

Sample Weighting and Expansion

**California Household Travel Survey 2012/13
for the
California Statewide Sample**

Planning Section
Metropolitan Transportation Commission
101 Eighth Street
Oakland, California 94607

October 2013

Table of Contents

I. Introduction	1
II. Strategy and Data Preparation	3
A. Definition of Weighting Districts	4
B. Imputation of Missing Values for Critical Variables	4
C. Assembly of Relevant Census 2010 and American Community Survey Data.....	5
D. Definition of Cross-Validation Tests	6
III. Exploring Census and Survey Characteristics (Appendix A).....	7
IV. Evaluation of Weighting Methods: Combined Sample (Appendix B).....	9
V. Evaluation of Weighting Methods: Weekday Sample (Appendix C)	12
VI. Evaluation of Weighting Methods: Saturday Sample (Appendix D).....	14
VII. Evaluation of Weighting Methods: Sunday Sample (Appendix E).....	15
VIII. Conclusions and Next Steps.....	16
IX. References	18

List of Tables and Figures

Figures/Maps

Figure 1. California “Super Counties” 19

Tables in Body of Report

Table 1. CHTS 2012/13 Bay Area, Non-Bay Area and Statewide Sample Households by Day of Week

Appendices

- Appendix A. Exploring Census and Survey Characteristics
- Appendix B. Household Level Model Validation: Combined Sample
- Appendix C. Household Level Model Validation: Weekday Sample
- Appendix D. Household Level Model Validation: Saturday Sample
- Appendix E. Household Level Model Validation: Sunday Sample

I. INTRODUCTION

This working paper is the first in a series for documenting procedures and results of the year 2012/2013 California Household Travel Survey for the California Statewide Sample (CHTS12/13). The purpose of this working paper, *Sample Weighting and Expansion*, is to describe procedures for weighting and expanding CHTS12/13 household and person files. Results of this sample weighting and expansion process are included in this working paper.

Four sets of weights are developed for this study:

- 1) Average Daily weights (for the combined samples);
- 2) Average Weekday weights (for the Monday through Friday samples);
- 3) Average Saturday weights (for the Saturday sample); and
- 4) Average Sunday weights (for the Sunday sample).

Working papers such as this report tend to be a “work in progress” and may be updated to incorporate other improvements, clarifications and analyses. Please check with MTC to obtain the most current version of this and other working papers.

What is Sample Weighting and Expansion? What is “Raking”?

Sample weighting is a technical necessity to account and correct for geographic and demographic biases in a survey. *Sample expansion*, on the other hand, is the process used to factor up survey records to represent aggregate demographic and travel characteristics. The weighting factors used in this analysis are essentially combined weighting and expansion factors.

“*Raking*” is a survey weighting methodology that uses different sets, or levels, of marginal control totals (typically from census data) to achieve a balanced representation of the population totals.

Weighting factors are applied to produce regional, aggregate estimates of travel by day of week, by trip purpose, by travel mode, by time of day and by market segment. The reader and data user should also recognize that even though CHTS12/13 is a very large survey of over 42,400 households (in California), it is still a “small sample survey” where the main intent and purpose of this data is for the estimation of disaggregate travel behavior models.

The CHTS2012/13 is the latest generation of household travel surveys conducted in California. The 2012/13 survey was a one-day travel/activity data from 42,431 California households. This included 9,719 sample households in the San Francisco Bay Area. Data was collected between February 1, 2012 and January 31, 2013. The non-Bay Area sample was collected for all 366 days, including weekends and holidays. The Bay Area

“add-on” sample was restricted to Tuesdays through Thursdays, over the same 12 month data collection period. NuStats Inc., of Austin, Texas, conducted CHTS 2012/13.

Table 1
CHTS 2012/13 Bay Area, Non-Bay Area and Statewide Sample Households by Day of Week

Day of Week	Sample Bay Area Households	% of Total	Sample Non-Bay Households	% of Total	Sample Calif. Total Households	% of Total
Monday	775	8.0%	5,013	15.3%	5,788	13.6%
Tuesday	2,149	22.1%	3,970	12.1%	6,119	14.4%
Wednesday	2,124	21.9%	4,014	12.3%	6,139	14.5%
Thursday	2,160	22.2%	4,092	12.5%	6,252	14.7%
Friday	878	9.0%	5,040	15.4%	5,918	14.0%
Saturday	717	7.4%	5,262	16.1%	5,979	14.1%
Sunday	916	9.4%	5,320	16.3%	6,236	14.7%
TOTAL	9,719	100.0%	32,712	100.0%	42,431	100.0%
Weekday Total	8,086	83.2%	22,130	67.6%	30,216	71.2%
Weekend Total	1,633	16.8%	10,582	32.4%	12,215	28.8%
Tuesday-Thursday Total	6,433	66.2%	12,077	36.9%	18,510	43.6%

