Using Synthetic Population Generation to Replace Sample and Expansion Weights in Household Surveys for Small Area Estimation of Population Parameters

Konstadinos G. Goulias
Professor, GeoTrans and Department of Geography
University of California Santa Barbara
goulias@geog.ucsb.edu

Srinath K. Ravulaparthy
Ph.D. Candidate, GeoTrans and Department of Geography
University of California Santa Barbara
srinath@geog.ucsb.edu

Karthik C. Konduri
Assistant Professor, Department of Civil and Environmental Engineering
University of Connecticut, Storrs
kkonduri@engr.uconn.edu

Ram M. Pendyala
Professor, Civil, Environmental, and Sustainable Engineering Program
School of Sustainable Engineering and the Built Environment
Arizona State University, Tempe
ram.pendyala@asu.edu

Paper 14-0501 accepted for presentation at the Transportation Research Board 93rd
Annual Meeting (January 12-16, 2014, Washington D.C.) and inclusion
in the Annual Meeting Compendium of Papers

GEOTRANS REPORT 2013-08-01

Using Synthetic Population Generation to Replace Sample and Expansion Weights in Household Surveys for Small Area Estimation of Population Parameters

Abstract: In this paper we illustrate the use of synthetic population generation methods to replace sample weights and expansion weights in household travel surveys. We use a combination of exogenous (US Census) and endogenous (the survey) data as the informants and in essence transfer information from the county level sample to the tracts. The method is based on a population synthesis approach called PopGen (PopGen 1.1, 2011) and is applied to the newly collected data in the California Household Travel Survey (CHTS). An illustration of using traditional sampling and expansion weights and synthetic population generation is illustrated at the tract level. We show synthetic population methods are able to recreate the entire spatial distribution of households and persons in small areas, recreate the variation that is lost when sampling. This method is capable of reproducing the variation in the real population and enables transferability without having to develop complicated methods. Moreover, it fills spatial gaps in data collection, produces a large database that is ready to be used in activity microsimulation, provides as byproducts sample and expansion weights, and offers the possibility to perform resampling for model estimation. However, additional testing and experimentation is also required.

Paper 14-0501 accepted for presentation at the Transportation Research Board 93rd
Annual Meeting (January 12-16, 2014, Washington D.C.) and inclusion
in the Annual Meeting Compendium of Papers

GEOTRANS REPORT 2013-08-01

INTRODUCTION

Sample weights (often claimed to be the inverse probabilities of selection for each observation) aim at reshaping the sample to make it look like a simple random draw of the total population. In this way descriptive statistics from the reshaped sample produce "more" accurate population estimates than the original unweighted sample. The survey design and associated statistical analysis literature offers a variety of solutions to perform this reshaping to counter a variety of design complications but to also fix mistakes in data collection. In post survey data collection we face mainly two options for reshaping the sample distributions of values to mimic the population and they are: a) develop weights for unit of data (household, person, activity, trip); and b) develop regression models that account for external and internal stratification and self-selection biases. The first method requires knowledge about the distribution of values of population parameters and careful monitoring of survey stages to detect self-selection and its determinants. The second method requires model specifications that include variables controlling the process of sample selection. Both approaches can become extremely complex when one needs to estimate not only the population means of parameters but also their variances and the relationships among different variables (Chung and Goulias, 1995, Gelman, 2007, Dumouchel and Duncan, 1983, Gelman and Carlin, 2002, Kish, 1992, Kitamura et al., 1993, Lu and Gelman, 2003, Ma and Goulias, 1997, Pendyala et al., 1993, Pfeffermann, 1993, Solon et al., 2013, Winship and Radbill, 1994).

In small area estimation with small area in this case intended as geographic area (statisticians use the word to also mean small segments of the population, which leads to similar issues) added complications include lack of local external to the survey data, cross-tabulation of variables that have many structural zeros, very few sample survey observations, and methods that are designed only for ideal situations (Chambers and Tzavidis, 2006, Datta et al., 2011, Fay and Herriott, 1979, Jiang et al., 2011, Li and Lahiri, 2010, Molina and Rao, 2010, Pfefferman and Sverchkov, 2007, Sinha and Rao, 2009, Tzavidis et al., 2010, You et al., 2012).

The third possibility of developing sample weights is to employ methods from synthetic population generation (Beckman et al., 1996, Chung and Goulias, 1997, Greaves and Stopher, 2000, Bowman, 2004 and references therein, Frick and Axhausen, 2004, Auld et al., 2009, Guo and Bhat, 2007). This is a method that is gaining wide acceptance in travel demand forecasting, it is a related method to multiple imputation (Rubin, 2009) and data augmentation (Schafer, 2010), it is feasible at any level of US Census geographic aggregation (block, blockgroup, tracts, traffic analysis zone), and attacks the problem of survey weighting in a way that solves multiple problems including transferability (Reuscher et al., 2002, National Cancer Institute, 2010). However, testing and experimentation with this method as a survey reshaping tool is not widespread and very few authors report their experience (Greaves and Stopher, 2000) but none in attempting small area estimation. In this paper we illustrate the method using data from the California Household Travel Survey that was completed in March 2013, show preliminary results in small area estimation and discuss next steps.

