STATE OF CALIFORNIA • DEPARTMENT OF TRANSPORTATION

TECHNICAL REPORT DOCUMENTATION PAGE

TR0003 (REV 10/98)

ADA Notice

For individuals with sensory disabilities, this document is available in alternate formats. For information call (916) 654-6410 or TDD (916) 654-3880 or write Records and Forms Management, 1120 N Street, MS-89, Sacramento, CA 95814.

1. REPORT NUMBER	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER
CA17-2963	n/a	n/a

4. TITLE AND SUBTITLE

Long Distance Travel in the California Household Travel Survey (CHTS) and Social Media Augmentation

5. REPORT DATE

05/15/2017

6. PERFORMING ORGANIZATION CODE

n/a

7. AUTHOR

Konstadinos Goulias

8. PERFORMING ORGANIZATION REPORT NO.

n/a

9. PERFORMING ORGANIZATION NAME AND ADDRESS

University of California Center on Economic Competitiveness in Transportation (UCCONNECT)
University of California Santa Barbara
Campus Address: 1832 Ellison Hall, UCSB, CA, 93106-4060
Campus Phone: 805-284-1597

10. WORK UNIT NUMBER

n/a

11. CONTRACT OR GRANT NUMBER

65A0529 TO 047

12. SPONSORING AGENCY AND ADDRESS

California Department of Transportation
Division of Research, Innovation and System Information
1727 30th Street, Sacramento, CA 95816

13. TYPE OF REPORT AND PERIOD COVERED

May 1, 2016 through April 30, 2017

14. SPONSORING AGENCY CODE

n/a

15. SUPPLEMENTARY NOTES

n/a

16. ABSTRACT

Major emphasis has been placed in recent years on intra-urban and intra-city travel within transportation research. However, this has left a large gap in conventional understanding of interregional and long distance travel. The primary goal of this research is to provide empirical evidence for conducting long distance travel behavior analysis using synthetic population methods; and identifying the determinants for long distance travel behavior. This study used the California Household Travel Survey (CHTS) data to improve the attractiveness indicators used by the long distance component of the statewide model. Our discoveries through this important research are as follows: 1) Travel differences among persons and households are mainly due to social and demographic characteristics of households with primary driver the household wealth and employment. Place of residence plays a major role in explaining long distance travel and this shows a more detailed analysis of opportunities for activities around the place of residence would inform long distract vehicle miles traveled contribution in a substantial way. The clear recommendation from this analysis is to design activity diaries that span multiple-days of complete households and a satellite survey that has diaries for an 8-week travel log that has added information about travel during the 8-week period and the people with whom travel happens. For Caltrans, this project intends to provide the following benefits/outcomes: 1) Baseline inventory for statewide travel demand; 2) Analysis of vulnerable segments of the population; and 3) Estimate of long distance travel, particularly as a constituent to statewide vehicle miles traveled.

17. KEY WORDS	18. DISTRIBUTION STATEMENT
Vehicles Miles Traveled, VMT, California Household Travel Survey, CHTS	This is a public university report. No restrictions.

19. SECURITY CLASSIFICATION *(of this report)*	20. NUMBER OF PAGES	21. COST OF REPORT CHARGED
Unclassified	97	

DISCLAIMER STATEMENT

This document is disseminated in the interest of information exchange. The contents of this report reflect the views of the authors who are responsible for the facts and accuracy of the data presented herein. The contents do not necessarily reflect the official views or policies of the State of California or the Federal Highway Administration. This publication does not constitute a standard, specification or regulation. This report does not constitute an endorsement by the Department of any product described herein.

For individuals with sensory disabilities, this document is available in alternate formats. For information, call (916) 654-8899, TTY 711, or write to California Department of Transportation, Division of Research, Innovation and System Information, MS-83, P.O. Box 942873, Sacramento, CA 94273-0001.

Long Distance Travel in CHTS and Social Media Augmentation

Konstadinos G. Goulias,
Adam W. Davis,
Elizabeth C. McBride
Krzysztof Janowicz
Rui Zhou