The proposed “combined sample / average daily” and “average weekday” raking scheme for the CHTS2012/13 statewide sample has six raking levels:

- 1) County (58) by Tenure (2);
- 2) County (58) by Age of Householder (5);
- 3) County (58) by Minority Status of Householder (2);
- 4) County (58) by Vehicles Available in Household (4);
- 5) Super-County (41) by Workers in Household (4); and
- 6) County (58) by Number of Persons in Household (5).

The census “marginal control totals” for expanding the CHTS 2012/13 are based on the Census 2010 “short form” data; and the Census Bureau’s American Community Survey (ACS) PUMS data, from the 2007/11 five-year PUMS database (the super-county-level rakes for workers in the household).

The “super-county” is constructed from the Public Use Microdata Sample areas, and is useful for weighting the smallest population counties. The 41 California super-counties are defined based on the Census 2000-based PUMAs, and are either the aggregations of PUMAs to the largest 34 counties, or the multi-county PUMAs.

Twenty-four of the 58 California counties are nested into seven “super-counties” as follows:

1. Del Norte / Siskiyou / Modoc / Lassen (4 counties)
2. Trinity / Tehama / Glenn / Colusa (4 counties)
3. Plumas / Sierra / Nevada (3 counties)
4. Yuba / Sutter (2 counties)
5. Mendocino / Lake (2 counties)
6. Monterey / San Benito (2 counties)
7. "Sierra Nevada" – Alpine, Amador, Calaveras, Inyo, Mono, Mariposa, Tuolumne (7 counties)

The proposed "average Saturday" and "average Sunday" raking scheme for the CHTS2012/13 statewide sample has six raking levels:

- 1) County (58) by Minority Status of Householder (2);
- 2) Super-County (41) by Age of Householder (5);
- 3) Super-County (41) by Vehicles Available in Household (4);
- 4) Super-County (41) by Workers in Household (4);
- 5) County (58) by Tenure (2); and
- 6) Super-County (41) by Number of Persons in Household (5).

The Saturday and Sunday weighting schemes are identical in structure to the Bay Area weekend weights documented in separate reports.

Six sets of weighting models are examined in this report:

- 0) Model #0. This is the consultant-estimated weights for the "combined sample / average daily sample";
- 1) Model #1. Weights for the "combined sample / average daily sample", but excludes the super-county by workers in household raking level;
- 2) Model #2. Weights for the "combined sample / average daily sample" that includes super-county by workers in household raking level;
- 3) Model #3. Weights for the "average weekday sample" with the same structure as model #2;
- 4) Model #4. Weights for the "average Saturday sample"
- 5) Model #5. Weights for the "average Sunday sample."

All of the models from #1 to #5 have corresponding "constrained" weights which are follow-up adjustments to extremely low and extremely high weights due to the nature of the complex raking schemes.

II. STRATEGY AND DATA PREPARATION

The strategy for CHTS12/13 was to implement a raking methodology based on previous efforts, and to take advantage of available computer program "macros" that simplifies the raking model application.

One of the key issues is the various date frames for the various elements: the survey was conducted in 2012-2013; the decennial Census 2010 short form is based on population as of April 1, 2010; and the American Community Survey (ACS) is a continuous sample conducted on a full-time basis since 2006.

Key issues include:

- A) Definition of weighting districts;
- B) Imputation of missing values for critical variables;
- C) Assembly of relevant census data;
- D) Definition of cross-validation tests.

A. Definition of Weighting Districts

The strategy for the weighting of CHTS12/13 is to use the 58 California counties, or the 41 California super-counties, for the weighting districts.

B. Imputation of Missing Values for Critical Variables

Imputation is the process of “filling in” data where there are missing values for variables of interest. The variables at the sample weighting and expansion stage of survey analysis are: geography of residence, household size, vehicles in household, tenure, workers in household, age of persons in household, age of householder, and race/ethnicity of householder. Some of these variables had no non-response issues (geography, household size, vehicles in household, workers in household).