The California Household Travel Survey (CHTS) sampled households from different segments of the population with different probabilities. This was done to obtain more precise information on the smaller subgroups in the population and to minimize the chance of missing a few population segments that are difficult to reach and interview. CHTS is used to calculate descriptive statistics that aim at accurately measuring the true values of parameters of interest in the population at a variety of geographic scales. The second type of analysis is the creation of behavioral models that contain relationships among variables of interest (e.g., number of cars in a household as a function of household income). The statistical literature claims that sometimes weights are needed, other times they are not influential, and some other times weights could even introduce bias in estimates (Pfeffermann, 1993, Winship and Radbill, 1994). Unfortunately, when this happens we are forced to develop complicated weights and perform weighted and unweighted estimation to study the sensitivity of our models. We may also be forced to compare different model specifications. If synthetic population is used for small area estimation, it makes more theoretical and practical sense to use this population for model estimation too. This paper clarifies some of these issues and provides options for weight creation. It also offers an illustration of the issues in small area estimation with examples of specific tracts in the Central Coast of California.

In the next section, we describe the data collection and sample selection steps. This is followed with a discussion of weight creation options and an illustration of small area estimation issues as well as the solution. The paper concludes with a summary and next steps for further analysis.

The 2010 CALIFORNIA HOUSEHOLD TRAVEL SURVEY

The 2010 California Household Travel Survey is designed with the new California policy framework in mind, taking into account the possible use of new technologies, as its Steering Committee clearly defines in the following paragraph.

"The purpose of the CHTS is to update the statewide database of household socioeconomic and travel behavior used to estimate, model and forecast travel throughout the State. Traditionally, the CHTS has provided multi-modal survey information to monitor, evaluate and make informed decisions regarding the State transportation system. The 2010 CHTS will be conducted to provide regional trip activities and inter-regional long-distance trips that will be used for the statewide model and regional travel models. This data will address both weekday and weekend travel. The CHTS will be used for the Statewide Travel Demand Model Framework (STDMF) to develop the information for the 2020 and 2035 GHG emission rate analyses, calibrate on-road fuel economy and fuel use, and enable the State to comply with Senate Bill 391 (SB 391) implementation. The CHTS data will also be used to develop and calibrate regional travel demand models to forecast the 2020 and 2035 Greenhouse Gas (GHG) emission rates and enable Senate Bill 375 (SB 375) implementation and other emerging modeling needs."

One objective for the data collected in this household travel survey is to develop a variety of newly formulated travel demand forecasting systems throughout the state and integrate land use policies with transportation policies. Very important for regional agencies is the provision of suitable data that inform a variety of new model developments including the activity-based models (ABM) and their integration with land use models at the State level and for each of the four major Metropolitan Planning Organizations (MPO). It is also the source of data for the many refinements of older four-step models and activity-based models in smaller MPOs and serves as the main source of data for behavioral model building, estimation of modules in other sustainability assessment tools, and the creation of simplified land use transportation models. Moreover, added details about car ownership and car type of households will be needed to develop a new set of models to more accurately estimate emissions of pollutants at unprecedented levels of temporal and spatial resolutions. In fact, CHTS meets the data needs criteria for a main core survey with satellite in-depth survey components we defined at an international travel survey methods conference recently (Goulias et al., 2013).

The CHTS databases include data collected by NUSTATS for the entire State of California and an added sample collected by Abt-SRBI for Southern California Association of Governments - SCAG. The databases include information about the household composition and facilities available, person characteristics of household members, and a single day place-based activity and travel diary. There are two components with the first component called the ***recruitment*** component and the second (in essence the diary) the ***retrieval***. In addition, a set of satellite type of surveys (a subset of households invited to participate in additional survey components) were also designed and administered to gain insights about the use of the transportation system (e.g., wearable GPS, GPS and OBD for cars) and to potentially complement and rectify travel information. Moreover, a long distance (for trips longer than 50 miles) extending for up to 8 weeks before the diary day was also designed and administered. For the SCAG region, Abt-SRBI designed and administered an ***add-on augment*** satellite survey containing questions that are specifically needed for model building at SCAG. The CHTS (NUSTATS and Abt-SRBI) sample selection is a combination of exogenously stratified random and convenience (see NUSTATS Final Report, 2013). This creates the need to identify ways that sample data can be "adjusted" to represent the resident population.

We define as "Universe of Addresses" all the addresses that are found within a tract. Similarly we define as "Resident Households" all the households that live within a tract. A sample of addresses is a portion of the universe of addresses and a sample of households is a portion of the resident households. The number of addresses in a tract should be higher than the number of households. For example, in Los Angeles County the following three tracts were delivered by a vendor of addresses from which samples were selected. The three tracts show that tract number 06037481101 has 1445 addresses, tract number 06037481102 has 1474 addresses and tract number 06037481103 has 1536 addresses. If in CHTS the vendor randomly selected from each tract 10 addresses, for tract number 06037481101 the probability of selection would be 10/1445 and the weight of each address 1445/10 = 144.5. In this way every selected address represents 144.5 addresses in that tract. If the same vendor selected another 5 addresses, the probability of selection would be (10+5)/1445 and the weight would be 1445/15 and every selected address representing 96.333 addresses. One could think to apply this weight to the households in CHTS in each tract but this does not apply to all selected households when other sample frames are added. For example, section 3.2.1 (NUSTATS, 2013) states:

