Blow Up: Expanding a Complex Random Sample Travel Survey

Peter R. Stopher and Cheryl Stecher

In April 1991 the Southern California Association of Governments contracted with the Applied Management and Planning Group to conduct an origin-destination survey of 15,700 households in five Southern California counties. The survey sample was stratified for each county by regional statistical areas (RSAs). Within each county, the sample was stratified also on housing type, vehicle ownership, and household size, yielding a 30-cell sampling matrix for each of the five counties but needing to be reported by the 49 RSAs in the study area. The overall sampling frame was thus a combination of 49 cells (RSAs) and 150 cells (stratification within county). The survey data were collected between April and June 1991. The subsequent file creation and analyses extended far enough into 1992 to enable the use of the 1990 census data to expand the sample to the population. The available census data for expansion consisted of the one-way frequencies of each of the three sociodemographic variables but did not provide any cross tabulations of these variables. The expansion and weighting procedure used to generate population estimates that match as closely as possible the characteristics measured in the 1990 census is described. It is shown that the procedure is relatively simple, even though the sampling procedure was complex. In addition, the sampling errors are reported and are compared with those that would have been obtained from a simple random sample of the entire region.

In April 1991 the Southern California Association of Governments (SCAG) contracted with the Applied Management and Planning Group to conduct an origin-destination survey of 15,700 households in five Southern California counties. The sampling methodology was based on random sampling of households, using random digit dialing (partly screened to eliminate blocks of business numbers and numbers listed in the Yellow Pages). The sampling methodology is based on both geographic and socioeconomic stratification, in the following procedure.

DESCRIPTION OF REGION AND SAMPLING REQUIREMENTS

The Los Angeles region is an urbanized area that covers three entire counties (Los Angeles, Orange, and Ventura), together with the western portion of two counties (San Bernardino and Riverside counties). This urbanized area contained about 14 million residents in 1991. The region extends from the Pacific Ocean on the west to the Mojave Desert on the east, and from San Diego County on the south to the mountain ranges on the north that run from San Bernardino at the eastern end to the coastal ranges at the western end. The area contains a number of major cities, including Los Angeles, Long Beach, Santa Ana, Anaheim, Riverside, San Bernardino, Ventura, and Oxnard. It is estimated that the region contains between 4 million and 5 million households. The majority of the population is located in Los Angeles County (about 8 million people), with the next largest group being in Orange County (about 2 million). The rest are spread almost equally between Ventura, Riverside, and San Bernardino counties.

For planning purposes, the region is divided into regional statistical areas (RSAs), of which there are 44 in urbanized areas and another 12 in rural areas not included as part of the urbanized region (the eastern portions of San Bernardino and Riverside counties and Imperial County to the south and east). The RSAs vary widely in both geographic extent and population, but they are apportioned among the counties in approximate proportion to the populations in each county. RSA boundaries are contiguous with county boundaries and also with 1980 census geography (i.e., each RSA contains whole census tracts and no partial tracts). There are 21 RSAs in Los Angeles County, 10 in Orange County, 6 in Ventura County, 4 in the urban portion of Riverside County, and 3 in the urban portion of San Bernardino County. (The survey area was extended in each of Riverside and San Bernardino counties to include some of the developing western desert areas of Palm Springs and the Coachella Valley, adding three RSAs in San Bernardino and three in Riverside. One RSA in Los Angeles County was deleted from the study because it was predominantly in the mountains and had only 700 listed telephone numbers. The total study area thus covered 49 RSAs.)

Initially, sampling requirements were computed for the region, using Smith's method (*1*) as modified by Stopher (*2*), using a geographic stratification by county and socioeconomic stratification by household size, vehicle ownership, and housing type. For this purpose, household size was divided into five categories, covering each size of household from one person through five persons and up; vehicle ownership was divided into three categories (no vehicle, one vehicle, and two or more vehicles); and housing type was divided into two categories: single-dwelling unit (SDU) and multiple-dwelling unit (MDU). This procedure used the sample standard deviations of trip rates from the 1967 survey but corrected the distribution of households from 1967 to the estimate (at that time) of 1990 population distribution among the categories.