Imputation was conducted separately on the Bay Area and the non-Bay Area samples. Results reported in the following paragraphs are for the non-Bay Area samples.

Tenure (owner-occupied versus renter-occupied households) is available on 32,609 of the 32,712 sample non-Bay Area households (99.7% response rate, and a 0.3% non-response rate). A simple deductive imputation was used for tenure, based on the detailed descriptions of the “other type of tenure” variable (O_OWN on the survey records). Basically the 105 sample households were assigned to “renter households.”

Race/ethnicity is coded on 83,457 out of 85,083 sample non-Bay Area persons (98.1% completion rate, or a 1.9% non-response rate). On a householder basis, 32,073 out of 32,712 householders have race/ethnicity coded (a 98.0% response rate). A “hot deck” imputation model was run for the five race/ethnicity categories, based on the SAS macro produced by Ellis (1). The race/ethnicity hot deck imputation used household size (5), the county-of-residence (58), and a random sort variable, as the imputation model.

Age is coded on 82,308 out of 85,083 sample non-Bay Area persons (96.7% completion rate, or a 3.3% non-response rate). A “hot deck” imputation model was produced on four sub-sets of the sample person files:

- 1) Students;
- 2) Workers;
- 3) Non-Workers;
- 4) Other/Not-Classified.

Student age was imputed using school level (e.g., high school students are very likely 14 to 17 years of age; college students are very likely 17 to 24 years of age, etc.)

The other groups were imputed based on the “relationship to head of household” variable. No geography classes were used in the hot deck imputation for age.

Household income is a critical variable that will require imputation for missing values, not only in terms of the household income category, but for the discrete household income values, as well.

A separate technical report in imputation for non-response will be prepared to fully document the imputation process and results.

C. Assembly of Relevant Census 2010 and American Community Survey Data

There is a temporal disconnect between the time period that the CHTS 2012/13 data was collected (February 2012 through January 2013) and available “marginal control total” data from the census. The strategy, at present time, is to weight and expand the 2012/13 survey to represent 2010 population characteristics. An option, a few years from now when 2012/13 ACS data is available, is to re-weight the survey to approximate the 2012/13 population characteristics.

The 2010 decennial Census is a 100 percent count of the American population as of April 1, 2010. Data from Census 2010 is only “short form” data, and doesn’t collect data on elements such as household vehicles, workers in household, journey-to-work, income, etc.

Census 2010 data elements of interest include: households, population in households, households by household size, households by age of householder, households by race/ethnicity of householder, and tenure.

The American Community Survey (ACS) is the new (2006 to present) census program that replaces the decennial census “long form.” Several options are available to this analysis: using the most recent five-year ACS data (2007-2011); the most recent three-year ACS (2009-2011); or any of the one-year ACS data products (2006 through 2011). Appendix Table A.1 through A.4 summarizes county and region-level number of households from the various census / ACS databases.

The weighting schemes require the use of Public Use Microdata Sample (PUMS) data, since there are no standard ACS tables on the distribution of households by workers in the household. (There are standard “American Factfinder” tables in the ACS that show households by the number of “workers” in the household. But the Census Bureau has defined the “workers in household” based on whether the worker reported going to work (“commuting”) during the reference week, thus neglecting to include “weekly absentees” as workers. The Census Bureau’s American FactFinder tables are more accurately described as “households by commuters in household.”)

There are multiple versions of ACS PUMS data: one-year PUMS (various years), three-year PUMS (various years), and five-year PUMS (2006-2010, and 2007-2011). The approach here is to use the largest, richest, most recent PUMS database, the 2007/11 PUMS. (The 2008/12 PUMS dataset is expected in March 2014).

D. Definition of Cross-Validation Tests

Cross-validation, in the context of travel survey analysis, is the process of comparing the expanded/weighted survey to other, independent variables not used in weighting, to gauge the quality of the weighting scheme. For example, for a weighting scheme based on geography, tenure, households by household size and households by vehicles available, relevant census-based cross-validation variables include:

- * Race/ethnicity of household population;
- * Household population by age and sex;
- * Student population by age, sex, and school status.

The results of the cross-validation may suggest the need to add a set of temporary or permanent adjustment/correction factors.

For this report on the California statewide sample and weights, no cross-validation tests were conducted. (This should be considered as a follow-up exercise.)

