

**Sample Weighting and Expansion  
Part I: Average Daily Weights**

**California Household Travel Survey 2012/13  
for the  
San Francisco Bay Area**

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

September 2013

## **Table of Contents**

I. Introduction .....	1
A. What is Sample Weighting and Expansion? What is Raking? .....	1
B. Household Travel Surveys in the San Francisco Bay Area (1965-2013) .....	1
II. Strategy and Data Preparation .....	5
A. Definition of Weighting Districts .....	6
B. Imputation of Missing Values for Critical Variables .....	6
C. Assembly of Relevant Census 2010 and American Community Survey Data.....	7
D. Definition of Cross-Validation Tests .....	9
III. Exploring Census and Survey Characteristics .....	10
IV. Validation and Evaluation of Weighting Methods .....	17
V. Next Steps .....	22
VI. References .....	23

## **List of Tables and Figures**

### Figures/Maps

Figure 1. Public Use Microdata Areas (PUMAs): based on Census 2010..... 24

### Tables in Body of Report

Table 1. CHTS 2012/13 Bay Area Sample Households by Day of Week

Table 2. Household Sample Weighting Methods: San Francisco Bay Area Travel Surveys: 1965 to 2013

Table 3. Share, Renter-Occupied Households of Total Households: Survey vs Census

Table 4. Raking Levels and Corresponding Appendix Tables

Table 5. Range of Weights by Weighting Model

Table 6. Percentile Distribution of Weights by Weighting Models

### Appendices

Appendix A. Exploring Census and Survey Characteristics

Appendix B. Household Level Model Validation: Models #0, #1, #1c

Appendix C. Person Level Model Validation: Models #0, #1, #1c

Appendix D. Household Level Model Validation: Models #2, #2c1, #2c2

Appendix E. Person Level Model Validation: Models #2, #2c1, #2c2

Appendix F. Person Correction Factors

## **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 conducted in the San Francisco Bay Area (CHTS12/13). The purpose of this working paper, *Sample Weighting and Expansion: Part I: Average Daily Weights*, 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 envisioned 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.

### **A. 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. Simpler versions of raking were used in expanding the BATS1990 and BATS2000 databases (BATS = Bay Area Travel Surveys); a more comprehensive raking scheme was used for expanding the CHTS12/13 data.

Weighting factors are applied to produce regional, aggregate estimates of travel 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 9,600 households (in the Bay Area), 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.

### **B. Household Travel Surveys in the San Francisco Bay Area (1965-2013)**

The 2012/13 California Household Travel Survey is the sixth in a series of major surveys conducted in the San Francisco Bay Area. The 2012/13 survey was managed as a

statewide project by the California Department of Transportation (Caltrans) with management assistance from the MTC. Previous surveys in the Bay Area were managed by the Metropolitan Transportation Commission (MTC) and the predecessor agency, the Bay Area Transportation Study Commission (BATSC) since 1965.

The 1965 BATS was conducted as a face-to-face survey of over 20,000 Bay Area households for their weekday travel, and another 10,000 households for their weekend daily travel. BATS '65 can be considered a traditional “home interview survey” in that the surveys were actually conducted at the home of the respondent. This origin-destination survey was conducted in-house by an expanded staff of the Bay Area Transportation Study Commission (BATSC).

The '65 BATS survey was weighted and expanded to 290-zone level estimates of households. No socio-economic stratifications were apparently used in the '65 BATS weighting scheme.

The 1981 BATS was the first telephone household travel survey conducted in the Bay Area. The '81 survey was a single weekday survey of 6,209 households, and 882 households for their weekend daily travel patterns. The survey was conducted by the consultant Crain and Associates of Menlo Park, California.

The '81 BATS survey was weighted and expanded to match 1980 Census counts of households by 45 weighting districts-of-residence, and by household size (1, 2, 3, 4, 5+ persons per household). The weighting districts included 15 districts within the City of San Francisco, and the 30 “superdistricts” in the other eight Bay Area counties (1).

The 1990 BATS was the next household travel survey conducted in the Bay Area. The '90 survey included a single-weekday component of 9,359 sample households and a multiple-weekday component of 1,479 households, or 10,838 total sample households. No data on weekend travel was collected in the 1990 BATS. The survey was conducted by the consultants E.H. White and Company of San Francisco with Nelson/Nygaard and Phase III Market Research.