These sampling requirements were based on the objective of achieving a sampling error less than or equal to that of the 1967 survey, with the criterion level being set as ±5 percent

P. R. Stopher, Louisiana Transportation Research Center, 4101 Gourrier Avenue, Baton Rouge, La. 70808. C. Stecher, Applied Management and Planning Group, 12300 Wilshire Boulevard, Suite 430, Los Angeles, Calif. 90025.

error at 95 percent confidence. This led to a sample size in each county that was quite modest, generally falling below that required for statistically accurate modeling of trip distribution and mode choice. The sample sizes also fell below those thought to be politically necessary for the acceptance of population and travel data in such a large urban region.

Through a combination of the political acceptability, available funding for data collection, and the distribution of funding among the counties, it was determined that a total sample of 15,800 households should be targeted for completion.

Sample Design

Initially, it was considered desirable to allocate the sample equally among all RSAs in the targeted area, thus resulting in an initial sample of 320 households from each RSA. This produced targets of 6,400 households in Los Angeles County, 3,200 in Orange County, 2,880 in Riverside County, and 1,600 in each of San Bernardino and Ventura counties. The rationale for equal sample sizes in each RSA was largely that it provides equal sampling errors in each RSA. However, the geographic sampling in this manner also provided a reasonable match between the sources of funding and the sample generated and ensured that the largest samples would be drawn from the most populous areas.

Within each county, it was then desired to ensure that the sample met the minimum requirements of sample size to achieve the maximum error of ±5 percent at 95 percent confidence for each cell of the three-dimensional socioeconomic matrix. Therefore, the expected sample was distributed among the socioeconomic categories, and the expected sample sizes were compared with those computed from the modified Smith's method. In Los Angeles and Orange counties, it appeared that all cells would exceed significantly the minimum sample size needed. However, it appeared that some cells in the other three counties would contain fewer than the desired number of samples.

The sampling methodology used was to draw a random sample from within each RSA and classify each household during the initial telephone contact into the appropriate housing type, vehicle ownership, and household size cell by county. As sampling approached 90 percent of the target, these totals were examined to determine if any were in danger of not being met. In the event that any cells were less than 90 percent complete while other cells already met or exceeded the sample requirements, remaining samples in that county were shifted from a purely random sampling into a stratified sampling, in which only households from the still-incomplete cells were added to the sample. This was done by using the classifying data in the initial telephone interview as a screening procedure and terminating households that fell into cells that had already reached their target levels. However, of the 16,000 households completing the survey, only 96 were recruited using this targeted approach.

Practical Aspects of Sampling Methodology

It is important to note that the survey mechanism used consisted of an initial telephone interview that classified a household and attempted to recruit the household for diary completion. Then diaries were mailed to all households that agreed to participate, following which the data in the diaries were retrieved through a telephone call. A number of households will refuse to participate at the outset of the first telephone contact, constituting an outright refusal. Others will terminate the interview before recruitment or will refuse to be recruited to complete the diaries. A third group will agree to be recruited but then will not complete diaries or will refuse to furnish the information from the diaries when called. A financial incentive was sent with the mailed-out diaries in order to achieve a higher response rate.

To allow for these various refusals and premature terminations, the initial telephone recruitment seeks to oversample fairly significantly in order to compensate for the losses of sample. The initial recruitment aimed at obtaining approximately 750 households in each RSA and at recruiting in each socioeconomic classification cell about 2.5 times as many households as were considered necessary to meet sampling requirements. However, when the response rates vary by cell from the recruited households (as is always the case), the final sample will be distributed differently from the recruited sample.

EXPANSION METHODOLOGY: SCAG HOUSEHOLD SURVEY

Overview

The methodology to expand the SCAG household survey data consisted of two primary steps: the first was to expand the actual survey responses to represent the total population of households, and the second was to reweight the expanded data to represent the proportion of households by size, housing type, and vehicle ownership. The first of these two steps is based on the number of responding households in each RSA compared with the total number of occupied households residing in the RSA (*3*). (There is a further step to the expansion within this process that accounts for the lack of response by individual household members within any household, where the intent is to expand the number of trips recorded to account for household members that did not complete diaries; however, this step was not performed as part of this study.)