Final Report
Contract Number: 64A0529
Task Order: 047

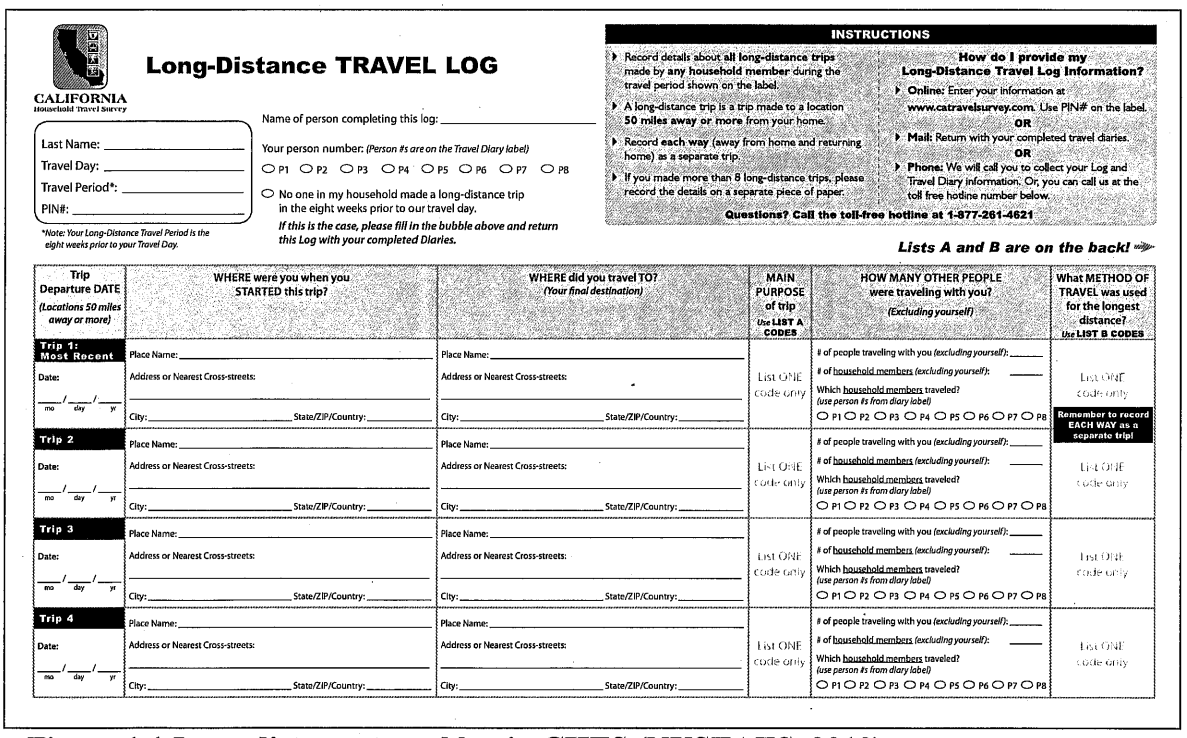
Submitted to CALTRANS&UCCONNECT

May 15, 2017
GeoTrans & STKO Laboratories, UCSB
Santa Barbara, CA

Table of Contents

1. Introduction

The objective of this report is to first review what we know from the literature about long distance travelers, analyze the contents of the long distance travel log of the California Household Travel Survey (CHTS), demonstrate the augmentation of the trip/tour records with destination attractiveness indicators, derive prototypical traveler profiles, and provide amore detailed analysis of long distance tours. The data are from a simplified travel log that asked respondents from households to report all the trips 50 miles or longer they made in the 8-weeks preceding the day they were assigned a full travel diary. The survey instrument used for this reporting is shown in Figure 1. In this report we identify a few issues with the data collected using this travel log, and these issues motivate us to also investigate the long distance travel reported in the daily diary. The range of variables that we can analyze depends heavily on the accuracy with which respondents reported their trips, and we found they were generally more accurate in the daily diary. However, the long distance travel log contains data that span longer periods than 24 hours and therefore a better record of trips with overnight stays away from home.



Long-Distance TRAVEL LOG

CALIFORNIA Household Travel Survey

Last Name: _____
Travel Day: _____
Travel Period*: _____
PIN#: _____

*Note: Your Long-Distance Travel Period is the eight weeks prior to your Travel Day.

Name of person completing this log: _____

Your person number: *(Person #s are on the Travel Diary label)*

○ P1 ○ P2 ○ P3 ○ P4 ○ P5 ○ P6 ○ P7 ○ P8

○ No one in my household made a long-distance trip in the eight weeks prior to our travel day.
If this is the case, please fill in the bubble above and return this Log with your completed Diaries.

INSTRUCTIONS

- Record details about **all long-distance trips** made by **any household member** during the travel period shown on the label.
- A long-distance trip is a trip made to a location **50 miles away or more** from your home.
- Record **each way** (away from home and returning home) as a separate trip.
- If you made more than 8 long-distance trips, please record the details on a separate piece of paper.

How do I provide my Long-Distance Travel Log Information?

- **Online:** Enter your information at **www.catravelsurvey.com**. Use PIN# on the label.
OR
- **Mail:** Return with your completed travel diaries.
OR
- **Phone:** We will call you to collect your Log and Travel Diary information. Or, you can call us at the toll free hotline number below.

Questions? Call the toll-free hotline at 1-877-261-4621

Lists A and B are on the back!

Trip Departure DATE *(Locations 50 miles away or more)*	WHERE were you when you STARTED this trip?	WHERE did you travel TO? *(Your final destination)*	MAIN PURPOSE of trip *Use* LIST A CODES	HOW MANY OTHER PEOPLE were traveling with you? *(Excluding yourself)*	What METHOD OF TRAVEL was used for the longest distance? *Use* LIST B CODES
Trip 1: Most Recent Date: ___/___/___ mo day yr	Place Name: _____ Address or Nearest Cross-streets: _____ City: _____ State/ZIP/Country: _____	Place Name: _____ Address or Nearest Cross-streets: _____ City: _____ State/ZIP/Country: _____	List ONE code only	# of people traveling with you *(excluding yourself)*: _____ # of household members *(excluding yourself)*: _____ Which household members traveled? *(use person #s from diary label)* ○ P1 ○ P2 ○ P3 ○ P4 ○ P5 ○ P6 ○ P7 ○ P8	List ONE code only **Remember to record EACH WAY as a separate trip!**
Trip 2 Date: ___/___/___ mo day yr	Place Name: _____ Address or Nearest Cross-streets: _____ City: _____ State/ZIP/Country: _____	Place Name: _____ Address or Nearest Cross-streets: _____ City: _____ State/ZIP/Country: _____	List ONE code only	# of people traveling with you *(excluding yourself)*: _____ # of household members *(excluding yourself)*: _____ Which household members traveled? *(use person #s from diary label)* ○ P1 ○ P2 ○ P3 ○ P4 ○ P5 ○ P6 ○ P7 ○ P8	List ONE code only
Trip 3 Date: ___/___/___ mo day yr	Place Name: _____ Address or Nearest Cross-streets: _____ City: _____ State/ZIP/Country: _____	Place Name: _____ Address or Nearest Cross-streets: _____ City: _____ State/ZIP/Country: _____	List ONE code only	# of people traveling with you *(excluding yourself)*: _____ # of household members *(excluding yourself)*: _____ Which household members traveled? *(use person #s from diary label)* ○ P1 ○ P2 ○ P3 ○ P4 ○ P5 ○ P6 ○ P7 ○ P8	List ONE code only
Trip 4 Date: ___/___/___ mo day yr	Place Name: _____ Address or Nearest Cross-streets: _____ City: _____ State/ZIP/Country: _____	Place Name: _____ Address or Nearest Cross-streets: _____ City: _____ State/ZIP/Country: _____	List ONE code only	# of people traveling with you *(excluding yourself)*: _____ # of household members *(excluding yourself)*: _____ Which household members traveled? *(use person #s from diary label)* ○ P1 ○ P2 ○ P3 ○ P4 ○ P5 ○ P6 ○ P7 ○ P8	List ONE code only