The '90 BATS survey was weighted and expanded to match 1990 Census counts of households by 34 superdistricts-of-residence, by household size (1, 2, 3, 4, 5+ persons per household), by tenure (owner, renter) and by vehicles available in the household (0, 1, 2, 3+ vehicles/household). Data from the 1990 Census was only available for households by household size by tenure, and for households by vehicles available by tenure, but not for all three variables combined. So, an “iterative proportional fitting” (IPF) procedure (one iteration only) was used to construct the 1990 BATS weights. This

weighting method is very similar to that used in Los Angeles for their 1991 household travel survey<sup>1</sup>. (2, 3)

The 1996 BATS was the first activity survey conducted in the Bay Area. The '96 survey was part of a larger Bay Bridge Congestion Pricing Demonstration Study. The survey collected two-day weekday and weekend travel and activity diaries from 3,678 households. Of these nearly 3,700 households, 1,654 households were from a regional random "control" sample, 1,857 households were screened for Bay Bridge corridor travel ("target" sample), and 167 households were households originally recruited in the 1990 MTC household travel survey. A congestion pricing stated preference (SP) survey was administered to a subset of 150 sample households. NuStats Inc., of Austin, Texas, conducted BATS '96.

The '96 BATS survey was weighted and expanded to match estimates of 1996 households by county-of-residence. Different weights were required for the "control" sample and the "target" sample.

The BATS2000 was the second household activity survey conducted in the Bay Area within five years. The 2000 survey collected two-day weekday and weekend travel/activity diaries from 15,064 households. Previous surveys were conducted in the spring and fall of the survey year. BATS2000 data was collected on a continuous basis, excluding holidays, between February 2000 and March 2001. Morpace International of Farmington Hills, Michigan conducted BATS2000.

The BATS2000 travel survey was expanded using a three-level raking scheme:

- 1) Households by PUMA of Residence (54) by Tenure (2) by Household Size (5)
- 2) Households by PUMA of Residence (54) by Tenure (2) by Vehicles Available (5); and
- 3) Households by PUMA of Residence (54) by Race/Ethnicity of Householder (8)

The first two-levels of the BATS2000 raking scheme were run for 16 iterations; the third level was a one iteration "correction" to the initial weights based on the first two levels (4). The weighting factors were applied to the first weekday day of each sample household. In the case of households with a Sunday-Monday pairing, the Monday trips were used, for purposes of developing "average weekday travel."

The PUMA is the "Public Use Microdata Area" defined for the Census 2000 "Public Use Microdata Sample" (PUMS). In the San Francisco Bay Area, the PUMAs are simple aggregations of Census 2000 census tracts, so are a convenient "district-level"

---

<sup>1</sup> The Los Angeles 1991 survey expanded households by districts (regional statistical areas) by household size, by vehicles available in household, by structure type (single-family versus multi-family). BATS '90 used tenure instead of structure type as a weighting dimension. See: Peter R. Stopher and Cheryl Stecher "Blow Up: Expanding a Complex Random Sample Travel Survey" in Transportation Research Record 1412, TRB, 1993, pp. 10-16 (2).

geography for use in sub-county analyses. For BATS2000, data from the Census 2000 “Summary File #3”, at the census tract level, was aggregated to the 54 Bay Area PUMAs, to produce the marginal control totals for raking the survey data.

The CHTS2012/13 is the latest generation of household travel surveys conducted in the Bay Area and 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. Of the 9,719 sample households, 8,085 provide weekday travel data; 717 provide Saturday data; and 916 households provide Sunday data. NuStats Inc., of Austin, Texas, conducted CHTS 2012/13.

**Table 1**  
**CHTS 2012/13 Bay Area Sample Households by Day of Week**

Day of Week	Sample Households	% of Total
Monday	775	8.0%
Tuesday	2,149	22.1%
Wednesday	2,124	21.9%
Thursday	2,160	22.2%
Friday	878	9.0%
Saturday	717	7.4%
Sunday	916	9.4%
TOTAL	9,719	100.0%
Weekday Total	8,086	83.2%
Weekend Total	1,633	16.8%
Tuesday-Thursday Total	6,433	66.2%

The proposed raking scheme for the Bay Area CHTS2012/13 sample has seven raking levels:

- 1) County (9) by Tenure (2) by Race/Ethnicity of Householder (5);
- 2) PUMA (55) by Tenure (2) by Minority Status of Householder (2);
- 3) County (9) by Tenure (2) by Workers in Household (4);
- 4) County (9) by Tenure (2) by Vehicles in Household (4);
- 5) PUMA (55) by Tenure (2) by Age of Householder (5);
- 6) County (9) by Number of Persons Age 20-29 in Household (3); and
- 7) PUMA (55) by Tenure (2) by Household Size (5).

The census “marginal control totals” for expanding the CHTS 2012/13 are based on the Census 2010 “short form” data (the four raking levels using the Census 2010 PUMAs as weighting districts; and the first raking level on county, tenure, race/ethnicity); and the Census Bureau’s American Community Survey (ACS) PUMS data, from the 2007/11 five-

year PUMS database (the county-level rates for vehicles, workers and persons age 20-29 in the household).