The second step consists of calculating weights from available data to correct for biases in the final samples. Biases arise from two principal sources: (a) self-selection by households concerning response to the survey, wherein experience has shown that households with two or third persons and at least one vehicle are more likely to respond than most other household groupings; and (b) intentional stratified sampling, wherein the sample design was aimed at ensuring certain minimum numbers of households of various subcategories being included in the sample.

Step 1: Expansion

In this step, two expansion activities should be undertaken. The first, which was not done in this study, is at the level of the household and the second, which was done in this study, is at the level of the RSA. For the first step, a trip rate per person would be calculated for each household type, household size, and vehicle ownership level within each county. This per person trip rate would be calculated by dividing the total trips recorded by persons in households of each category and dividing these by the number of diaries from which the sum was obtained. Households would then be identified in which the number of completed diaries retrieved was less than the number of persons reported as living in the household (but not including any persons who completed a diary and returned it indicating that on that day they either did not leave home or were out of town). The person trip rate would be multiplied by the number of individuals that did not respond with a diary in each household category, and this number would be added into the household record.

The second step in the expansion is to compute the ratio of the occupied households in each RSA to the number of responding households in the RSA. This ratio is used as a multiplier for all households in the RSA. For example, the data show that RSA 7 in Los Angeles County had 20,192 occupied households in 1991, and the survey obtained responses from 324 households in that RSA. Therefore, the expansion factor applied to all the households in RSA 7 is 20,192/324, or 62.32. In other words, each surveyed household in RSA 7 represents 62.32 actual households. But even though it appears obvious that this should be the next step in the process, there are good reasons to make this the final step of the expansion process, when the reweighting data are only partially known.

Step 2: Reweighting

Ideally, the second step would be conducted by using three-way cross tabulations of the census data, updated to 1991, showing the actual distributions of households by size, housing type, and vehicle ownership. Although sampling was performed at the county level against these categories, it is more accurate to calculate weights at the RSA level because the bias in the raw survey data stems from both the stratified sampling and the differential response rates that can be expected to differ from RSA to RSA. If the 1990 census totals were available by RSA, then this step is also a simple one, in which the numbers of households of each class as shown by the census for each RSA would be divided by the number of that class found in the sample for each RSA. This factor would be multiplied by the sample number, thus re-creating the actual number of households in each class in each RSA.

Because the 1990 census statistics showing cross tabulations of these three variables were not available in time for use in this expansion effort, an alternative strategy was used. The only census data that were available in the appropriate time frame were the one-way totals of households by housing type, household size, and vehicle ownership (from the May release—STF3). Cross tabulations of any variables were not available.

In the recruitment process for households, five dispositions were possible for each contacted household: the household member who was contacted

1. Refused to answer any questions;
2. Refused to answer questions of household size and automobile ownership, although he or she may have answered prior questions;
3. Refused to participate in the diary survey but did answer the questions on household size and automobile ownership;
4. Answered the telephone interview questions and agreed to participate in the diary survey, from which diaries were not, however, retrieved; or
5. Answered the telephone interview questions and agreed to participate in the diary survey, and diaries were successfully retrieved.

The fifth of these groups is the one that provided the raw survey data. The first two groups provide no usable data. The third and fourth groups provide data from which distributions can be computed of households by household size and vehicle ownership (housing type was not collected from Group 3). Adding Groups 3, 4, and 5 and stratifying by RSA, household size, and vehicle ownership, distributions can be obtained that are less biased than those from the fifth group alone. Determining the proportions of each type of household in each RSA from the summed groups divided by the proportion in the fifth group alone provides a reweighting coefficient that can be applied to each household in each cell of the matrix within each RSA, when census data are lacking. In this project, however, these weights were not used because census data were available in time to be used.

Because census data currently available provide only the one-way distributions on the control variables, an iterative row-and-column balancing (iterative proportional fitting or the Furness method) is used to correct the two-dimensional matrices obtained by taking each of the variables two at a time to create the most probable underlying cross classification, thereby producing weights to redistribute households in each class. In this method, the row and column entries are balanced alternately in iterative steps until the iterations converge to a stable set of cell values that sum to the desired row and column control totals.