Figure 1.1 Long distance travel log in CHTS (NUSTATS, 2013)

Past studies of long distance travel have found that commuting by people who sought out lower cost housing is a major contributor to long distance travel, and that higher income and employed persons travel more, but there are multiple shortcomings in the literature that we seek to address here. The literature contains a variety of definitions for "long distance" travel, including ones based on distance (e.g., longer than 50 miles, 100 miles, or longer than 100 kilometers) and travel time (e.g., 40 minutes). Long distance travel researchers have considered a variety of indicators including number of long distance trips, activity before and/or after commute, mode used, time of day of trip, and destination (Georggi and Pendyala, 2000, Axhausen, 2001, Beckman and Goulias, 2008, LaMondia and Bhat, 2011, Caltrans, 2015, Shahrin et al., 2014, Holz-Rau et al., 2014). Most studies did not address trip chaining (e.g., people going to a work place, then to a leisure destination, and then back home). Very little analysis is also found in differentiating trips with an overnight stay, despite the important differences between these trips and daily commuting. The choice of analysis in past studies was presumably driven by: a) an emphasis in the literature on trips to and from work; and b) a focus on a single trip by an individual person as the unit of analysis instead of multiple trips from household members.

This focus on commute trips is also reflected in the multitude of person factors used to explain variation in travel behavior in past research (Table 1.1). Table 1.1 also shows household and location characteristics that have been considered as determinants of long distance travel behavior. It is also important to note that a few researchers (de Abreu et al., 2006, 2012) consider long distance travel, car ownership, and residential and job location (and the distance between the two) as a system of joint decisions that are best explained using methods that can disentangle the complex relationships among all these behavioral facets. From this viewpoint, long distance travel (particularly for commuters) cannot be separated from the choice of work and home location and should be modeled jointly.

Table 1.1 A selection of variables used to explain long distance travel in past studies

Person	Household	Location
Age	Household size	Destination region
Gender	Household Income	Characteristics at destination model
Education	Type of household (single, couple etc)	Opportunities for activities
Occupation	Home ownership type	Leisure employment
Employment	Presence and number of children	Living area density and diversity
Ethnicity	Number of persons with driver's license	Destination area density and diversity
Income	HH Annual Income	Accessibility of origin and destination
Vehicle Ownership	HH Car Ownership	Distance to CBD
Life Cycle Stage	Household structure	Jobs-housing balance
Availability of company car	Childcare	Availability of different modes and their characteristics at home location
Length of employment	Child-related travel	Availability of different modes and their characteristics at work location
Foreign birth		Property value
Attitudes about attachment to activities		

The review in Mitra (2016) is particularly useful in mapping recent literature on long distance travel and its determinants. His findings are exactly what one would expect: age, gender, education, employment and occupation, car ownership, household structure, place of residence and workplace as well as housing cost and accessibility influence long distance travel in a variety of ways. His analysis also shows that developing traveler profiles at the level of a household (rather than the individual) is a better choice to understand how and why long distance travel happens, and our analysis follows this lead.

In another analysis of CHTS, Bierce and Kurth (2014) identified an issue of underreporting of repetitive trips in the 8-week long distance data. In essence, long distance commuters did not report all their commuting trips. We find that this underreporting may also exist for longer trips, though less severely than it does for shorter ones. Identifying the correct mix of distances and overall volume of travel is particularly important when one seeks to estimate the contribution of VMT from long distance travel to California estimates of VMT (see also Chapman, 2007).

2. Long Distance Travel Taxonomy

Figures 2.1 and 2.2 provide the context and framework of our analysis of long distance travel in California. We use a similar taxonomy to the one used by the DATELINE project in Europe.

Travel patterns of the long-distance log are classified by the presence or absence of an overnight stay. This allows us to identify travel that requires added transactions such as hotel reservations, meal planning, modes other than personal vehicles, and rearrangement of time allocation within a household. Travel with an overnight stay is further classified by single versus multiple overnight stays. We also classify travel by destination location, since this affects the sorts of outside data we can use to augment the travel record: a) all trips within California; b) all trips within the United States (but at least one trip out of California); and c) at least one trip outside the United States. Other indicators that aid in exploring the data include: weekday versus weekend travel and the degree of complete information available. Unlike other studies we do not discard any trip records (e.g., missing income records) in this analysis to avoid introducing more biases that are currently present in the CHTS data. The geographic distribution of the long-distance trips found in the 8-week log and daily diary are displayed in the maps of Appendices C, D, E. and F. The next section provides an overview of the analysis of trips and tours to understand the data we have and to guide subsequent work.