Table 2 summarizes the weighting schemes used in Bay Area household travel surveys between 1965 and 2013.

**Table 2**  
**Household Sample Weighting Methods**  
**San Francisco Bay Area Travel Surveys: 1965 to 2013**

1965	Zone-of-Residence (290)
1981	District-of-Residence (45) by Household Size (5)
1990	District-of-Residence (45) by Tenure (2) by Vehicles in Household (4) District-of-Residence (45) by Tenure (2) by Household Size (5)
1996	County-of-Residence (9)
2000	PUMA-of-Residence (54) by Tenure (2) by Household Size (5) PUMA-of-Residence (54) by Tenure (2) by Vehicles Available (5) PUMA-of-Residence (54) by Race/Ethnicity of Householder (8)
2012/13	County-of-Residence (9) by Tenure (2) by Race/Ethnicity of Householder (5) PUMA-of-Residence (55) by Tenure (2) by Minority Status of Householder (2) County-of-Residence (9) by Tenure (2) by Workers in Household (4) County-of-Residence (9) by Tenure (2) by Vehicles in Household (4) PUMA-of-Residence (55) by Tenure (2) by Age of Householder (5) County-of-Residence (9) by Number of Persons Age 20-29 in Household (3) PUMA-of-Residence (55) by Tenure (2) by Household Size (5)

## **II. STRATEGY AND DATA PREPARATION**

Previous efforts at weighting and expanding Bay Area travel surveys favored an incremental approach that tested approaches ranging from the simple to the most sophisticated. 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.

The strategy was to build up a set of weights similar to previous surveys, that is: expanding households by district of residence, by household size, by vehicles available, by tenure.

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

Geographic weighting dimensions for previous Bay Area travel surveys have included travel analysis zones (290 zones in BATS '65); MTC's 34 superdistricts (BATS '90); split superdistricts in the over-sampled San Francisco County in BATS '81 (45 weighting districts); the nine Bay Area counties for the smallest survey (BATS '96); and the 54 Census 2000 PUMAs (Public Use Microdata Areas) in BATS2000.

The strategy for the weighting of CHTS12/13 is to use the Census 2010 PUMAs, defined after tract-level population data was published by the Census Bureau in 2011. After Census 2000, 54 PUMAs (minimum population of 100,000) were defined for the nine-county Bay Area. After Census 2010, 55 PUMAs were defined.

The Census 2010 PUMAs were used since the CHTS12/13 data was geo-coded to Census 2010 geography. This means that the 2010 PUMA codes could be easily appended to the CHTS12/13 data records.

One of the drawbacks to using the Census 2010 PUMAs is that data from either the Census 2010 "short form" databases; or the American Community Survey (ACS), has not been published at these new PUMAs as of summer 2013. (This is slightly confusing, since current PUMAs included in the ACS PUMS files, or Census 2010 standard tabulations, are based on the Census 2000-based PUMAs).

A map showing the Census 2010-based 55 Public Use Microdata Areas is provided in Figure 1.

#### 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).

Tenure (owner-occupied versus renter-occupied households) is available on 9,700 of the 9,719 sample Bay Area households (99.2% response rate, and a 0.8% 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 19 sample households were assigned to “renter households.”

Race/ethnicity is coded on 23,494 out of 24,030 sample Bay Area persons (97.8% completion rate, or a 2.2% non-response rate). On a householder basis, 9,497 out of 9,719 householders have race/ethnicity coded (also 97.8% response rate). A “hot deck” imputation model was run for the five race/ethnicity categories, based on the SAS macro produced by Ellis (5). The race/ethnicity hot deck imputation used household size (5), the PUMA-of-residence (55), and a random sort variable, as the imputation model.

Age is coded on 22,978 out of 24,030 sample Bay Area persons (95.6% completion rate, or a 4.4% non-response rate). A “hot deck” imputation model was produced on four sub-sets of the sample person files:

- 1) Students (172 missing age out of 4,959 total sample students);
- 2) Workers (603 missing age out of 12,529 total sample workers);
- 3) Non-Workers (277 missing age out of 6,013 total sample non-workers);
- 4) Other/Not-Classified (97 missing age out of 529 “other” persons).

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.

Hot deck imputation was also performed on other variables used in the validation of the sample/expansion weighting factors: sex (male, female); number of jobs per worker; number of hours worked per job; and driver’s license status.

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, tenure, and household population by age and sex.

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.2.1 summarizes county and region-level number of households from the various census / ACS databases. The decennial census counts 2,608,023 households in the nine-county Bay Area; the one-year (2011) ACS, 2,599,927 households; the three-year 2009-11 ACS, 2,582,449 households; and the five-year 2007-11 ACS, 2,577,480 households.