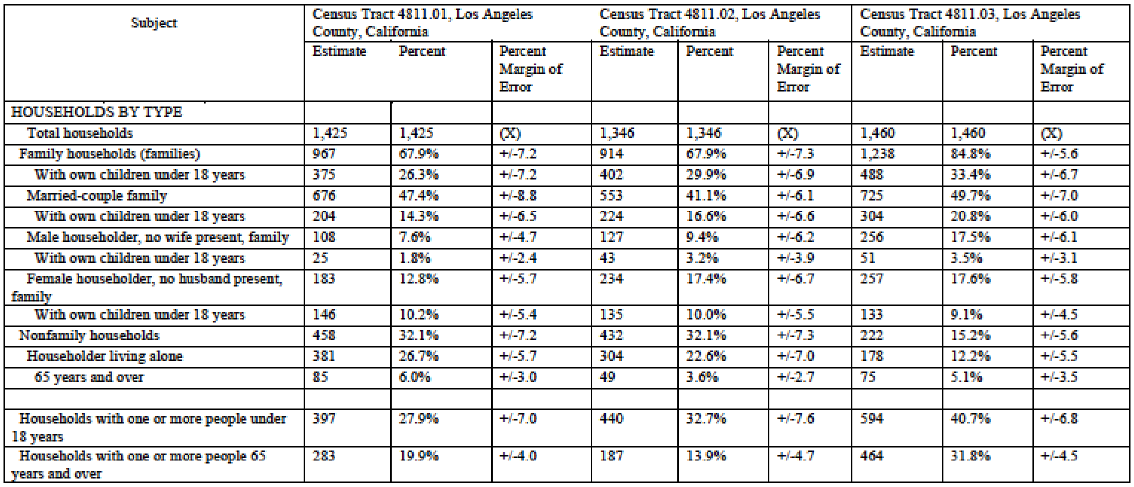
> "An Address-based sampling frame approach was used. An Address-based sample is a random sample of all residential addresses that receive U.S. Mail delivery. Its main advantage is its reach into population groups that typically participate at lower-than-average levels, largely due to coverage bias (such as households with no phones or cell-phone only households). For efficiency of data collection, NuStats matched addresses to telephone numbers that had a listed name of the household appended to the sampled mailing addresses. This sampling frame ensured coverage of all types of households irrespective of their telephone ownership status, including households with no telephones (estimated at less than 3% of households in the U.S.). In order to better target the hard-to-reach groups, the address-based sample were supplemented with samples drawn from the listed residential frame that included listed telephone numbers from working blocks of numbers in the United States for which the name and address associated with the telephone number were known. The “targeted” Listed Residential sample, as available from the sampling vendor, included low-income listed sample, large-household listed sample, young population sample, and Spanish-surname sample (to name a few). As expected, this sample was used to further strengthen the coverage of hard-to-reach households. The advantage of drawing sample from this frame is its efficiency in conducting the survey effort—being able to directly reach the hard-to-reach households and secure their participation in the survey in a direct and active approach. Both address and listed residential samples were procured from the sample provider – Marketing Systems Group (MSG) based in Fort Washington, PA. The survey population was representative of all households residing in the 58 counties in California. According to 2010 Census data, the survey universe comprised 12,577,498 households. Table 3.2.3.1 provides the distribution of households by counties and by MPO/RTPA. As shown in the table, 83% resided in four MPO regions (spread over 22 counties) – 46% in SCAG, 21% in MTC, 9% in SANDAG, and 7% in SACOG. The remaining 17% households reside in 36 counties in California."

Moreover, additional rules were used to identify tracts with households of interest (e.g., most likely hard to reach). Applying a naive weighting procedure would lead to biases because the final list of households with complete data are the result of many intermediate events from their address selection to final data delivery. A few of these events include inability to contact, refusal to participate and so forth. The outcome of these actions is a function of the characteristics of the household and the person responding to the survey. It also depends on the ability of the data collection company to convince respondents to participate and the different skills of the company's interviewers. It is usually not feasible to know the characteristics of these persons for such a large survey and this increases the uncertainty about self-selection of the final set of respondents in the recruitment stage. From many other studies we know that self-selection is usually systematic, which means the survey participants are systematically different (in their observed and unobserved characteristics) than the non-participants and obviously different from the population of interest. For example, we knew from the pilot survey in Fall 2011 that CHTS is attracting a sample that does not represent the age and gender distribution of California.

A separate issue is our uncertainty about the population characteristics at the tract level. We should keep in mind that ACS (Table 2, shows data from the three tracts mentioned above with identification numbers 4811.01, .02. and .03) is a survey and because of this it also carries an error with its estimates. As can be seen from Table 1 this error is substantial (see the third column of each tract of Table 1). However, the relative "closeness" between the number of addresses the vendor reports and the number of households is noteworthy. For example, in tract 4811.01 we see in ACS 1425 households and in the sampling frame used to select sample units we find 1450 addresses. The other two tracts show larger differences. Presumably this is due to housing units with unique address that are not occupied by households and possible errors in the sample frame and ACS data.

Independently of data errors about the statistical universe these differences show that ***addresses*** and ***households*** are not only conceptually different but also numerically different in each tract in a differential way (i.e., for some tracts we have larger differences and for others we have smaller differences). This is the main reason that we need to include at least two stages in weight creation. The **first stage** estimates the probability of selection for each address (the inverse of which is the sample weight) and the **second stage** aims to adjust the household and person probability of selection based on the information available externally such as ACS (or other control totals developed by a local agency demographic prediction procedure).

Table 1. Three Tracts in the American Community Survey (2007-2011).