III. EXPLORING CENSUS AND SURVEY CHARACTERISTICS (APPENDIX A)

The purpose of this section is to detail the exploratory analysis of potential census “marginal control total” data, and the corresponding patterns from the CHTS 2012/13 in California. This will highlight the critical biases in the survey that are correctable using appropriate weighting schemes.

Detailed data tables included in Appendix A are reported in this section. As appropriate, census data sources (Census 2010 “short form” data versus American Community Survey 2007/11) are cited.

The survey provides detailed information on 42,431 California households. This is out of the 12,577,498 households counted in Census 2010, for a “sampling rate” of 0.34% (42431 / 12577498) (Table A.5). The simplest expansion model is to apply a weight of 296.4 to all households.

The sampling rate ranges from a low of 0.16% in Orange and San Diego Counties to 4.23% in Alpine County (Table A.5). The number of sample households ranges from 21 in Alpine County to 8,219 in Los Angeles County.

Tenure (owner-occupied versus renter-occupied households) is the most severe, though correctable bias in the household travel survey. The sampling rate of renter-occupied households is 0.17 percent, comparing to 0.47 percent of owner-occupied households (Table A.6). T

The consultant report details efforts to contact and recruit difficult-to-reach households. And the survey literature abounds with new evidence on the proliferation of cell-phone only households, and the difficulty in recruiting these households for social science surveys.

Household travel surveys tend to under-sample the very small one-person households, and the very large five-or-more person households. The sampling rate ranges from a low of 0.20 percent of five-or-more person households, to a high of 0.45 percent of two person households (Table A.9). One-person households have the second lowest sampling rate at 0.31 percent.

Analyzing households by the age of the householder is a new strategy employed by MTC in evaluating travel surveys. (The “householder” is the first person listed in either the Census records, or the survey records.) The Census 2010 provides standard tables on households by age of householder, broken down by ten year cohorts: age 15 to 24, age 25 to 34, etc., up to age 85-and-over householders. The Census 2010 data was aggregated into five age categories which are all approximately 18 to 22 percent of total households in California. The sampling rate ranges from a low of 0.15 percent of the

youngest householders (age 15 to 34); and a high of 0.58 percent of the householders age 55 to 64.

Analysis of households by the race and ethnicity of the householder was based on Census 2010 data, and survey data, aggregated to five major categories:

- 1) White, non-Hispanic householders;
- 2) Black, non-Hispanic householders;
- 3) Asian / Native Hawaiian or Other Pacific Islander, non-Hispanic householders;
- 4) Other (other race, American Indian/Alaskan Native, two-or-more race), non-Hispanic householders; and
- 5) Hispanic/Latino householders (any race).

Data from the Census, and the travel surveys collect separate information on the “race” of persons (white, black, Asian, other); and the Hispanic status (“ethnicity”) of the person (Hispanic/Latino, or not Hispanic/Latino).

White householders are 52 percent of the Census 2010 population; and 69 percent of the CHTS 2012/13 sample (Table A.13). This is an obvious oversampling which is correctable using appropriate sampling weights.

Data from the 2007/11 American Community Survey (ACS) Public Use Microdata Sample (PUMS) was used to analyze households by number of workers in the household. Standard tables (from American Factfinder) were extracted from the 2007/11 ACS for the households by number of vehicles available. (The statewide totals for the ACS household are slightly different than the Census 2010.)

The survey tends to under-sample households with three-or-more worker households (0.24%), and over-sampling the one-worker households (0.37 percent) (Table A.19).

Zero vehicle households (0.26 percent) are the most under-sampled compared to 0.46 percent of two-vehicle households (0.40 percent) (Table A.16).

Zero vehicle households are 7.7 percent of California households according to the 2007/11 ACS. This compares with 5.8 percent of the CHTS 12/13 sample households. This again is an important, but correctable bias.

The ordering of the raking levels is not important, except for the final, last raking level. The “last rake” will show the best fit, comparing the modeled, expanded households to the “marginal control totals.” For previous and current MTC weighting approaches, the focus is on obtaining accurate estimates of households by geography by household size, in order to ensure the best approximation of total household population.

The first raking level, county by tenure, is based on Census 2010-based marginal control totals (Table A.6). For the “combined sample / average daily” model, there are no problems with missing values of sample households by county by tenure.

The second raking level is county by age of householder (Tables A.10-A.12). There are no sample Alpine County households with householders age 15 to 34, so the marginal control total is collapsed/combined for Alpine, for householders age 15 to 44.