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 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 tables are more accurately described as “households by commuters in household.”)

There is a standard ACS table on households by vehicles available by tenure (American Factfinder, Table #B25044), but the PUMS database was used for all ACS-based raking levels. (The original concept was to use households by vehicles available by workers in household by tenure, but the scarce nature of certain markets, e.g., multi-worker households with zero vehicles, required the splitting of the raking levels for workers and vehicles 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.

ACS data elements of interest include: workers in household, vehicles in household, student population, household population by age and school level, and workers by part-time / full-time employment status.

#### 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.

### **III. EXPLORING CENSUS AND SURVEY CHARACTERISTICS**

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 the Bay Area. 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 9,719 Bay Area households. This is out of the 2,608,023 households counted in Census 2010, for a “sampling rate” of 0.37% (9,719 / 2,608,023) (Table A.1.1). The simplest expansion model is to apply a weight of 268.3 to all households.

The sampling rate ranges from a low of 0.31% in San Francisco and Alameda Counties to 0.65% in Napa County (Table A.1.2). The number of sample households ranges from 317 in Napa County to 2,136 in Santa Clara 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 0.18 percent, comparing to 0.52 percent of owner-occupied households (Table A.1.3). The under-sampling of renter-occupied households is a continuing, and progressively worse problem in Bay Area household travel surveys. The following Table 3 shows how the renter household samples have changed in Bay Area households since 1980:

**Table 3**  
**Share, Renter-Occupied Households of Total Households: Survey vs. Census**

Year(s)	Census	Survey
1980/81	44.2%	49.4%
1990	43.6%	37.0%
2000	42.3%	30.6%
2010/12/13	43.8%	21.5%

(See Table A.3)

The census shows a steady share of renter households at 42 to 44 percent of all Bay Area households, 1980 to 2010. The 1981 survey actually over-sampled renter households relative to the 1980 Census (49.4 percent versus 44.2%). The survey share of renter households has gotten progressively worse, now at 21.5% survey versus 43.8% census in the latest generation of travel surveys.

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 0.21 percent of five-or-more person households, to a high of 0.47 percent of two person households (Table A.1.4). One-person households have the second lowest sampling rate at 0.32 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 the Bay Area. The sampling rate ranges from a low of 0.13 percent of the youngest householders (age 15 to 34); and a high of 0.65 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 53 percent of the Census 2010 population; and 74 percent of the CHTS 2012/13 sample (Table A.1.6). This is an obvious oversampling which is correctable using appropriate sampling weights. The smallest group is the “other” category (other race, American Indian / Alaskan Native, two-or-more race) which is 2.9 percent of households in Census 2010, compared to 3.0 percent of the survey households.

The sampling rate, by race/ethnicity of householder ranges from a low of 0.17 percent of Asian and Black householders, to 0.52 percent of white, non-Hispanic householders.

Sampling rates were also analyzed by population density level, using the Census 2010 gross population density (persons per square mile), for the following six density categories:

- 1) Rural (less than 500 persons per square mile);
- 2) Rural-Suburban (500 to 1,000 persons per square mile);
- 3) Disperse Suburb (1,000 to 6,000 persons per square mile);
- 4) Dense Suburb (6,000 to 10,000 persons per square mile);
- 5) Urban (10,000 to 20,000 persons per square mile); and
- 6) Urban Core (20,000 or more persons per square mile).

Sampling rate decreases with increasing density. The sampling rate ranges from a low of 0.29 percent in the urban core neighborhoods in the Bay Area to a high of 0.49 percent in the rural neighborhoods (Table A.1.7).

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, and by number of vehicles available in the household. (Note the slightly lower regional total households, 2,577,480, based on the ACS compared to Census 2010.)

The survey tends to under-sample households with zero worker households (0.29 percent), and three-or-more worker households (0.29%), and over-sampling the two-worker households (0.43 percent) (Table A.1.8).

Zero vehicle households (0.26 percent) are the most under-sampled compared to 0.46 percent of two-vehicle households (0.46 percent) (Table A.1.9).

Zero vehicle households are 9.6 percent of Bay Area households according to the 2007/11 ACS. This compares with 6.5 percent of the CHTS 12/13 sample households. This again is an important, but correctable bias.

Sampling rates for the 55 Bay Area PUMAs (Public Use Microdata Areas based on Census 2010) are reported in Table A.10. Sampling rates range from a low of 0.20 percent in the Newark/Union City/West Fremont PUMA, and the Hayward PUMA; to a high of 0.65 percent in the Napa County PUMA.