Subject	Census Tract 4811.01, Los Angeles County, California			Census Tract 4811.02, Los Angeles County, California			Census Tract 4811.03, Los Angeles County, California		
	Estimate	Percent	Percent Margin of Error	Estimate	Percent	Percent Margin of Error	Estimate	Percent	Percent Margin of Error
HOUSEHOLDS BY TYPE									
Total households	1,425	1,425	(X)	1,346	1,346	(X)	1,460	1,460	(X)
Family households (families)	967	67.9%	+/-7.2	914	67.9%	+/-7.3	1,238	84.8%	+/-5.6
With own children under 18 years	375	26.3%	+/-7.2	402	29.9%	+/-6.9	488	33.4%	+/-6.7
Married-couple family	676	47.4%	+/-8.8	553	41.1%	+/-6.1	725	49.7%	+/-7.0
With own children under 18 years	204	14.3%	+/-6.5	224	16.6%	+/-6.6	304	20.8%	+/-6.0
Male householder, no wife present, family	108	7.6%	+/-4.7	127	9.4%	+/-6.2	256	17.5%	+/-6.1
With own children under 18 years	25	1.8%	+/-2.4	43	3.2%	+/-3.9	51	3.5%	+/-3.1
Female householder, no husband present, family	183	12.8%	+/-5.7	234	17.4%	+/-6.7	257	17.6%	+/-5.8
With own children under 18 years	146	10.2%	+/-5.4	135	10.0%	+/-5.5	133	9.1%	+/-4.5
Nonfamily households	458	32.1%	+/-7.2	432	32.1%	+/-7.3	222	15.2%	+/-5.6
Householder living alone	381	26.7%	+/-5.7	304	22.6%	+/-7.0	178	12.2%	+/-5.5
65 years and over	85	6.0%	+/-3.0	49	3.6%	+/-2.7	75	5.1%	+/-3.5
Households with one or more people under 18 years	397	27.9%	+/-7.0	440	32.7%	+/-7.6	594	40.7%	+/-6.8
Households with one or more people 65 years and over	283	19.9%	+/-4.0	187	13.9%	+/-4.7	464	31.8%	+/-4.5

Combined weights of this type will give the proper "importance" for each household. In this way estimation of descriptive statistics represents the resident population and behavioral models are based on data without major biases. In addition, separate weights can be developed to function as expansion weights reproducing the total numbers we obtain from the US Census. For practical purposes we will name ***sampling weights*** the multipliers of each unit/record in CHTS that give differential importance to each record to account for unequal probability of selection and ***expansion weights*** the weights that reproduce the entire population. Multiplying each record by its sampling weight(s) maintains the sample size and it is the right weight to use for model estimation, otherwise, all statistics are inflated in an artificial way.

Stages of Sample Selection

In sample selection we should distinguish between the selection done by the analysts and the selection done by each household to participate in the survey (self-selection). The stages below describe each stage and the sample selection steps.

First Stage: Selection by address and listed phone numbers

Although the original plan for CHTS was to recruit households from an exclusively address based sampling frame (called ABS in the NUSTATS report), in early 2012 a second sampling frame was added that is based on listed telephone numbers (to the best of our understanding). Moreover, NUSTATS performed an active sampling redesign as the survey administration was progressing. This was considered necessary because response rates were extremely low. However, this creates a major problem in estimating probability of selection (and its inverse that is usually the sampling weight) because the number of observations does not have a well defined universe from which a sample unit is selected. In fact, there are a few complications distancing the CHTS sample from a random sample. During the step of selecting addresses, analysts select a household by a combination of criteria that includes the address in a tract, the tract within a geographic stratum, and four main household characteristics treated as targets. These are Hispanic last name, low income (<$25,000), large household (household size >3), young household (Age<25), and for some transit classified tracts added sampling to increase transit use in the sample as well as zero vehicle households. In addition, other segments were added from special recruitment (see Table 6.4.1, NUSTATS, 2013). This complicates the process and the uncertainty about the value of the denominator of the calculation of probability of selection. Moreover, after households were selected for the first contact, they refused or were not available for a variety of reasons to participate. Non-responding households are also systematically different than responding households introducing another incidence of self-selection bias in CHTS.

Second Stage: Retrieval Management

Indicative of the process followed is the following paragraph from the NUSTATS final report (section 4.2.13):

"Sample management for retrieval was an on-going and hands on task that often times required supervisory and management staff to discuss sample segments or even specific households on the best approach to finalize the household. Some of the considerations taken into account included whether the household had been called during the day of the week and time of the day when the recruitment interview took place, whether calls had been spread out across times of the day and days of the week, whether any or too many messages had been left, or whether the household needed to be finalized as non-completed and needed to be replaced. The Strata and Quotas definition module in VOXCO allowed NuStats to manage subsets of the sample and to open or close access to any stratum or subset as needed. It also allowed NuStats to apply quotas or ceilings to control the maximum number of completed interviews by stratum, and the rate at which they were attained. This module was used for tracking goals and in sample management by assisting in the release or withholding of specific sample segments. Many of the sample management activities already described were made possible by a specific strata definition that existed in the Quota Management module. The starting point of making this sample control tool work was to specify a set of criteria or strata, upon which sample controls or quotas were to be applied. For the CHTS, quotas also were used to monitor household distribution across travel days to obtain a proportional distribution of days of the week and across weeks and months during the full year of data collection. There were situations in which there was a need to regulate or balance the rate at which a group of strata were filled during the course of the project. To achieve this, a probability or weight was assigned to a lagging stratum, so that the system would increase the rate by which sample from that strata was released. This process was critical for achieving goals on time, for example when the deadline for closing out a wearable MTC scheduling date was approaching, adding a weight to the wearable MTC sample ensured more of this sample was called to increase the chances to meet the goal on time."

The paragraph above shows that pragmatic considerations motivate the combination of probability sample selection with non-probability (quota) sample selection. This introduces biases by aiming at meeting quotas and violating the probability-based selection of sample units. Solutions for this type of process exist and a simple "raking" weight creation may mask biases but unfortunately in survey administration may also be the only option. However, there are many post-stratification and raking weight creation options. A sample of post-stratification and raking methods and issues can be found in Deming and Stephan, 1940, Deville et al., 1993, Kalton et al., 1998, Gelman and Carlin, 2002, and Gelman, 2007, among many others. In addition, households that agree to participate in the diary portion (called complete households in CHTS) are systematically different than the households selected in the first stage. The options presented in the next section aim at rectifying these biases. Under an ideal weight creation scenario we should develop weights for every decision point of sample selection. This requires a very carefully and meticulously documented sample monitoring process that in parallel also monitors the corresponding population characteristics. For example, for each tract we need to have reports of the number of addresses and the number of listed numbers by each group from which a sample unit (address and household) was selected. In addition, the population characteristics when they are known are also not known with certainty (e.g., ACS is a survey with its own set of data collection errors and cross-classification of household income by ethnicity and household size of the resident population is not available not even at the tract level).