Figure 2.1 One-Day Journey

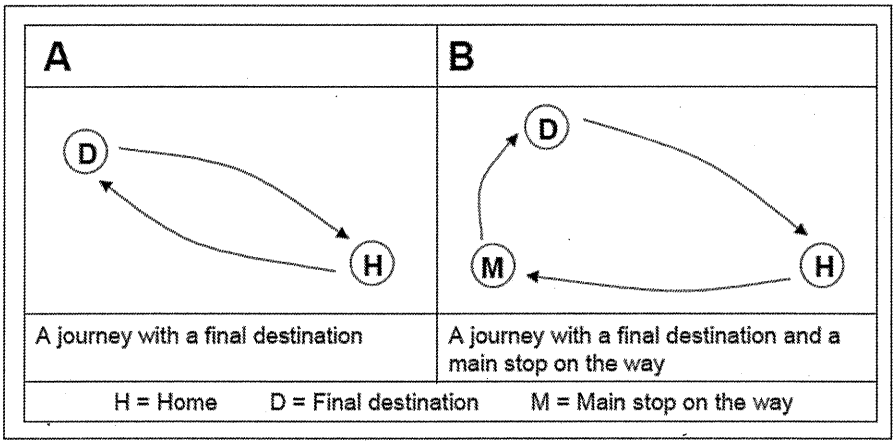
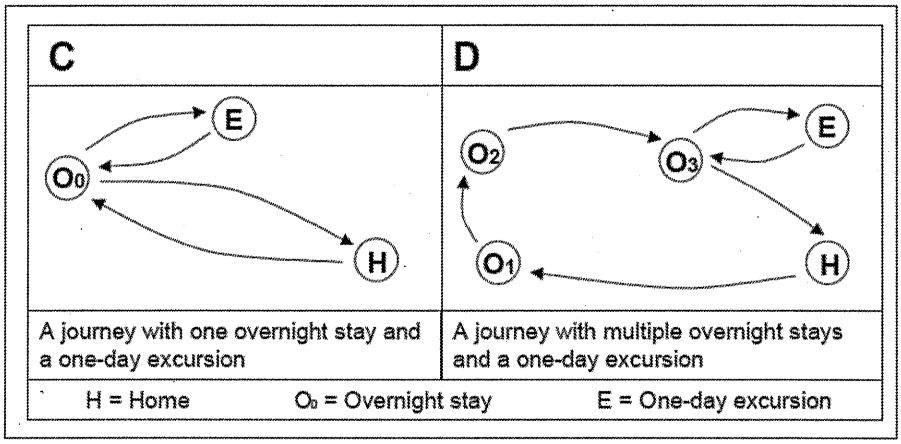


Figure 2.2 Multi-Day Journey



Figures 2.1 & 2.2 Taxonomy of long distance travel (source DATELINE project, Broeg et al., 2003)

3. Trip-based and Tour-based Analysis and Data Completeness

This section reviews the data available in the long distance travel log with specific focus on data quality, data extraction techniques, and some basic findings from the dataset. The eight-week long distance travel log in the CHTS provides valuable data but because it was measured separately from the main travel diary, this dataset is less complete and consistent. The limitations and biases in this data necessitate additional quality checking. This project presents us with two key questions about data inclusion:

- Which records should be included for general analysis of the relationship between long distance travel and household/person/regional characteristics?
- What threshold of completeness should be used to decide which households should form the seed of a synthetic population for long distance travel?

In this section we demonstrate that most records in the eight-week long distance travel log can be used in the analysis of household and long distance trip types, but that the daily diary should be used to estimate long distance VMT and PMT in the synthetic population, since it provides a more complete record of long distance travel per day.

Eight Week Log Completeness

The CHTS households dataset contains two indicators that pertain to long distance travel:

- "Do you have a completed Long Distance Log to refer to?"
- "We would like to gather a list of all long distance trips made by anyone in your household in the eight weeks prior to your travel day. Remember that a long distance trip is any one-way trip of more than 50 miles. Tell me about the first long distance trip."

Responses to these questions should align with each other and with the actual record of long distance trips, but for many households they do not, as shown in Table 3.1. Of the 42,431 households for which we have any usable data, 22,692 reported having a complete long distance log, 16,965 did not have a complete log, and 2,448 reported never having received a copy of the log. The Long Distance file contains 68,193 trip records from 18,008 households. Most long distance trips are reported by households that gave consistent answers to questions about long distance travel (86.6% of households are in

this category - cells shaded green). Unfortunately, thousands of households did not provide consistent information. 12.8% of households answered in ways that suggest a strong possibility of incomplete data (cells shaded yellow, e.g. they said they did not have a complete LD log but did report some trips). A small number of households gave severely inconsistent data (cells shaded orange, e.g., said they did not have a complete log and said they made no long distance trips, but nevertheless reported at least one LD trip) or responded either Don't Know or Refuse for one or more relevant questions (cells in red).

Table 3.1 Relationship between reported LD log completeness and reported LD trips

Do you have complete LD log?	Tell me about your first LD trip.	Households with an LD trip	Households without an LD trips	LD Trips	Trips / HH with Trip
1-Completed	1-Yes	14,667	1	59,174	4.03
1-Completed	2-No LD Trips	61	7,952	186	3.05
1-Completed	8/9-DK/RF	11	0	26	2.36
2-Not Completed	1-Yes	2,694	1	7,430	2.76
2-Not Completed	2-No LD Trips	110	14,136	329	2.99
2-Not Completed	8/9-DK/RF	23	1	28	1.22
3-Never Received	1-Yes	359	0	843	2.35
3-Never Received	2-No LD Trips	6	2,081	10	1.67
3-Never Received	8/9-DK/RF	2	0	3	1.50
8/9-DK/RF	1-Yes	62	0	140	2.26
8/9-DK/RF	2-No LD Trips	5	248	5	1.00
8/9-DK/RF	8/9-DK/RF	8	3	8	1.00

In order to avoid biasing the results by excluding large numbers of households from the analysis, it may be beneficial to keep most records (possibly those shaded green and yellow) for general analysis of the relationship between household characteristics and long distance travel. One way to diminish the effect that incomplete records would have on the results would be to base this analysis on indicator variables (e.g., does this household make ANY overnight trips?) instead of on counts (e.g., how many overnight trips did this household make?).