Other PUMAs with low sampling rates include the East Valley of San Jose (0.21 percent); and BayView/Hunters Point (0.22 percent). Other PUMAs with high sampling rates include Menlo Park/East Palo Alto (0.59 percent); Palo Alto/Mountain View (0.55 percent); Redwood City/San Carlos (0.54 percent) Walnut Creek/Lamorinda (0.53 percent); Sebastopol/Healdsburg/Sonoma (0.53 percent); and Berkeley/Albany (0.52 percent).

The first set of tables in Appendix A show the “one-dimensional” characteristics of census and survey data (Tables A.1 through A.3). The next of tables show the “two-

dimensional” characteristics, cross-classifying variable by geography and other socio-economic strata.

Households by county-of-residence by household size, census and survey, is summarized in Table A.4. The “survey weights” (the inverse of the “sampling rate”) ranges from a low of 119.4 weight for two-person households in Napa County, to a high of 822.4 for five-or-more person households in San Francisco County. The county patterns tend to mimic the regional pattern: under-sampling of the very small and very large households; over-sampling the two through four-person households.

Households by county-of-residence by tenure are shown in Table A.5. The sample weights range from a low of 119.1 for owner-occupied Napa County households to 646.0 for renter-occupied Alameda County households. Alameda County also has the highest weight for owner-occupied households by county (222.8).

Households by county-of-residence by age of householder are summarized in Table A.6. The sampling weights range from a low of 86.4 for Napa County householders age 55 to 64, to a high of 980.6 for San Francisco householders age 15 to 34. Another very high value is the 874.1 weight for Alameda County householders age 15 to 34.

Households by county-of-residence by race/ethnicity of householder are shown in Table A.7. Weights range from a lower of 128.4 for white, non-Hispanic households in Napa County to 1,256.5 for Asian/non-Hispanic households, also in Napa County.

Households by county-of-residence by number of workers in the household (ACS compared to survey) are shown in Table A.8. The weights range from a low of 138.6 for two-worker households in Napa County; to a high of 623.5 for three-or-more worker households in San Francisco.

Households by county-of-residence by number of vehicles in the household (again, ACS compared to survey) are shown in Table A.9. The weights range from a low of 125.0 for two vehicle households in Napa County; to a high of 511.8 for zero-vehicle households in Santa Clara County.

Two-way cross-classifications by PUMA-of-residence, by socio-economic variables, are provided in Tables A.10 through A.13 (household size, tenure, age of householder, race/ethnicity of householder). There is a pattern of no sample households for various minority householder categories in several of the PUMAs (Table A.13). This suggests the need to “collapse” certain categories if PUMA by race/ethnicity of householder is used as a weighting/raking level.

Cross-classifications by socio-economic variables (regional level, no geography breakout) is provided in Tables A.14 through A.16

Regional households by household size by tenure are shown in Table A.14. The sample weights from a low of 152.8 for two-person owner-occupied households to 763.2 for five-or-more person renter-occupied households.

Regional households by age of householder by tenure are shown in Table A.15. Weights range from 128.5 for owner-occupied households, age 55 to 64 householders; to a high of 1,139.9 for renter-occupied household, age 15 to 34 householders.

Regional households by race/ethnicity of householders are shown in Table A.16. Weights range from a low of 147.7 for owner-occupied, white/non-Hispanic households; to a high of 1,143.9 for renter-occupied, Asian/non-Hispanic householders. Other large weights include 807.9 for renter-occupied, black/non-Hispanic holds; and 585.5 for renter-occupied, Latino/Hispanic households.

The next set of tables (Tables A.17 through A.25) are “three-way” cross-classification of census and survey data, by county or PUMA-of-residence; by tenure; and by the other key socio-economic variables.

Several of these “three-way” cross-classifications provide the basis for the “marginal control totals” used in the proposed raking models.

Table 4 summarizes the raking level, and corresponding appendix table, used in developing the various raking/weight models.

**Table 4**  
**Raking Levels and Corresponding Appendix Tables**

Table	Raking Level Description
A.19	County (9) by Tenure (2) by Race/Ethnicity (5)
A.24	PUMA (55) by Tenure (2) by Minority Status (2)
A.20	County (9) by Tenure (2) by Workers in Household (4)
A.21	County (9) by Tenure (2) by Vehicles in Household (4)
A.23	PUMA (55) by Tenure (2) by Age of Householder (5)
A.27	County (9) by Number of Persons Age 20 to 29 (3)
A.22	PUMA (55) by Tenure (2) by Household Size (5)

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 by race/ethnicity, is based on Census 2010-based marginal control totals (Table A.19). The only problem is there are no sample renter-

occupied Asian/Pacific Islander households in Napa County. In this instance, the marginal control total was “collapsed” to include both Napa County renter and owner-occupied households for Asian/Pacific Islander householders.