There are, however, alternate options from a more pragmatic viewpoint that on the one hand rectify some of the survey sample biases and on the other hand offer a way to combine data from the two data collection companies as well as the two main sources of information: the CHTS sample and ACS. These options have similarities with the method used in CHTS (i.e., two stage address/listed and raking) but with the added flexibility to employ many additional variables as external controls and to use different geographic subdivisions to match the Traffic Analysis Zones of MPOs while at the same time using a combined database of the NUSTATS and Abt-SRBI core data. In summary, there are three main reasons to create sample weights (and expansion weights) and reshape the sample to resemble the residents population and they are:

- Weights to represent addresses in each tract
- Weights to adjust household and person characteristics of each tract
- Weights to counter selectivity bias between recruitment and retrieval

The options presented below aim at addressing each of these three reasons. All options below require to first merge the household data from the two data collection consultants into one database that is harmonized. The same should be done for the person data as well as all other databases. We also distinguish below between complete and incomplete households. The incomplete households herein are the households for which we do not have all the information about activity and travel.

Before listing the options a few words about the PopGen population synthesis (PopGen 1.1.) are in order. To create a synthetic population PopGen uses two main sources of information: a) a sample that functions as "seed" to provide records that can be repeatedly used as the "donor" records to replicate and produce the final population; and b) target distributions of variables that function as controls (called constraints or marginals). The algorithm first creates the equivalent of a contingency table of key variables at the household level (e.g., household size, household income) and the person level (e.g., age categories, gender). These are provided by the seed survey sample. Then, "weights" are developed that reproduce correctly the target distributions of the control variables. After this, a routine randomly selects donor households that satisfy these weights and allocates them to a geographic unit keeping track of how many times each donor is selected. The software preserves the original donor identification record number allowing to carry over all its characteristics (including any other variables).

Option A. Merge all the complete data and use the final complete households and persons from the consultants. Develop weights at the TAZ/Tract level using MPO provided sociodemographic estimates (e.g., in this case SCAG has a complex process of computing these together with local jurisdictions) employing PopGen. These will be expansion weights because PopGen creates a synthetic population (see Ye et al., 2009). Use as descriptive statistics the characteristics of the synthetic population. We can distinguish between two slightly different variants: Option A.1- Use as seed cross-tabulations in SCAG-CHTS and Option A.2 - Use as seed the address weighted cross-tabulations in SCAG-CHTS.

Option B. First merge all the complete data and use the final complete households and persons from the consultants. Then, develop a first set of weights for US Census tract residence using the data provided by the original vendor of addresses. Finally, Use PopGEN as a raking algorithm and as seed the weighted cross-tabulations from the second step above.

Option C. First merge all the data including the partially complete to form one household level and one person level database. Then, develop a first set of weights for tract residence using the data provided by Abt-SRBI. After this step create a probability model with dependent variable taking the value of 1 if the household responded to the diary portion and 0 otherwise. Use as explanatory variables in this nonlinear regression variables that explain willingness to participate and these can include location of the household. Take the inverse of the probability and use it as household sample weight. Finally, use PopGen as the raking procedure with seed the weighted households with weights from the second and third steps.

A comparison among Options A, B, and C will show if we need the more labor intensive B and C options. In the next section an example of Option A.1 is offered to illustrate PopGen's ability to reproduce external data.

Illustration of Option A.1

As a proof of concept in this section we report findings from a small pilot project that we recently completed to illustrate the ability of PopGen to function as sample rectifier in small geographic areas. We selected six tracts in Santa Barbara county and San Louis Obispo county. This was done for two reasons: a) we are familiar with the area; and b) CHTS collected a very small sample for these two counties due to administrative/management reasons. We will use these six tracts as case studies for which we will synthetically generate the households and persons they include as residents and their characteristics.

In this pilot experiment we use as "seed" database the records of 1,282 (in the database) and 1,023 (in the final report of NUSTATS) households that participate in CHTS from the Counties of San Louis Obispo and Santa Barbara. For control distributions we use the ACS 2006-2010 reported data for each tract in this same region. The four control variables are: Household size (with values 1, 2, 3, 4 or more), household income (with categories Less than $9,999, $10,000 - $24,999, $25,000 – $34,999, $35,000-$49,999, $50,000 - $74,999, $75,000-$99,999, $100,000-$149,999, $150,000 - $199,999, $200,000 or above), age (25 or below, >25 and <40, >=40 and <50, >=50 and <65, and >=65 and above), and gender (male and female). It is the combination of the categories of these variables that PopGen is attempting to replicate (finding how many combinations correspond to the population that generated the control totals of the data used as targets). This is a very old problem in statistics that can be solved with the Iterative Proportional Fitting (IPF) algorithm and in modeling software the Expectation - Maximization (EM) algorithm (Fienberg,1970, Meng and Rubin, 1993). Similar algorithms are used in the derivation of weights in the procedure known as "raking." PopGen has a very robust algorithm with different options (PopGen 1.1, 2011, Ye et al., 2009, Bar-Gera et al., 2009). NUSTATS in its final deliverable also used a raking algorithm to adjust weights and control for distributions at larger geographical areas than tracts.