The third raking level is county by minority status of householder (white, not-hispanic versus all minority groups, combined) (Table A.13). For the “combined sample / average daily” model, there are no problems with missing values of sample households.

The fourth raking level is county by number of vehicles in household (Table A.14-A.16). Two counties, Alpine County and Mariposa County, have no sample zero-vehicle households. So, the marginal control totals are collapsed for Alpine and Mariposa, to represent the 0-and-1 vehicle households.

The fifth raking level is super-county by workers in household. (Table A.17-A.19). There are no missing values, so no need to collapse marginal control total categories.

The sixth and last raking level is county by household size (Table A.7-A.9). The only problem is no 5+ person households sampled for Sierra County. So, the marginal control totals for Sierra are collapsed/combined to 4+ person households.

This collapsing, or combining of marginal control totals, is a technical necessity since raking procedures won’t work if there are marginal control totals, and nothing to “rake”. (The raking procedures might work if there are samples without corresponding marginal control totals, but the careful imputation for missing values controlled for this possibility).

IV. EVALUATION OF WEIGHTING MODELS: COMBINED SAMPLE (APPENDIX B)

Five weighting models, for the “combined sample / average daily sample” are examined in this section. Detailed data tables are included in Appendix B.

The study consultant developed a set of weights, which are denoted as “Model #0” weights in this report. MTC staff developed four sets of raking models/weights, denoted as Model #1, Model #1c, Model #2, and Model #2c. The “c” stands for “constrained”.

Model #0 weighting procedures are documented in the study consultant report (2). It is a five-level raking scheme, conducted at the statewide level for all households. The raking levels are:

- 1) Statewide households (1) by household size (4);

- 2) Statewide households (1) by household income (6);
- 3) Statewide households (1) by workers in household (4);
- 4) Statewide households (1) by vehicles in household (4); and
- 5) County-of-residence (58)

In Model #0 the statewide “marginal control totals” are derived from the one-year 2011 American Community Survey (ACS). The county-of-residence “marginal control totals” are from the 2007/11 (five-year) ACS. Though not explicit in the documentation, these are standard tables from the American Community Survey available from American FactFinder. It isn’t apparent that PUMS was used in the study consultant’s analysis.

(As such, the “workers per household” available from ACS standard tabulations is more precisely “commuters per household” since the Census Bureau is excluding 2 to 3 percent of the workers who are “weekly absentees”, in the definition of “workers per household”).

Household income is a variable with a typically larger share of item non-response, about 10 percent non-response. The study consultant performed an imputation on household income, using a mean of household income based on a combination of tenure, household size and vehicle availability. “A mean of each combination was calculated and applied to the refused income values for the relevant category.” (2) (This is a little confusing, since there are no “mean household incomes” included in the survey file; only “household income categories”.)

The study consultant also produced person weights, starting with the household weights, and raking for different characteristics. This means that there are different person weights within a multi-person household. The person-level raking levels used in Model #0 are:

- 1) Statewide persons in household (1) by Hispanic/Latino status (2);
- 2) Statewide persons in household (1) by Race (4);
- 3) Statewide persons in household (1) by Age (5);
- 4) Statewide persons in household (1) by Employment Status (2); and
- 5) County-of-residence (56) (Alpine, Amador and Calaveras Counties were combined).

The study consultant documentation also provides background on the imputation procedures used for race/ethnicity and age.

Model #1 and #2 weights were produced by MTC staff adapting SAS “raking” macros produced by Izrael, Hoagland and Battaglia (3).

Model #1 includes five raking levels:

- 1) County (58) by Tenure (2);
- 2) County (58) by Age of Householder (5);
- 3) County (58) by Minority Status of Householder (2);
- 4) County (58) by Vehicles in Household (4);
- 5) County (58) by Household Size (5).

Five sets of marginal control total files are required for this five level raking model. The two variables required in these files are an index variable (denoting the categories), and the control total value. The index variable is a composite of the two categories used in any given level, for example:

County * 10 + Tenure * 1.

An index of "11" in this example is Alameda County households (county=1), owner-occupied (tenure=1). An index of "1152" in this example is Yuba County households (county=115), renter-occupied (tenure=2), Census FIPS codes are used for county codes.