The original concept was to have a raking level of households by PUMA-of-residence by tenure by race/ethnicity (5) of householder. Given the sparseness of the sample, and the desire to include a PUMA-level, race/ethnicity raking level, MTC added a PUMA-of-Residence (55) by tenure (2) by minority status (2), raking level to complement the county-level level (Table A.24). The “minority status” is either 1) white, non-Hispanic; or 2) all other groups, combined.

The third raking level is county by tenure by number of workers in the household (0, 1, 2, 3+) (Table A.20). The marginal control totals are derived from the American Community Survey 2007/11 PUMS.

The fourth raking level is county by tenure by number of vehicles available in the household (0, 1, 2, 3+) (Table A.21). Again, the marginal control totals are derived from the American Community Survey 2007/11 PUMS. The original concept was to jointly rake the distribution of households by county by tenure by workers in household by vehicles in household. This was problematic in that the ACS data shows very few households in certain categories (e.g., multi-worker households with zero vehicles, and multi-vehicle households with zero workers.) An alternative strategy may be to use the ACS standard tabulations and create a raking level by PUMA-of-Residence (55) by Tenure (2) by vehicles available in the household (4).

The fifth raking level is PUMA-of-residence by tenure by age of householder (Table A.23). There are ten instances (out of 550 possible marginal control totals) where there are zero sample households. In these instances, the marginal control totals were re-defined by combining the owner-occupied and renter-occupied number of households, within the problematic PUMA, age of householder combinations.

The sixth raking level is county of residence by the number of persons age 20 to 29 in the household (0, 1, 2+) (Table A.27). This raking level was not included in Raking Model #1, but was introduced in Raking Model #2 to correct some of the underweighting of persons age 20 to 29. The PUMA/tenure/age of householder is very important in improving the age balance of the survey, and this sixth raking level complements and improves on the estimates of 20 to 29 year old persons, who are probably a very mobile population group.

The seventh and last raking level is PUMA-of-residence (55) by tenure (2) by household size (1, 2, 3, 4, 5+) (Table A.22). The marginal control totals are from the Census 2010 “short form.” There are twelve instances (out of 550 possible marginal control totals) where there are zero sample households. As was the case with other raking levels, the

owner-occupied and renter-occupied marginal control totals were collapsed, or combined, for the problematic cells.

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).

Additional tables in Appendix A show the regional age-sex distribution in both tables and figures. If the sampling weights were produced at a person level, the weights would range from a low of 143.8 for males age 60 to 69; to a high of 635.6 for females age 20 to 29 (Table A.26). Other high sampling weights are males age 20 to 29 (620.1) and males age 30 to 39 (549.4). Other low sampling weights are for females age 60 to 69 (153.8), females age 50 to 59 (178.5), and males age 50 to 59 (189.3).

Classic "age-sex pyramids" are illustrated in Figures A.1 and A.2. Figure A.1 shows data for the Bay Area based on Census 2010, showing a fairly regular "bottle-shaped" pattern associated with a large baby boomer population. Figure A.2 shows the data from the CHTS 2012/13 for the Bay Area: an "hourglass" shape, with very thin shares for persons age 20 to 34. Perhaps the persons age 20 to 34 are a "high mobility" group, and are difficult to contact and retain in a household travel survey effort. This is an obvious but correctable, to a certain extent, bias.

#### **IV. VALIDATION AND EVALUATION OF WEIGHTING METHODS**

Six weighting methods are examined in this section. Detailed data tables are included as Appendices B, C, D and E.

The study consultant developed a set of weights, which are denoted as “Model #0” weights in this report. MTC staff developed five sets of raking models/weights, denoted as Model #1, Model #1c, Model #2, Model #2c1, and Model #2c2. The “c” stands for “constrained”.

Model #0 weighting procedures are documented in the study consultant report (6). It is a five-level raking scheme, conducted at the statewide level for all Bay Area and non-Bay Area 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)

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.” (6) (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 weights were produced by MTC staff adapting SAS “raking” macros produced by Izrael, Hoagland and Battaglia (7).

Model #1 includes six raking levels:

- 1) County (9) by Tenure (2) by Race/Ethnicity of Householder (5);
- 2) PUMA (55) by Tenure (2) by Minority Status of Householder (2);
- 3) County (9) by Tenure (2) by Workers in Household (4);
- 4) County (9) by Tenure (2) by Vehicles in Household (4);
- 5) PUMA (55) by Tenure (2) by Age of Householder (5); and
- 6) PUMA (55) by Tenure (2) by Household Size (5).