Table 2 lists the six tracts used in this experiment. We list only person characteristics because PopGen reproduces the exact number of households by design. The ACS 2006-2010 is the estimate of number of persons in each tract and provides the control totals for this exercise. PopGen synthesis is the number of persons recreated by PopGen. Sample CHTS is the number of observations CHTS contains for each tract. Expanded CHTS is the multiplication of the Sample CHTS (number of persons of the CHTS sample that reside in each tract) by the person expansion weight provided by NUSTATS.

Table 2 Comparison of Persons Replication per Tract

Tract Number	(A)Persons in ACS 2006-2010	(B)Persons in PopGen 1.1	Difference (A) - (B)	CHTS Sample	(C) Expanded CHTS Sample	Difference (A)-(C)
06083 001000 Santa Barbara	5409	5362	47	14	10535	-5126
06083 000900 Santa Barbara	3085	2883	202	11	3346	-261
06083 002906 Goleta	4013	3822	191	25	4563	-550
06083 002402 Santa Maria	11793	11000	793	14	9577	2216
06079 010703 Morro Bay	3739	3780	-41	31	2241	1498
06079 011102 San Louis Obispo	5306	5395	-89	28	2272	3034

As one would expect the expansion weights developed by NUSTATS are reproducing the number of persons in these tracts. PopGen on the other hand does not agree always with the total number of persons estimated by ACS but it is much closer than the CHTS expanded sample because it attempts to recreate the persons corresponding to the ACS number of households (the number of households are reproduced exactly).

The second part of the illustration (Table 3) examines the age distribution replication by CHTS (unweighted and weighted sample) and the generated synthetic population by PopGen. We report only the person characteristics of Tract number 06079011102.

Table 3 Age Distribution within Tract 06079011102

	ACS 2006-2010		PopGen		CHTS Sample		CHTS Expanded	
AGE	Frequency	Percent	Frequency	Percent	Frequency	Percent	Frequency	Percent
<=25	2383	44.9	2426	45.0	4	14.3	369	16.2
>25 & <40	1273	23.9	1299	24.1	3	10.7	372	16.4
>=40 & <50	490	9.23	531	9.8	4	14.3	266	11.7
>=50 & <65	799	15.05	787	14.6	14	50.0	897	39.5
>=65	361	6.80	352	6.5	3	10.7	368	16.2
Total	5306	100.0	5395	100.0	28	100.0	2272	100.0

The age distribution produced by PopGen closely resembles the ACS age distribution because it uses it as target. The age distribution of the weight expanded CHTS is not reproducing ACS at this level of geography because it was not intended for this purpose and used a different geography to derive sample and expansion weights. This shows that Metropolitan Planning Organizations in California should not use these weights if they need to perform data analysis and model estimation at the tract level (this is the level that Traffic Analysis Zones are often defined).

It should also be noted that PopGen, before creating a synthetic population, creates weights for each combination of control variables. This translates into weights for each unit in the sample used as seed. When the sample used as seed is the CHTS survey, PopGen in essence produces the equivalent of raking weights for CHTS balancing the information at the household level with the information at the person level. MPOs can use these weights to perform weighted sample descriptive analyses at higher levels of geography (e.g., the county level).

An additional benefit in using synthetic population generation with seed information from a survey is also our ability to "carry over" other variables that were not used in developing weights and marginal control total targets. For example, we use the information in the synthetic population to estimate the number of trips per person per day at each tract analyzed here. Table 4 shows this estimate for each tract of Table 2.

Table 4 Estimates of Daily Trips per Person

Tract Number	Daily Trips/ person	Synthetic Persons	Std. Deviati on	In CHTS	Respon- dents	Weighted CHTS
900	3.39	2882	2.905	3.64	11	3.57
1000	3.50	5360	2.995	3.93	14	3.27
2402	3.48	10983	3.054	2.21	14	2.12
2906	3.57	3801	3.029	2.75	24	2.55
10703	3.40	3779	2.908	3.61	31	3.18
11102	3.99	5393	3.717	4.29	28	4.07
Total		32198			122	

The average number of daily unweighted person trips in CHTS are: in Santa Barbara 3.43 (sd = 2.96) by 1044 persons and in San Louis Obispo are 3.20 (sd = 2.83) trips by 1898 persons. When we apply the NUSTATS sampling weights for Santa Barbara we get 3.31 (sd = 2.92) trips by 1201 persons and the average number of trips in San Louis Obispo is 3.18 (sd=2.78) trips by 797 persons. The synthetic population performed an averaging of the values derived from the small samples borrowing information from the entire county. On the other hand weighted CHTS amplifies the already skewed sample of each tract (see Tracts 2402 and 2906).

Regression Models and Sample Weights

The majority of the models used in activity-based and trip-based approaches to travel demand forecasting are based on regression methods (this includes the discrete choice models). In model estimation as mentioned in the introduction we are aware of the need and risks of using weights. In model estimation based on data that are not collected in a completely random way we have two schools of thought: a) regression models do not need weighted sample if one includes all the variables that were used to stratify or otherwise bias the sample (Winship and Radbill, 1994); and b) regression models should use weights to account for non equiprobablity sample data collection (DuMouchel and Duncan, 1983, Little, 1991). A more cautious strategy would be to use model specifications that include as explanatory variables place of residence, household income, age of respondents and age of householder, and ethnicity to account for sample selection used in CHTS (these are also the sample weight creation by raking dimensions). Then, perform estimation with and without sampling weights and examine in detail regression statistics. If an MPO adheres to the second school of thought and uses sampling weights in regression model estimation, the weights produced by PopGen and attached to each household and person used in the seed could be employed. In addition, when commercially available software is used attention should be paid on how the variance of all estimates is computed. In this way, models enriched in their estimation with an array of variables that account for sample selection and sample weights are more likely to represent the population. However, there is a third method that we believe is superior to this.

