multiway table match the truth exactly (within rounding). The distribution of variables not used in the construction of the multiway table, such as household size, are reasonably estimated as evidenced in Fig. 1. Also, the joint distributions of two or more variables not used in the multiway table construction, such as household size and the number of vehicles, are estimated well by this procedure.

Acknowledgements—This work is part of the TRANSIMS project which is sponsored by The Department of Transportation, The Environmental Protection Agency and The Department of Energy. The authors would like to thank the referees and the Editor whose comments improved an earlier version of this paper.

REFERENCES

Axhausen K. W. and Gärling T. (1991) *Activity-based Approaches to Travel Analysis: Conceptual Frameworks, Models and Research Problems*. U.S. Department of Commerce, National Technical Information Service, TSU Ref:628.

Bhat C. R. and Koppelman F. S. (1993) A conceptual framework of individual activity program generation. *Transpn Res. A* **27**, 433–446.

Breiman I, Friedman J. H., Olshen R. A. and Stone C. J. (1984) *Classification and Regression Trees*. Wadsworth and Brooks, Monterey, CA.

Census (1992a) *Census of Population and Housing, 1990*; Summary Tape File 3 on CD-ROM Technical Documentation/prepared by the Bureau of the Census. The Bureau, Washington.

Census (1992b) *Census of Population and Housing, 1990*; Public Use Microdata Sample U.S. Technical Documentation/prepared by the Bureau of the Census. The Bureau, Washington.

Deming W. E. and Stephan F. F. (1940) On a least squares adjustment of a sampled frequency table when the expected marginal tables are known. *Annals Mathl Stats* **11**, 427–444.

Dugway G., Jung W. and McFadden D. (1976) *SYNSAM: A Methodology for Synthesizing, Household Transportation Survey Data*. Working Paper No. 7618, University of California, Berkeley.

Gärling T., Kwan M. and Colledge R. G. (1994) Computational-process modelling of household activity scheduling. *Transpn Res. B* **28**, 355–364.

Ireland C. T. and Kullback S. (1968) Contingency tables with given marginals. *Biometrika* **55**, 179–188.

Kitamura R. (1988) An evaluation of activity-based travel analysis. *Transpn.* **15**, 9–34.

Little R. J. A. and Wu M. M. (1991) Models for contingency tables with known marginals when target and sampled populations differ. *J. Stat. Assoc.* **86**, 87–95.

Papacostas C. S. and Prevedouros P. D. (1993) *Transportation Engineering and Planning*, (2nd Ed). Prentice Hall, Englewood Cliffs, NJ.

Purvis C. L. (1994) Using 1990 Census Public Use Microdata Sample to estimate demographic and automobile ownership models. *Transpn Res. Rec.* **1443**, 21–29

Recker W. W. (1994) A household activity pattern problem: general formulation and solution. *Transpn Res. B* **29** 61–77.

Recker W. W., McNally M. G. and Root G. S. (1986a) A model of complex travel behavior: Part I—theoretical development. *Transpn Res. A* **20**, 307–318.

Recker W. W., McNally M. G. and Root G. S. (1986b) A model of complex travel behavior: Part II—an operational model. *Transpn Res. A* **20**, 319–330.

Smith L., Beckman R., Anson D., Nagel K. and Williams M. (1995) *TRANSIMS: TRansportation ANalysis and SIMulation System*. Los Alamos National Laboratory Unclassified Report, LA-UR-95-1664, Los Alamos, NM 8744.