Six sets of marginal control total files are required for this six 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 three categories used in any given level, for example:

$$\text{County} * 100 + \text{Tenure} * 10 + \text{Race/Ethnicity} * 1.$$

An index of “111” in this example is Alameda County households (county=1), owner-occupied (tenure=1), white/not-Hispanic (race/ethnicity=1). An index of “9725” in this example is Sonoma County households (county=97), renter-occupied (tenure=2), Hispanic/Latino (race/ethnicity=5). 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 (66.0) and ceiling (2500.0) to Model #1 weights. This is denoted as Model #1c (or Model #1, constrained).

The validation and cross-validation of weighting Model #0, Model #1, and Model #1c then started, and is summarized in Appendices B (household-level) and C (person-level) tables.

Results of Model #1c appeared very promising, until MTC staff performed cross-validation tests based on the age and sex of the survey respondents relative to Census 2010. Model #1c, at the regional level, under-estimates males age 20 to 29 by 32 percent; and females age 20 to 29 by 25 percent (Table C.4). The original concept was that the “age of householder” raking level would produce the best estimates of population by age; the results suggested a novel approach: adding a raking level adjusting to the number of “twentysomethings” in the household.

The distribution of households by the number of persons age 20 to 29 is hardly a standard census tabulation, so the 2007/11 ACS PUMS was used to produce marginal control total of households by county-of-residence (9) by persons age 20-29 in household (0, 1, 2+).

Model #2 simply adds a “sixth” raking level to Model #1. The seven raking levels in Model #2 are:

- 1) County (9) by Tenure (2) by Race/Ethnicity of Householder (5);
- 2) PUMA (55) by Tenure (2) by Minority Status of Householder (2);
- 3) County (9) by Tenure (2) by Workers in Household (4);
- 4) County (9) by Tenure (2) by Vehicles in Household (4);
- 5) PUMA (55) by Tenure (2) by Age of Householder (5);
- 6) County (9) by Number of Persons Age 20-29 in Household (3); and
- 7) PUMA (55) by Tenure (2) by Household Size (5).

Similar to Model #1, Model #2 weights were constrained using floors and ceilings. Two options were produced:

- 1) Model #2c1, constraining the weights to 66.0 floor and 2500.0 ceiling; and
- 2) Model #2c2, constraining the weights to 49.0 floor and 3000.0 ceiling.

**Table 5**  
**Range of Weights by Weighting Model**

	Median	Mean	Minimum	Maximum
Model #0	254.69	265.22	8.42	1,236.39
Model #1	147.30	268.34	7.73E-08	4,772.94
Model #1c	147.30	268.36	66.00	2,500.00
Model #2	144.09	268.34	7.38E-08	4,772.94
Model #2c1	144.09	268.41	67.00	2,500.00
Model #2c2	144.09	268.38	49.00	3,000.00

**Table 6**  
**Percentile Distribution of Weights by Weighting Models**

	1st Percentile	5th Percentile	95th Percentile	99th Percentile
Model #0	16.08	84.87	481.14	586.75
Model #1	3.69	41.22	897.88	1,982.07
Model #1c	66.00	66.00	897.88	1,982.07
Model #2	3.16	38.59	914.95	1,998.85
Model #2c1	67.00	67.00	914.95	1,998.85
Model #2c2	49.00	49.00	914.95	1,998.85

What is interesting is the range of weights in simpler models versus more complex raking models. Model #0 has only 76 total “marginal controls” in the five layer raking scheme, at the statewide level. Model #1 has 1,554 “marginal control totals” in a six layer raking scheme; Model #2 has 1,581 “marginal control totals” in a seven layer raking scheme.

It is apparent that simpler raking schemes have less extreme weights than more complex raking schemes. On the other hand, the validation of these raking models, in terms of comparing the goodness of fit, model versus observed, and is superior in the complex models versus the simpler model.

Appendix B summarizes the household level validation of Models #0, #1 and #1c, at the county and regional level. Regional level results are provided in Table B.1.1 through B.1.6. Model #1 perfectly matches the Census 2010 count of households by household size, given that the last raking level is household size by tenure (Table B.1.1.) Model #1 tenure is slightly different than Census 2010 control totals due to the need for collapsing categories at several PUMA/tenure/household size combinations. Comparison for the other household categories (householder age, householder race/ethnicity, workers/household, vehicles/household) show reasonable matches, Model #1 relative to Census 2010 or ACS data.

Appendix C summarizes the person level validation of Models #0, #1 and #1c, at the county and regional level. Household population by race/ethnicity is not easily derivable from Census 2010 standard tabulations (Table C.1). (Data is published for household population white, not-Hispanic; and Hispanic/Latino, but not for the other three minority groups.) The regional age-sex distribution for Model #1c (Table C.4) suggested the need for additional household controls on population by age. Household population by the eight “person types” used in the current travel model system are shown in Tables C.5 and C.6.1-C.6.9. Both Models #1 and #1c show a very reasonable estimate of workers, non-workers, and students, relative to Census/ACS observed totals.