In this project, the first reweighting was for dwelling unit type, using the RSA totals of SDUs and MDUs, adjusted for occupied units, as the control. Adjustment factors were obtained from the final iteration and were multiplied through all cells to yield new totals of the sample data, from which automobile ownership statistics were obtained. In the second step, automobile ownership was adjusted through the same procedure by RSA, after which automobile ownership and housing type by county were readjusted to produce county totals for each category that matched the census data. New composite expansion factors were derived from this and applied to the original distribution of households, from which new cell totals were determined and new statistics produced on automobile ownership and housing type.

Again, because the cross-tabulation data were not available, the next item to become available from census releases was that of household size distributions that were provided by SCAG at the RSA level. Using the cell values produced by the preceding factoring step, totals were computed by RSA and county for households in the five household size groups. These were factored to produce the correct RSA totals, following which two successive applications were made of the

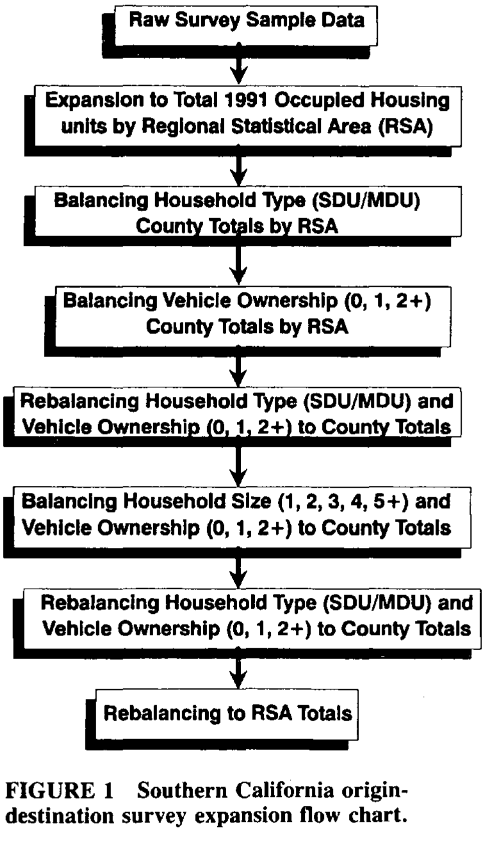
Furness method, first to produce household size by automobile ownership and then to produce household size by housing type, after incorporating the adjustments from the previous step.

In the final step, the composite factors were applied to the original cells and the RSA population totals were rebalanced to total RSA expanded population. (It would have had the same effect to have performed all of the preceding adjustment steps on unfactored data, using the proportions of households in each category instead of the absolute numbers, and then factoring the resulting RSA sample populations to the total census populations. Mathematically the results are identical, and there is some appeal in seeing figures throughout the adjustment process that are of the order of magnitude of the total population.)

The procedure described here yielded expanded data that were, for all five counties, less than 1 percent different from the countywide control totals on each of the three categorical variables and that represented an exact match to the census population totals. Figure 1 shows a summary of the steps used to expand the data.

Example

An example of the application of this methodology may be helpful. The example uses one specific RSA to track the effects of the steps in the expansion process. At the time that the expansions were performed, census data were available to provide the 1990 estimates of population, number of households, and distributions of households by vehicle ownership, housing type, and household size. No cross tabulations of these attributes were available, however. Table 1 gives the actual numbers of households recruited that completed diaries for all members of the household present on the day of the survey, by category.



Raw Survey Sample Data
Expansion to Total 1991 Occupied Housing units by Regional Statistical Area (RSA)
Balancing Household Type (SDU/MDU) County Totals by RSA
Balancing Vehicle Ownership (0, 1, 2+) County Totals by RSA
Rebalancing Household Type (SDU/MDU) and Vehicle Ownership (0, 1, 2+) to County Totals
Balancing Household Size (1, 2, 3, 4, 5+) and Vehicle Ownership (0, 1, 2+) to County Totals
Rebalancing Household Type (SDU/MDU) and Vehicle Ownership (0, 1, 2+) to County Totals
Rebalancing to RSA Totals

FIGURE 1 Southern California origin-destination survey expansion flow chart.

According to census information, updated to 1991, there were 21,672 housing units in RSA 71, of which 15,909 were SDUs and 5,763 were MDUs. The estimated number of occupied housing units in 1991 was 20,192.