After a region-wide synthetic population is created using as seed the CHTS data, multiple random samples can be extracted. Then, for model estimation we can use these random samples from this synthetic population. A repeated resampling from the synthetic population and estimation of models will provide an empirical test of the variability in regression coefficients. It is expected that different samples will lead to similar coefficients. However, as in many applications of this type it is always a good idea to include in the explanatory variables as many indicators as possible that can capture the selectivity bias in the survey design. For example, household characteristics (size and age composition), household income, and ethnicity will control for the initial selection of households. A variety of location indicators will also account for the county of residence stratification. For example, in SimAGENT using as explanatory variables the accessibility indicators that were developed at fine spatial resolution (Chen et al., 2011, Lei et al., 2012) or other spatial classification methods (Henson and Goulias, 2011, Mohammadian and Zhang, 2007) will eliminate the need to account for county of residence.

SUMMARY

Considerable complications emerging from extremely low response rates and the need to avoid major disasters in data collection create the need to develop sampling weights that are based on only partial information with unknown precision and unusable small area estimation of population parameters. In this paper we illustrate the use of synthetic population generation methods to fill the gap. We use a combination of exogenous (US Census) and endogenous (CHTS database) as the informants and in essence transfer information from the county level sample to the tracts in a similar line of estimation as Simpson and Tranmer (2005) demonstrated using IPF. Three possible options are also presented as variants of the use of synthetic population as the core methodology. There are many advantages doing this because a synthetic population:

a) recreates the entire spatial distribution of households and persons in small areas;
b) recreates the variation that is lost when sampling and possibly mimics the variation in the real population;
c) enables transferability without having to develop complicated methods and fills spatial gaps in data collection;
d) produces a large database that is ready to be used in activity microsimulation;
e) provides as byproducts sample and expansion weights; and
f) offers the possibility to perform resampling for model estimation.

Additional testing and experimentation is required to provide guidelines on the use of the methods here and to make comparisons with alternate procedures. This is left as a future task.

Acknowledgments

SCAG and BMC provided support for the latest development of PopGen. CALTRANS, SCAG, MTC, CEC and other agencies provided funding for data collection in CHTS. Funding for this paper was provided by the University of California Transportation Center and the University of California Multicampus Research Program Initiative on Sustainable Transportation. SCAG also funded the first author to serve as management consultant in the CHTS project. This paper does not constitute policy of any government agency.

REFERENCES

Auld, J., Mohammadian, A., Wies, K., 2009. Population synthesis with sub-region-level control variable aggregation. ASCE Journal of Transportation Engineering 135 (9), 632–639.

Bar-Gera H., KC Konduri, B Sana, X Ye, RM Pendyala (2009) Estimating survey weights with multiple constraints using entropy optimization methods. Paper presented at Transportation Research Board 88th Annual Meeting and included in the CD ROM proceedings.

Beckman, R.J., Baggerly, K.A., McKay, M.D., 1996. Creating synthetic baseline populations. Transportation Research, Part A 30, 415–429.

Bowman J (2004) A Comparison of Population Synthesizers Used in Microsimulation Models of Activity and Travel Demand. Working Paper. <http:// jbowman.net/> (Accessed 07/03/2013).

Chambers, R. and Tzavidis, N. (2006). M-quantile models for small area estimation. Biometrika, 73, pp. 597-604.

Chen Y., S. Ravulaparthy, K. Deutsch, P. Dalal, S.Y. Yoon, T. Lei, K. G. Goulias, R. M. Pendyala, C. R. Bhat, and H. Hu. (2011) "Development of indicators of opportunity-based accessibility." *Transportation Research Record: Journal of the Transportation Research Board* 2255, no. 1,. pp. 58-68.

Chung, J. and K. G. Goulias (1995) Sample selection bias with multiple selection rules: An application with residential relocation, attrition, and activity participation in the Puget Sound transportation panel. *Transportation Research Record*, 1493, pp. 128-135.

Chung, J. and K.G. Goulias (1997) Travel demand forecasting using microsimulation: Initial results from a case study in Pennsylvania. *Transportation Research Record* 1607,pp. 24-30.

Datta, G. S., Hall, P. And Mandal, A. (2011). Model selection for the presence of small-area effects and applications to area-level data. Journal of the American Statistical Association, 106, pp. 362-374.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* 11, pp. 427--444.

Deville, J. C., Särndal, C. E., & Sautory, O. (1993). Generalized raking procedures in survey sampling. *Journal of the American Statistical Association*, *88*(423), 1013-1020.

DuMouchel,W.H., Duncan, G. J. (1983). Using Sample Survey Weights in Multiple Regression Analyses of Stratified Samples. Journal of the American Statistical Association 78, pp. 535–543.

Fay, R.E. and Herriot, R. A. (1979). Estimates of income for small places: an application of James-Stein procedures to census data. Journal of the American Statistical Association, 74, pp. 269-277.

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 907-917.

Frick, M., Axhausen, K.W., 2004. Generating Synthetic Populations using IPF and Monte Carlo Techniques: Some New Results. In: Paper Presented at the 4th Swiss Transport Research Conference.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science*, *22*(2), 153-164.

Gelman, A. and Carlin, J. B. (2002). Poststratification and weighting adjustments. In Survey Nonresponse (R. M. Groves, D. A. Dillman, J. L. Eltinge and R. J. A. Little, eds.) 289–302. Wiley, New York.

Goulias, K. G., Pendyala, R. M., & Bhat, C. R. (2013). Keynote—Total Design Data Needs for the New Generation Large-Scale Activity Microsimulation Models. In *Transport Survey Methods: Best Practice for Decision Making*. (Zmud, J., & Lee-Gosselin, M. Eds), Emerald Group Publishing.