The weights of the Model #1 raking were evaluated, examining the extreme values at both the high and low ends. Extremely low weights, say less than 1.0, imply that the "sample household is so common that it shouldn't have been sampled." This is ridiculous, and very low weights suggest that those households are irrelevant to any analysis. On the other hand, very high weights may be necessary for rarer households, say, renter-occupied, minority, and very large households with very few vehicles. The decision was to develop a weighting model to complement Model #1 by placing a floor (2.0) and ceiling (2500.0) to Model #1 weights. This is denoted as Model #1c (or Model #1, constrained).

The statistical distribution (mean, median, minimum, maximum, 1st percentile, 90th percentile) by county for models #0, #1, and #2 are summarized in Tables B.1 through B.3.

Model #0 weights range from a low of 0.75 to a high of 1717.8. Model #1 weights ranged from 0.28 to 4627.2; and model #2 weights ranged from 0.31 to 5220.8.

Results of Model #1c appeared promising, but lacked control of the workers in household variable as included in previous MTC work. The challenge was to use a new geography, the "super-county" to provide relevant ACS PUMS data on households by workers in household.

Model #2 simply adds a "sixth" raking level to Model #1. The six raking levels in Model #2 are:

- 1) County (58) by Tenure (2);
- 2) County (58) by Age of Householder (5);

- 3) County (58) by Minority Status of Householder (2);
- 4) County (58) by Vehicles in Household (4);
- 5) Super-County (51) by Workers in Household (4); and
- 6) County (58) by Household Size (5).

Similar to Model #1, Model #2 weights were constrained using floors and ceilings. Model #2c constrains the weights to a 2.0 floor and 2500.0 ceiling.

Statewide validation results for Model #0, #1 and #1c are shown in Table B.4.1 – B.4.5 and county-level results in Table B.5. Both models #1 and #1c show a very good match on statewide households by tenure, by minority status of householder, by household size, by age of householder, and by vehicles in household. The Model #1 of households by household size is almost a perfect match to Census 2010 control totals since households by county by household size is the last raking level included in this model. Model #0 significantly over-estimates owner-occupied households, under-estimates minority households, underestimates younger (15-34) households, and over-estimates middle-aged (55-64) households. Model #0 does a fairly good job on statewide households by household size and households by vehicles in household.

Statewide validation results for Model #0, #2 and #2c are shown in Table B.6.1 – B.6.6, and county-level results in Table B.7. The results at the statewide and county level are very satisfactory.

Appendix Table B.8 through B.15 is a summary of the “person correction factors” used to adjust the household weights for the very large (five-or-more person) households. This technique was also used in adjusting previous Bay Area Travel Surveys. This is a fix since the average size of 5+ persons in the Census (6.063 persons/5+ person household) is slightly higher than the average size in the weighted survey (5.584 persons/5+ person household). These correction factors are calculated at the county-of-residence, and may range from 0.92 to 1.20. These person correction factors are applied equally to all members within the 5+ person household, ensuring that the final person factors do not vary within each sample household.

V. EVALUATION OF WEIGHTING MODELS: WEEKDAY SAMPLE (APPENDIX C)

Two weighting models (Model #3, #3c), for the “average weekday sample” are examined in this section. Detailed data tables are included in Appendix C.

It is necessary to develop separate weights for households reporting weekday travel (as opposed to weekend travel) to estimate reliable aggregate estimates of average weekday travel. This also supports the travel model development efforts to base the disaggregate travel models on weekday patterns, supported by available weekday transit and highway levels-of-service databases.

Weekend travel data is useful in obtaining data on average Saturday and average Sunday travel patterns (VMT, trips by travel purpose by time of day, modal shares, etc.) and is useful in developing annualization factors to convert average weekday travel behavior forecasts to average daily (or annual) travel estimates. Rarely, if ever, are scarce resources spent on estimating disaggregate travel models for Saturday or Sunday travel.

In the CHTS 2012/13, 30,216 sample households out of 42,431 sample households provided weekday travel data. The weekday sample households in the Southern California (N=10,536) and San Francisco Bay Area (N=8,086) are very likely large enough to support the estimation of disaggregate travel demand models. The weekday sample households in the San Diego region (N=1,142), Sacramento region (1,621) and the Monterey Bay region (N=1,380) may or may not be sufficient to estimate robust disaggregate travel demand models. The sample sizes in the other California counties is probably sufficient for estimating overall aggregate travel patterns.

The comparison of census-based marginal control totals to the weekday sample households is a technical necessity to determine if there are additional missing values that need to be combined/collapsed (Tables C.1 through C.15).