Because the population synthesis is intended to estimate statewide long distance travel totals, data used in it should meet a higher standard of completeness. Households with an incomplete LD log reported an average 1.3 fewer trips than did households with complete LD logs. This may indicate that they took fewer long distance trips, but it seems likely that this is at least partially a result of them excluding or forgetting some trips. Including these households may cause this analysis to underestimate total long distance travel. Additionally, using synthetic populations should smooth out any biases caused by excluding households with incomplete records if true long distance travel is related to variables included in the process of synthesizing the population.

Extracting Tour Characteristics

Among the households that provided a complete long distance travel log, there is considerable variability in the level of detail provided. Some long distance logs include a step-by-step record of every trip taken as part of a long distance tour (e.g. home → SFO, SFO → LHR, LHR → hotel in London, and then back again for six total trips); other logs include a single entry for each tour (e.g., one trip home → London). Though both of these records provide useful information, they cannot be treated in the same way. Individual unconnected trips can be assumed to be independent of each other, whereas individual legs on a single tour cannot. On the other hand, complete tour records provide more information, such as duration, links between modes, and the mix of trip purposes that cannot be extracted from a single trip. The criteria used to evaluate potential tours and the variables that can be extracted from each tour are summarized in Table 3.2.

The process of extracting tours from the LD dataset relies on the assumption that long distance tours start and end at home, since this appears to match the way most households filled out the log. To convert the record of trips to a set of tours classified by data quality, we follow these steps (for each "near" distance comparison, a threshold of 1km was used to account for any imprecision in location reporting and geocoding): (1) For each trip, calculate distance from home for origin and destination; (2) If a trip starts near the household's home location it is flagged as a *from home* trip; if a trip ends near home, it is flagged as a *to home* trip; (3) Each household's trips are sorted by date, and the

cumulative sums of trips *from* and *to home* is calculated for each trip in such a way that the cumulative sum of trips *to home* doesn't increment until the next trip; (4) Trips by a household that have matching cumulative total of trips *from* and *to home* are grouped into tours, and all trips without matches are declared single trips; (5) For all trips that are part of a multi-trip tour, distance from destination to the next origin is computed; tours are considered fully continuous if each destination matches the next origin in either name or location (again within 1km); (6) Tours are considered complete if they have full continuity and the first origin and last destination match the household's home location; all other tours are considered partially complete. Numbers of trips and tours detected by this process are shown in Table 3.3. Most trips group cleanly into complete tours, but many do not.

Table 3.2 Tour completeness criteria and usable variables.

Tour Completeness Category	Criteria	Potentially Usable Characteristics
Single Trip	• One trip • No continuity with other LD trips reported by this household	• Origin and destination locations and distance from home • Single mode used • Single trip purpose
Partial Tour	• Multiple trips in sequence • Trip destinations generally match next origin, but discontinuity is allowed • Start and/or end may not be at home	• Duration (floor only) • Multiple locations • Mix of modes • Mix of purposes
Complete Tour	• Multiple trips in sequence • Trip destinations always match next origin • Origin of first trip and destination of last trip are at home	• Duration (precise) • Multiple locations • Mix of modes • Mix of purposes

Table 3.3 Number of long distance trips by completeness and length.

	Single Trip	Partial Tour	Complete
Single Day	--	970	7,688
One Overnight	--	593	1,924
2-6 Overnights	--	1,545	5,207
7+ Overnights	--	1,979	2,122
Unknown Duration	18,298	172	678
Total Tours	18,298	5,259	17,619
Total Trips	18,298	12,690	37,194
Households with at least one such trip	9,513	3,560	9,540

Despite the limitations imposed by the quality of the eight-week long distance travel log data, it is easy to identify some trends. Households with higher income are much more likely to make long distance trips, as shown in Figure 3.1, and households in some parts of the State engage in long distance travel at much higher rates than in others. Tables 3.4a and 3.4b show the rate of long distance trip-making in the sample split by home county and trip length. This table shows that while there is considerable spatial variation, regional factors likely influence long distance travel. Survey respondents who live in relatively dense Southern California urban counties like Los Angeles and San Diego are less likely to make long distance trips, perhaps because they have access to a wide range of opportunities with a trip of less than 50 miles from home. Urban counties in Northern California make long distance trips at a somewhat higher rate than those in the south, and this difference is especially pronounced for long trips. Additionally, respondents drawn from Northern California's outer suburban counties (Napa, Sonoma, and Santa Cruz) are very likely to make long distance trips, perhaps since they can access a much wider range of employment and retail opportunities by traveling a little more than 50 miles to San Francisco. In part because of the small number of households drawn from each rural county, small rural counties have widely variable reported rates of long distance travel (an issue that synthetic population generation is designed to address), but appear to be relatively overrepresented at the top of the table. This is no surprise, since in rural areas, trips of over 50 miles are often necessary to access any opportunities not available in the immediate vicinity of home.