The initial expansion step undertaken was to expand households to the total in each RSA, on the basis of the census figures. The expansion factor for RSA 7 was 62.321 (equal to the ratio of the actual number of occupied housing units, 20,192, divided by the number of completed households, 324). The resulting distribution of households is presented in Table 2.

It should be noted that zero entries in the original table (Table 1) can never change through the expansion process. The next step was to adjust all RSAs by county to the county-level totals of MDUs and SDUs. For Los Angeles County, the total numbers of SDUs and MDUs were 1,516,956 and 1,493,449, respectively. However, the initial expansion of the data, as performed in the step shown in Table 2, generated totals of 1,734,606 SDUs and 1,276,022 MDUs. The Furness method was applied to the RSAs to rebalance to the correct totals, and stable results were obtained after 10 iterations. Applying the resulting adjustment factors produced RSA 7 values given in Table 3.

After this step, census data were obtained on vehicle ownership levels by county. This showed that the expanded data at this point contained 7.05 percent households owning zero vehicles, compared with 11.2 percent in the census; 34.6 percent households owning one vehicle, compared with 35.8 percent in the census; and 58.4 percent households owning two or more vehicles, compared with 53.0 percent in the census. The next step in the procedure was to apply the Furness method by RSA to the vehicle ownership figures. This converged after nine iterations, following which the joint distribution of households by housing type and vehicle ownership for the county was rebalanced to provide jointly the correct totals for each category of housing type and vehicle ownership. The results of this step, which also caused a shift away from the correct totals of households in each RSA, are presented in Table 4.

The next step in the procedure was to balance the county RSAs with the census distribution of household size. Table 5 gives the comparison of the expanded data and the census data, where the difference in the number of households is a result of growth from 1990 to 1991.

Applying the census percentages in Table 5 to the expanded data, new targets were determined for the distribution of households by household size. Two successive applications of the Furness method were then used: in the first, the balancing was done to the household size distribution and the vehicle ownership distribution; in the second, balancing was done to the vehicle ownership and housing type distributions. The results of these steps are presented in Table 6.

One step remained at this point, which was to return the total number of households in the RSA to the original number

TABLE 1 Final Completed Households for RSA 7 by Category

Household Size	Vehicle Ownership						Total
	0		1		2		
	SDU	MDU	SDU	MDU	SDU	MDU	
1	0	1	17	12	6	4	40
2	0	0	6	6	92	36	140
3	0	0	4	3	64	7	78
4	0	0	0	0	42	4	46
5+	0	0	1	0	18	1	20
Total	0	1	28	21	222	52	324

TABLE 2 Initial Expansion Results for RSA 7

Household Size	Vehicle Ownership						Total
	0		1		2		
	SDU	MDU	SDU	MDU	SDU	MDU	
1	0	62	1059	748	374	249	2492
2	0	0	374	374	5734	2244	8726
3	0	0	249	187	3989	436	4861
4	0	0	0	0	2617	249	2866
5+	0	0	62	0	1122	62	1246
Total	0	62	1744	1309	13836	3240	20191

TABLE 3 Second Adjustment of Household Distribution in RSA 7 Using Housing Type

Household Size	Vehicle Ownership						Total
	0		1		2		
	SDU	MDU	SDU	MDU	SDU	MDU	
1	0	78	979	941	345	314	2657
2	0	0	345	470	5296	2822	8933
3	0	0	230	235	3684	549	4698
4	0	0	0	0	2418	314	2732
5+	0	0	58	0	1036	78	1172
Total	0	78	1612	1646	12779	4077	20192

TABLE 4 Rebalanced Distribution for RSA 7 Based on Joint County Distribution of Housing Type and Vehicle Ownership

Household Size	Vehicle Ownership						Total
	0		1		2		
	SDU	MDU	SDU	MDU	SDU	MDU	
1	0	136	1168	1028	346	288	2966
2	0	0	412	514	5307	2589	8822
3	0	0	275	257	3692	504	4728
4	0	0	0	0	2423	288	2711
5+	0	0	69	0	1038	72	1179
Total	0	136	1924	1799	12806	3741	20406

TABLE 5 Comparison of Countywide Household Size Data to Expanded and Adjusted Data at Step 3