Several additional marginal control totals required collapsing in the weekday sample, due to zero sample households:

- 1) No minority households in Sierra County;
- 2) No 3+ vehicle households in Alpine County;
- 3) No zero-vehicle households in Modoc County; and
- 4) No zero-vehicle households in Mono County.

Model #3 is the “average weekday model” and has the same raking levels as the “combined sample” Model #2. The six raking levels in Model #3 are:

- 1) County (58) by Tenure (2);
- 2) County (58) by Age of Householder (5);
- 3) County (58) by Minority Status of Householder (2);
- 4) County (58) by Vehicles in Household (4);
- 5) Super-County (51) by Workers in Household (4); and
- 6) County (58) by Household Size (5).

Similar to previous models, Model #3 weights were constrained using floors and ceilings. Model #3c constrains the weights to a 2.0 floor and 4000.0 ceiling.

The Model #3 weights range from a low of near zero ($2.96E-4$ in Alpine County) to a high of 8,042.5 in Sacramento County (Table C.16). Counties with very low (<2.0) weights include Alpine, Colusa, Del Norte, Sierra and Trinity Counties. Counties with very high

weights include Sacramento (8,042.5), Orange (7,231.4), San Diego (5,728.3) and San Francisco (4,806.4).

Statewide validation results for Model #3 and #3c are shown in Table C.17.1 – C.17.6, and county-level results in Table C.18. The results at the statewide and county level are very satisfactory.

The last set of tables in Appendix C (Table C.19 – C.22) document the “person correction factors” used to adjust the person weights in the very large 5+ person households, again, applied at the 58 county level.

VI. EVALUATION OF WEIGHTING MODELS: SATURDAY SAMPLE (APPENDIX D)

Two weighting models (Model #4, #4c), for the “average Saturday sample” are examined in this section. Detailed data tables are included in Appendix D.

The weighting for the Saturday data is based on 5,979 sample households in the statewide survey. This is a simple sampling rate of 0.05 percent of California households, with a simple statewide weight of 2,103.6.

Model #4 and #4c has six raking levels:

- 1) County (58) by Minority Status of Householder (2);
- 2) Super-County (41) by Age of Householder (5);
- 3) Super-County (41) by Vehicles Available in Household (4);
- 4) Super-County (41) by Workers in Household (4);
- 5) County (58) by Tenure (2); and
- 6) Super-County (41) by Household Size (5).

The decision to use the super-county versus county geographic level in the Saturday raking was based on a thorough review of the sample data at both county and super-county level (Table D.1 through D.15). Two raking levels: households by minority status of householder and households by tenure, could readily support the county-level rakes. Even so, the Saturday raking scheme has twelve occurrences where marginal control totals needed combining due to lack of survey samples.

The statistical properties of the Model #4 weights, by county (mean, median, minimum, maximum, 1st percentile, 90th percentile) is summarized in Table D.16. The weights range from a low of near zero (2.45E-8) in San Mateo County to a high of 43,860.4 in Contra Costa County.

The decision on setting floors and ceilings on the Model #4 weights was to use just a 2.0 floor, with no cap/ceiling on the weights. Various alternatives were tested, but the 2.0

floor / no cap alternative does the best job in retaining the total number of California households.

Statewide household validation of the Saturday weighting models is provided in Tables D.17.1 through D.17.6. Saturday weighting results at the county level are shown in Table D.18. The results are very satisfactory.

A “person correction factor” of 1.07825 is applied to the Saturday, 5+ person household weights to derive a person weight. This was not done at the county-level (as was done for the combined sample and weekday sample) due to the smaller Saturday sample size.

VII. EVALUATION OF WEIGHTING MODELS: SUNDAY SAMPLE (APPENDIX E)

Two weighting models (Model #5, #5c), for the “average Sunday sample” are examined in this section. Detailed data tables are included in Appendix E.

The weighting for the Sunday data is based on 6,236 sample households in the statewide survey. This is a simple sampling rate of 0.05 percent of California households, with a simple statewide weight of 2,016.9.

The structure of the Sunday weighting models is the same as the Saturday weighting models. Model #5 and #5c has six raking levels:

- 1) County (58) by Minority Status of Householder (2);
- 2) Super-County (41) by Age of Householder (5);
- 3) Super-County (41) by Vehicles Available in Household (4);
- 4) Super-County (41) by Workers in Household (4);
- 5) County (58) by Tenure (2); and
- 6) Super-County (41) by Household Size (5).