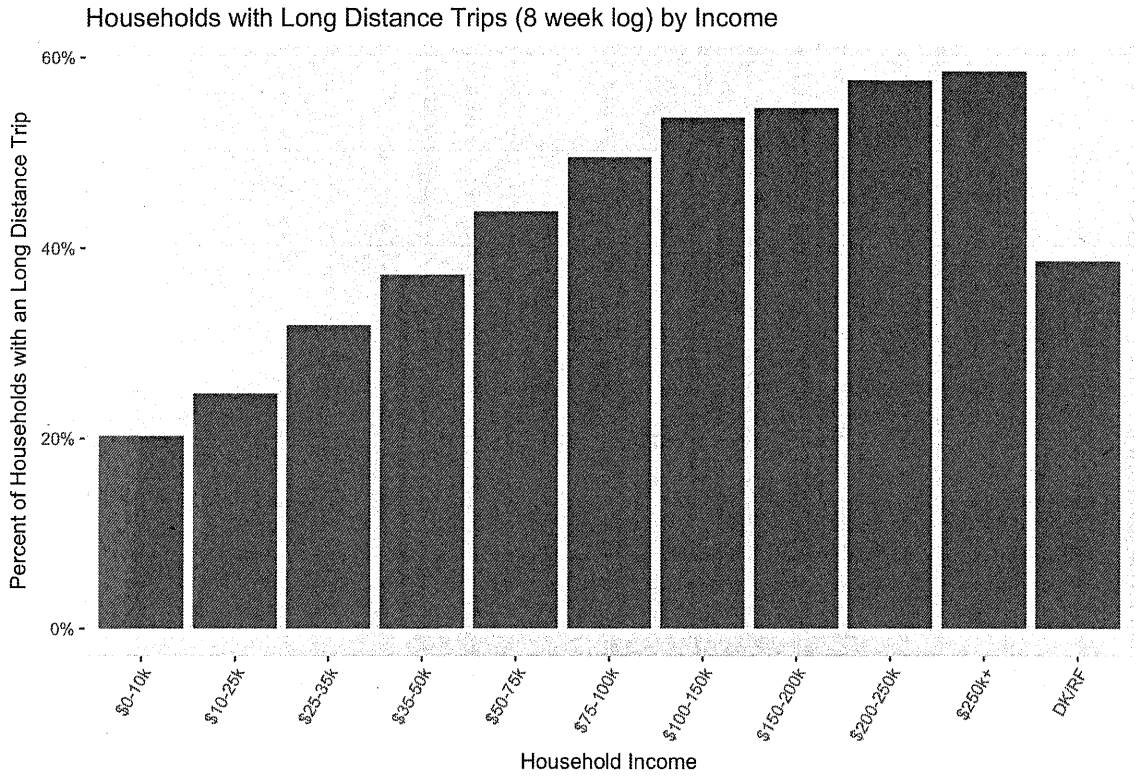


Figure 3.1 Long distance trips and household income

Table 3.4a Counties sorted by sample percentage of households with long distance travel in eight week log (upper half)

County	Households in Sample	%HH with any LD trip	LD trip without overnight	LD trip with 1 overnight	LD trip with 2-6 nights	LD trip with 7+ nights
Mono	107	59.81%	46.73%	7.48%	10.28%	15.89%
Inyo	189	58.20%	35.98%	10.05%	23.28%	13.76%
Mendocino	175	56.00%	42.86%	8.57%	15.43%	6.86%
Modoc	111	54.95%	44.14%	4.50%	13.51%	10.81%
Sonoma	870	54.83%	40.11%	5.75%	17.47%	11.15%
Lassen	152	53.95%	44.74%	5.92%	16.45%	11.84%
Plumas	150	53.33%	43.33%	8.00%	12.67%	6.67%
Siskiyou	212	52.36%	40.57%	6.13%	14.62%	7.55%
San Benito	268	52.24%	43.28%	6.34%	12.31%	7.09%
Santa Cruz	674	52.08%	35.61%	6.68%	15.58%	12.02%
Placer	481	50.52%	32.64%	7.28%	14.35%	9.77%
Colusa	107	50.47%	38.32%	7.48%	16.82%	2.80%
Tuolumne	193	49.74%	36.27%	5.70%	13.99%	9.33%
El Dorado	414	49.52%	35.27%	6.28%	14.98%	12.80%
Napa	317	49.21%	35.33%	6.62%	10.73%	11.99%
Trinity	175	49.14%	37.14%	4.00%	10.29%	8.57%
Amador	182	48.90%	37.36%	7.14%	12.64%	9.89%
Del Norte	189	48.68%	32.80%	5.82%	14.29%	8.99%
Santa Clara	2136	48.60%	34.41%	5.15%	16.25%	10.91%
Nevada	188	48.40%	38.30%	7.45%	15.43%	15.43%
Marin	461	48.37%	33.41%	4.12%	15.62%	11.28%
San Luis Obispo	847	47.93%	31.52%	6.73%	15.47%	9.45%
Lake	182	47.80%	40.66%	4.40%	10.44%	3.30%
Calaveras	176	47.73%	35.80%	6.25%	7.95%	7.39%
Contra Costa	1389	47.37%	34.34%	5.26%	14.69%	10.08%
San Mateo	1142	47.20%	32.84%	6.22%	16.64%	11.65%
Santa Barbara	435	46.44%	31.95%	5.29%	13.56%	7.82%
Butte	360	46.39%	33.33%	5.83%	13.06%	7.78%
Sierra	59	45.76%	37.29%	6.78%	8.47%	5.08%

Table 3.4b Counties sorted by sample percentage of households with long distance travel in eight week log (lower half)