The last set of tables in Appendix C review the distribution of student population by age and by school level. Model #1c does a good job in estimating high school (+5.4 percent) and undergraduate college (+2.1%) enrollment; but over-estimates grade school (K-8) (+15.0 percent); and under-estimates graduate school enrollment (-31.3 percent) (Table C.7.3). The age-sex pyramid and the graduate school enrollment tend to reinforce the need to add a household age based raking level.

Appendix D summarizes the household level validation of Models #2, #2c1, and #2c2, at the county and regional level. The Appendix D tables follow the same numbering sequence at the Appendix B tables. What is interesting to note is that Model #2c2 (49/3000 floor/ceiling) does a slightly better job in matching observed census data than Model #2c1 (67/2500 floor/ceiling). The results at the regional and county level are reasonable for all three weighting/raking schemes.

Appendix E summarizes the person level validation of Models #2, #2c1, and #2c2, at the county and regional level. These models reduce the under-estimates of the 20 to 29 year old population relative to the Model #1 series (Table E.2 through E.4). The 20 to 29 year old population is 12 percent low, in Model #2c2, compared to 29 percent low in Model #1c. Graduate student population is still too low by 24.6 percent in Model #2c2, compared to 31.3 percent in Model #1c.

The last set of tables in Appendix E, Tables E.16 through E.20, examine average (mean) characteristics by market segment: average household size, average workers per household, average students per household, and the average age of persons in household. Data is compared to American Community Survey 2007/11, from the ACS PUMS records. Data is summarized for the ACS and the six raking models. These tables are interesting to show how these regional mean characteristics vary by the different household classifications as used in this study.

Appendix F 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 the 1990 and 2000 Bay Area Travel Surveys. This is a fix since the average size of 5+ persons in the Census (5.979 persons/5+ person household) is slightly higher than the average size in the weighted survey (5.515 persons/5+ person household). These correction factors are calculated at the PUMA-of-residence, and may range from 0.86 to 1.26. 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. Correction factors for Models #1 and #1c are documented in Appendix Table F.1 through F.4; and for Models #2, #2c1, and #2c2 in Appendix Tables F.5 through F.8.

## **V. NEXT STEPS**

The recommendation is that Model #2c2 weights are the final MTC weights on the Bay Area households in the CHTS 2012/13 database. This is for the “combined” sample (N=9,719 sample households) which includes travel for both weekdays and weekend days.

The next step in this analysis is to prepare separate weights for the “weekday sample households” (N=8,086); the “Saturday sample households” (N=717); and the “Sunday sample households” (N=916). It is envisioned that the “combined sample weights” will be used when estimating auto ownership level models, and other analyses focusing on the demographics of the household. For purposes of estimating aggregate “average weekday” travel characteristics, only data from the “weekday sample households” would be used.

Future steps in the analysis of the CHTS 2012/13 travel survey for the Bay Area 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 MTC 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.

New weighting/raking methods have also been developed for the entire statewide CHTS 2012/13 databases. Technical reports documenting these methods will be produced. In addition, 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.

Further research on raking methods will be undertaken as time permits. Options may include simplifying some of the three-dimensional raking schemes (e.g., omitting in tenure in the PUMA by tenure by household size) to analyze the impacts on extreme weights, and raking model closure. The macro may also allow for specifying “floors” and “ceilings” for weights, which might prove useful in future efforts.

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

## **VI. REFERENCES**

1. Working Paper #4: 1981 MTC Travel Survey: Sample Weighting. Metropolitan Transportation Commission, Berkeley, CA, June 1982.  
(<http://dataportal.mtc.ca.gov/working-paper-4--1981-mtc-travel-survey-sample-weighting.aspx> )
2. Peter R. Stopher and Cheryl Stecher. "Blow Up: Expanding a Complex Random Sample Travel Survey" in Transportation Research Record 1412, TRB, 1993, pp. 10-16.
3. Sample Weighting and Expansion: Working Paper 2: 1990 MTC Travel Survey. Metropolitan Transportation Commission, Oakland, CA, Revised June 1993.
4. Sample Weighting and Expansion: Working Paper #1: Bay Area Travel Survey 2000 (BATS2000). Metropolitan Transportation Commission, Oakland, CA, June 2003.  
(<http://dataportal.mtc.ca.gov/working-paper-1--bay-area-travel-survey-2000-sample-weighting-and-expansion.aspx>)
5. 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>)
6. "2010-12 California Household Travel Survey: Final Report" Versions 1.0, NuStats Research Solutions, Austin, Texas, June 14, 2013.
7. 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](http://www.abtassociates.com/PDFS/258_25.aspx))