The Sunday raking scheme has nine occurrences where marginal control totals needed combining due to lack of survey samples. Four of these categories are due to lack of Sunday zero-vehicle households in four super-counties: Butte, Marin, Napa, and the Plumas/Sierra/Nevada super-county.

The statistical properties of the Model #5 weights, by county (mean, median, minimum, maximum, 1st percentile, 90th percentile) is summarized in Table E.16. The weights range from a low of near zero (2.54E-6) in Mendocino County to a high of 35,559.1 in Santa Clara County.

The decision on setting floors and ceilings on the Model #5 weights was to use just a 2.0 floor, with no cap/ceiling on the weights. Various alternatives were tested, but the 2.0 floor / no cap alternative does the best job in retaining the total number of California households.

Statewide household validation of the Sunday weighting models is provided in Tables E.17.1 through E.17.6. Sunday weighting results at the county level are shown in Table E.18. The results are very satisfactory.

A “person correction factor” of 1.09322 is applied to the Sunday, 5+ person household weights to derive a person weight. This was not done at the county-level (as was done for the combined sample and weekday sample) due to the smaller Sunday sample size.

VIII. CONCLUSIONS AND NEXT STEPS

The recommendation is that Model #2c weights are the final weights on the statewide households in the CHTS 2012/13 database. This is for the “combined” sample (N=42,431 sample households) which includes travel for both weekdays and weekend days. Model #2c applies a 2.0 floor and 2500.0 cap/ceiling to the Model #2 weights.

Model #3c is the recommended model for the “weekday sample” of the California Household Travel Survey (N=30,216). Model #3c applies a 2.0 floor and a 4,000.0 cap/ceiling to the Model #3 weights.

Model #4c is the recommended model for the “Saturday sample” of the CHTS 12/13 (N=5,979). Model #4c applies a 2.0 floor, and no ceiling, to the Model #4 weights.

Model #5c is recommended for the “Sunday sample” of the CHTS 12/13 (N=6,236). Model #4c applies a 2.0 floor, and no ceiling, to the Model #4 weights.

Future steps in the analysis of the CHTS 2012/13 travel survey include detailed processing of the unlinked trip records to produce a linked trip, tour and sub-tour files. The product of this trip linking/chaining process will be both traditional linked trip files, as used in trip-based travel demand models; and tour-based travel files, for supporting the current and future generation of travel behavior models.

Procedures to impute missing values will be documented in separate technical reports. Other “data cleaning” notes will be included in MTC staff notes and technical documentation.

The recommended household and person-level weights will be extracted and provided to CHTS 2012/13 data users. Appropriate metadata will be developed to assist the data user. In addition, the MTC produced weights for the San Francisco Bay Area sample will be merged with the California Statewide weights, to produce a consolidated file with all appropriate weights. The Bay Area raking models were more detailed, with raking

conducted at the 55 PUMA level in the nine Bay Area counties, as opposed to just the county-level in the statewide weights.

Further research on raking methods will be undertaken as time permits. Options may include simplifying some of the two-dimensional raking schemes to analyze the impacts on extreme weights, and raking model closure.

Procedures to impute missing trips and tours may be required, and will probably be included in future technical reports.

Other follow-on work that should be considered include:

- 1) Evaluation of statewide, regional and county population by age and sex;
- 2) Evaluation of statewide, regional and county population by race/ethnicity;
- 3) Evaluation of statewide, regional, county population of students and workers;
- 4) Evaluation of PUMAs (or Super-PUMAs) as weighting districts for Los Angeles County;.

IX. REFERENCES

1. Bruce Ellis "A Consolidated Macro for Iterative Hot Deck Imputation" Proceedings of the North East SAS Users Group (NESUG), 2007.
(<http://www.nesug.org/proceedings/nesug07/po/po03.pdf>)
2. "2010-12 California Household Travel Survey: Final Report" Versions 1.0, NuStats Research Solutions, Austin, Texas, June 14, 2013.
3. David Izrael, David Hoagland and Michael Battaglia "A SAS Macro for Balancing a Weighted Sample" Paper #258-25, Proceedings of the Twenty-Fifth Annual SAS Users Group International Conference, Cary, NC, 2000.
(http://www.abtassociates.com/PDFS/258_25.aspx)