County	Households in Sample	%HH with any LD trip	LD trip without overnight	LD trip with 1 overnight	LD trip with 2-6 nights	LD trip with 7+ nights
Merced	474	45.36%	34.39%	5.70%	9.28%	6.75%
Alameda	1699	45.32%	31.78%	5.59%	13.89%	10.71%
Yolo	246	45.12%	34.15%	8.13%	13.01%	9.35%
Monterey	1022	44.23%	32.68%	5.58%	13.11%	8.90%
San Francisco	1076	44.14%	30.86%	5.48%	14.22%	12.17%
Sacramento	825	44.12%	31.27%	5.09%	12.00%	9.21%
Sutter	168	44.05%	31.55%	4.17%	10.71%	5.36%
Mariposa	148	43.92%	31.76%	4.05%	10.81%	9.46%
Stanislaus	552	43.66%	33.88%	3.62%	10.51%	6.88%
Ventura	1211	43.52%	31.21%	5.53%	12.14%	8.42%
Solano	627	42.42%	32.22%	4.78%	10.37%	7.18%
Humboldt	321	42.37%	26.79%	7.17%	13.71%	8.72%
Madera	311	42.12%	31.51%	6.75%	11.25%	4.82%
Glenn	182	41.76%	32.42%	7.69%	9.89%	4.40%
San Joaquin	629	41.49%	32.91%	4.13%	10.02%	6.20%
Shasta	250	41.20%	28.80%	2.80%	13.20%	9.60%
Tulare	799	40.93%	28.79%	5.01%	11.26%	6.26%
Yuba	205	40.49%	27.32%	5.85%	10.24%	4.88%
San Bernardino	1703	40.22%	29.95%	3.05%	8.75%	5.34%
Tehama	177	39.55%	25.42%	6.78%	11.86%	5.65%
Riverside	1701	39.39%	29.69%	3.23%	8.00%	4.64%
Alpine	23	39.13%	30.43%	8.70%	0.00%	4.35%
Imperial	480	38.96%	32.08%	3.96%	7.92%	4.58%
Orange	2401	38.61%	26.53%	3.08%	10.50%	7.21%
Fresno	1115	38.57%	26.28%	4.22%	11.66%	6.28%
Kern	1544	38.47%	30.12%	4.73%	9.78%	6.09%
San Diego	1688	38.45%	25.06%	4.32%	10.31%	8.53%
Kings	294	35.71%	23.47%	3.74%	10.20%	3.40%
Los Angeles	8219	35.14%	24.35%	3.77%	8.92%	6.34%

Long Distance Travel Distance Comparison

The eight-week travel log is designed to capture information about relatively infrequent travel that a daily diary would miss. As noted in Section 1, it appears to miss shorter, more frequent trips, notably long distance commuting (Bierce and Kurth, 2014). By comparing the data in the long distance log with long distance trips recorded in daily diaries, we can estimate the overall scale of this dataset's distance bias.

To estimate the distance bias, we first must place the daily diary and the eight week log on roughly equal footing in terms of data specificity and time span. Each person submitted an individual diary, whereas the eight-week log was filled out at a household level, so some trips appear multiple times in the daily diaries. As discussed above, a significant number of households excluded trip-by-trip detail from the eight week logs, providing only the general arc of the tour instead of information about each leg and/or only including the trip away from home but not the return trip. To address these inconsistencies, this comparison includes only the longest trip with a destination over 50 miles from home made by each household on a given day in either dataset. Although we have computed route distances for trips in the daily diaries, we base this comparison on straight line distance from home, to allow us to include trips overseas. Once we extract the maximum daily distance for each household, we group the distances into bands and calculate the number of household-days that occur in each band. Totals from the eight-week log are divided by 56 (the number of days in eight weeks) to convert them to per-day estimates. The results from this comparison are shown in Table 3.5

Table 3.5 Ratio of long distance trips per day predicted by daily diary to eight-week log (longest trip per household per day, only including trips away from home)

Distance	*Ratio (from home)*	*Ratio (trip length)*
50-75 miles	6.29	6.81
75-100 miles	5.01	5.05
100-200 miles	4.34	3.94
200-500 miles	2.97	2.28
500-1000 miles	2.38	2.35
1000-3000 miles	1.88	2.05
over 3000 miles	1.34	1.91

Table 3.5 shows the ratio between the number of long distance trips in the daily diary and the number of long distance trips per day in the eight-week log. This ratio is calculated both for the maximum distance from home and for the longest single trip. As the table shows, the daily diary predicts between 6 and 7 times as many trips per household per day as the long distance log in the 50-75 mile range, and this undercounting persists to a lesser extent in all distance bands regardless of whether maximum trip length or maximum distance from home is used. This indicates that the daily diary provides the most accurate representation of the number of long distance trips made by households in the CHTS independent of the length of the trip, so it should be used as the basis for estimating long distance VMT with the synthetic population. However, to understand the determinants of long distance travel when overnight stays are involved requires the use of data from the 8-week travel log. Before moving to the analysis of determinants a study of the potential biases due to self-selection in reporting is needed.

4. Household Patterns

The first analysis of long-distance household travel patterns aims at exploring the differences between households that reported long-distance and those that did not.

Table 4.1 Comparison between households with and without long distance travel information

	HH Size	# of Workers	# of Students	# of Drivers	# of Cars	# of Daily Trips in Diary	Total HHs in CHTS
Households that Report Long Distance Travel	2.59	1.31	0.65	1.98	2.02	9.36	18,008
Households that do not Report* Long Distance Travel	2.56	1.15	0.64	1.78	1.75	7.50	24,423

*We use the term "do not report" because potentially any household can make long distance trips and we cannot distinguish the event of not making any long distance trips from the event of making but not reporting long distance trips in the 8-week period assigned to the household.

Table 4.1 above shows that only 18,008 (42.44%) from the total of 42,431 CHTS households provide information about their long distance travel, and this is sufficient to perform a variety of data analysis and traveler profiling. The averages shown in Table 4.1 also show that households that reported making long distance trips in the eight weeks preceding their interview are significantly different from the households that did not. For example, households with more workers tend to also report long distance trips, as do households that own more cars.

Preliminary Binary and Count Models

A better way to compare the two groups here is to estimate a binary regression model (Logit) using the decision to report long distance trips as the dependent variable and various household characteristics as explanatory variables. Table 4.2 reports the findings of the model and identifies the significant differences between the two groups. Positive

coefficient estimates signify that households are more likely not to report long distance travel. For example, larger households are less likely to report long distance travel, but households with more workers, students, and owning more cars are more likely to report long distance travel, which counteracts the first effect. Interestingly, Hispanic households are more likely to report long distance travel but when the interview was done in Spanish they are more likely not to report long distance travel. Household income shows almost monotonic trends with higher income households more likely to report long distance trips, matching the summary shown in Figure 3.1. Households that refused to report income behaved similarly to those in the category of $50-$75 thousand annual income and the category of "don't know" are closer to the lower income categories. Home ownership and type of living arrangements show the expected sign with people who do not own their home also not reporting long distance travel. Mitra (2016) shows that people who want to buy a home usually move to less expensive areas, which lengthens their commute. This is further confirmed by the coefficient of the living in single home variable of Table 4.2, which indicates an increased propensity for long distance travel. Households that live in mobile homes and large apartment complexes also tend not to report long distance trips. All this shows that households with long distance trips in the CHTS 8-week travel log are significantly different from the households that do not have long distance travel. For this reason caution should be exercised in interpreting the findings of long distance traveling not to generalize to the entire California population without employing methods that can alleviate the data collection biases of the survey discussed in section 3. Moreover, when we estimate regression models of long distance characteristics and include some of the long distance participation determinants (household size, household income, residential arrangements, household composition) we will account in an implicit way for this participation in the long distance component self-selection bias and it is worth exploring this method further.