Household Size	Expanded Data		Census Data	
	Number	Percent	Number	Percent
1	770,964	25.61	745,661	24.95
2	1,122,512	37.28	835,043	27.94
3	496,478	16.49	474,760	15.89
4	379,740	12.61	417,815	13.98
5+	240,935	8.00	515,148	17.24
Total	3,010,628	100.00	2,988,427	100.00

TABLE 6 Revised Distribution of Households for RSA 7 After Step 5

Household Size	Vehicle Ownership						Total
	0		1		2		
	SDU	MDU	SDU	MDU	SDU	MDU	
1	0	136	1085	1037	311	281	2850
2	0	0	304	412	3791	2009	6516
3	0	0	261	265	3399	504	4429
4	0	0	0	0	2576	332	2908
5+	0	0	146	0	2141	161	2448
Total	0	136	1796	1714	12218	3287	19151

in Table 2. The results of this are given in Table 7, and the final expansion factors are given in Table 8. Table 9 presents a comparison for the entire county of the numbers of households in each category of the three variables and the percentage differences, after applying the final adjusted expansion factors.

Comparing Table 7 with Table 2, it can be seen that the adjustment procedure has made significant changes to the distribution of households by the three categorization variables. As might be expected from an examination of the adjustments required, population has been shifted out of the households that own two or more vehicles (both SDU and MDU) and added into households that own zero or one. In addition, household sizes of one, four, and five-plus have each increased, and the middle two size groups have both decreased. Table 8 indicates that the expansion factors for each cell are quite markedly different, ranging from 41 to 161. This shows the clear need for a more complex expansion procedure than would have occurred using a simple expansion to the RSA population of occupied housing units.

Finally, an examination of Table 9 indicates that the iterative row and column balancing (Furness method) has produced results that are within less than 1 percent error on all county-level demographics. The largest errors occur on house-

TABLE 7 Final Expanded and Adjusted Household Distribution in RSA 7

Household Size	Vehicle Ownership						Total
	0		1		2		
	SDU	MDU	SDU	MDU	SDU	MDU	
1	0	143	1143	1093	327	296	3002
2	0	0	320	434	3997	2118	6869
3	0	0	275	279	3583	531	4668
4	0	0	0	0	2716	350	3066
5+	0	0	153	0	2257	169	2579
Total	0	143	1891	1806	12880	3464	20184

TABLE 8 Final Adjusted Expansion Factors for RSA 7

Household Size	Vehicle Ownership					
	0		1		2	
	SDU	MDU	SDU	MDU	SDU	MDU
1	0	135.7385	63.80844	86.41374	51.87328	70.25033
2	0	0	50.68585	68.64223	41.20522	55.80292
3	0	0	65.3332	88.47868	53.11284	71.92903
4	0	0	0	0	61.32925	83.05626
5+	0	0	146.316	0	118.9481	161.0876

TABLE 9 Comparison of Demographic Distributions After Expansion

Variable	Census	Expanded Sample	Percent Difference
Housing Type			
SDU	1,516,956	1,517,091	0.0089
MDU	1,493,449	1,493,537	0.0059
Vehicle Ownership			
0	337,562	337,559	-0.0009
1	1,078,985	1,078,983	-0.0002
2+	1,594,081	1;594,086	0.0003
Household Size			
1	751,201	757,387	0.8235
2	841,247	844,549	0.3925
3	478,287	476,292	-0.4171
4	420,919	418,549	-0.5631
5+	518,975	513,852	-0.9871

hold size, because that was the variable against which adjustments were made furthest back in the process. With that exception, the county totals are replicated to within less than 1/100 percent.

REFERENCES

1. Smith, M. E. Design of Small-Sample Home-Interview Travel Surveys. In *Transportation Research Record 701*, TRB, National Research Council, Washington, D.C., 1979, pp. 29–35.
2. Stopher, P. R. Small-Sample Home-Interview Travel Surveys: Application and Suggested Modifications. In *Transportation Research Record 886*, TRB, National Research Council, Washington, D.C., 1982, pp. 41–47.
3. Stopher, P. R., and A. H. Meyburg. *Survey Sampling and Multivariate Statistics for Social Scientists and Engineers*. D. C. Heath and Co., Lexington, Mass., 1979, Ch. 4.

Publication of this paper sponsored by Committee on Transportation Data and Information Systems.