Greaves, S.P., Stopher, P.R., 2000. Creating a synthetic household travel/activity survey – rationale and feasibility analysis. Transportation Research Record, 1706, pp. 82–91.

Guo, J.Y., Bhat, C.R., 2007. Population Synthesis for Microsimulating Travel Behavior. Transportation Research Record 2014, pp. 92–101.

Henson K. and K. Goulias (2011) Travel Determinants and Multiscale Transferability of National Activity Patterns to Local Populations. *Transportation Research Record: Journal of the Transportation Research Board, No. 2231*, Transportation Research Board of the National Academies, Washington D.C., 2011, pp. 35-43.

Jiang, J., Nguyen, T. and Rao, J. S. (2011). Best predictive small area estimation. Journal of the American Statistical Association, 106, pp. 732-745.

Kalton, Graham, Ismael FloresCervantes, Hui Zheng, Roderick JA Little, Changbao Wu, Ying Luan, Hao Lu et al. "Weighting methods." *New Methods for Survey Research* (1998) - available at http://www.asc.org.uk/publications/proceedings/ASC1998Proceedings.pdf#page=89 (accessed November 2013).

Kish, L. (1992). Weighting for unequal Pi . Journal Official Statistics 8. pp. 183–200.

Kitamura, R., R. M. Pendyala, and K. G. Goulias (1993). Weighting Methods for Choice-Based Panel Correlation and Initial Choice. In Transportation and Traffic Theory, Elsevier Science Publishers, 1993, pp. 275–294.

Lei, T. L., Chen, Y., & Goulias, K. G. (2012). Opportunity-Based Dynamic Transit Accessibility in Southern California. *Transportation Research Record: Journal of the Transportation Research Board*, *2276*(1), 26-37.

Li, H. and Lahiri, P. (2010). An adjusted maximum likelihood method for solving small area estimation problems. Journal of Multivariate Analysis, 101, pp. 882-892.

Little, R. J. A. (1991). Inference with survey weights. J. Official Statistics 7. pp. 405–424.

Lu, H. and Gelman, A. (2003). A method for estimating design based sampling variances for surveys with weighting, poststratification and raking. J. Official Statistics 19. pp. 133–151.

Ma, J. and K. G. Goulias (1997) Systematic self-selection and sample weight creation in panel surveys: The Puget Sound transportation panel case. *Transportation Research,* 31A (5), pp. 365-375.

Meng, X. L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*(2), 267-278.

Mohammadian, A., Zhang, Y., 2007. Investigating the transferability of national household travel survey data. Transportation Research Record 1993, 67–79.

Molina, I. and Rao, J. N. K. (2010). Small area estimation of poverty indicators. Canadian Journal of Statistics, 38, pp. 369-385.

National Cancer Institute (2010) Model-Based Small Area Estimates of Cancer Risk Factors & Screening Behaviors [homepage on the Internet]. National Cancer Institute (U.S.); [cited 2013 June 30]. Methodology for the Model-Based Small Area Estimates. Available from: http://sae.cancer.gov/understanding/methodology.html.

NUSTATS (2013) 2010-2012 California Household Travel Survey Final Report: *Version 1.0. June 14. Submitted to the California Department of Transportation. Austin, TX.*

Pendyala, R. M., K. G. Goulias, and R. Kitamura (1993). Development of Weights for a Choice-Based Panel Survey Sample with Attrition. Transportation Research A, Vol. 27A, No. 6, pp. 477–492.

Pfeffermann, D. (1993). The role of sampling weights when modeling survey data. International Statistical Review 61, pp. 317–337.

Pfeffermann, D. and Sverchkov, M. (2007). Small-area estimation under informative probability sampling. Journal of the American Statistical Association, 102, pp. 1427-1439.

PopGen 1.1 (2011) A Synthetic Population Generator for Advanced Microsimulation Models for Travel Demand. Arizona State University. Tempe, AZ.

Reuscher, T.R., Schmoyer, R.L., Hu, P.S., 2002. Transferability of nationwide personal transportation survey data to regional and local scales. Transportation Research Record 1817, 25–32.

Rubin, D. B. (2009). *Multiple imputation for nonresponse in surveys* (Vol. 307). Wiley.

Schafer, J. L. (2010). *Analysis of incomplete multivariate data*. CRC press.

Simpson, L., & Tranmer, M. (2005). Combining sample and census data in small area estimates: iterative proportional fitting with standard software. *The Professional Geographer*, *57*(2), pp. 222-234.

Sinha, S. K. and Rao, J. N. K. (2009). Robust small area estimation. Canadian Journal of Statistics, 37, pp. 381-399.

Solon, G., S. J. Haider, and J. Wooldridge (2013) What Are We Weighting For? NBER Working Paper No. 18859.

Tzavidis, N., Marchetti, S. and Chambers, R. (2010). Robust estimation of small-area means and quantiles. Australian and New Zealand Journal of Statistics, 52, pp. 167-186.

Winship, C., Radbill, L. (1994). Sampling Weights and Regression Analysis. Sociological Methods & Research 23(2), pp. 230–257.

Ye, X., K.C. Konduri, R.M. Pendyala, B. Sana, and P. Waddell (2009) A Methodology to Match Distributions of Both Household and Person Attributes in the Generation of Synthetic Populations. DVD Compendium of Papers of the 88th Annual Meeting of the Transportation Research Board, TRB, Washington, D.C.

You, Y., Rao, J. N. K. and Hidiroglou, M. (2012). On the performance of self benchmarked small area estimators under the Fay-Herriot area level model. Survey Methodology, 39(1), pp. 217-229.