Table 4.2 Binary Logit Model
(dependent variable: 0 = reported long distance and 1 did not report long distance trips)

	Estimate	Std. Error	z value
(Intercept)	1.256	0.077	16.324
Household Size	0.078	0.015	5.347
# Workers	-0.034	0.014	-2.397
# Students	-0.096	0.017	-5.676
# Cars	-0.101	0.013	-7.650
Interview in Spanish	0.610	0.063	9.749
Hispanic Household	-0.198	0.028	-7.190
Household Income is			
$10,000 to $24,999	-0.168	0.070	-2.389
$25,000 to $34,999	-0.390	0.073	-5.365
$35,000 to $49,999	-0.534	0.070	-7.585
$50,000 to $74,999	-0.736	0.069	-10.695
$75,000 to $99,999	-0.921	0.070	-13.100
$100,000 to $149,999	-1.055	0.071	-14.967
$150,000 to $199,999	-1.076	0.077	-13.983
$200,000 to $249,999	-1.179	0.089	-13.225
$250,000 or more	-1.199	0.088	-13.662
Income Don't Know	-0.223	0.091	-2.444
Income Refused	-0.702	0.076	-9.180
Does not own home	0.141	0.033	4.326
Lives in single family home	-0.082	0.033	-2.462
Lives in mobile home	0.190	0.070	2.723
Lives in bldg with 20+ apartments	0.210	0.053	3.984

We turn now to the analysis of the 18,008 households that reported long distance travel and explore the composition of these long distance travel records. Table 4.3 provides summary statistics of the long distance travel behavior of these households. For 17,123 households we were able to compute tour durations and provide indicators regarding the day of the week when the tour started. There is substantial variation in the total number of trips reported and the distribution of this variable is skewed as shown by the difference between the mean number of trips (3.72) and the median (2.00). This is due to a few observations with many trips, which reflects both that some people took complex multi-

destination trips and that some households reported their long distance trips in much more detail than others did. The number of tours does not show the same extreme patterns.

Table 4.3 Travel characteristics of the 18,008 long distance reporting households

Statistic	N	Mean	St. Dev.	Min	Median	Max
# of Tours	18,008	2.29	2.19	1	2	49
# of Trips	18,008	3.79	3.72	1	2	50
All trips in CA	18,008	0.67	0.47	0	1	1
All trips in the USA	18,008	0.94	0.24	0	1	1
Tours start in weekend	17,123	0.79	1.05	0	1	16
Tours start in weekday	17,123	1.5	1.84	0	1	45
Tours with single trip	18,008	1.02	1.65	0	1	48
Tours partially complete	18,008	0.29	0.78	0	0	22
Complete data	18,008	0.98	1.45	0	1	25
# Tours with no overnight	17,123	1.51	1.89	0	1	47
# Tours with one overnight	17,123	0.15	0.49	0	0	15
# Tours with 2 to 6 overnights	17,123	0.39	0.76	0	0	15
# Tours with 7+ overnights	17,123	0.24	0.53	0	0	6

The number of trips and the number of tours are count data, which makes linear regression models (that assume the dependent variable to be continuous) inappropriate. In addition, Poisson and Negative Binomial regression are created for counts of events in a given time window. The counts of trips and tours refer to a specific interval of 8-weeks. For this reason we can analyze the data and correlate them with household characteristics with a Poisson-type regression model. However, a Poisson regression model imposes an assumption about the data generating process (aka traveler behavior) that is restrictive: the average rate of number of trips (or number of tours) is equal to the variance. Without adding too many complications one can assume that the occurrence of these trips and tours follow a negative binomial distribution. This allows the variance to be different than the mean and offers a more flexible regression model. Table 4.4 shows two models. The first model identifies which variables are significant in explaining variation in the number of long distance trips among the 18,008 households that provided

information. The second model does the same for tours. Positive coefficients show the specific group of households is more likely to have a higher number of trips or tours.

Most of the variables are significantly different than zero indicating they are good predictors of the propensity to have long distance travel and the tendencies are what we expect. Households of higher income are more likely to have long distance trips and tours, and car ownership has a positive effect on trip making. Living in a single family home (as opposed to apartment or mobile home) is also indicating that wealth motivates long distance travel. This is further supported by the indicator of dwelling ownership. Households that do not own their home are more likely to have lower numbers of long distance trips and lower number of long distance tours (captured by the negative and significantly different than zero coefficients in both models of Table 4.4). The number of persons in the household (household size) and the number of workers have negative and significant coefficients indicating the possibility of intra-household constraints in traveling far from home. The average number of long distance trips predicted by the negative binomial models is 3.786 per household in the 8-weeks of interview and the observed is the same because this type of regression model reproduces the observed means. Figure 4.1 shows the distribution of the observed and predicted values for long distance trips among the 18,008 respondents that have a long distance travel log. It is clear in the figure that trips in pairs and in fours are the most often seen in the sample. However, the large mass of observations in the "1" trip category is of concern and needs further scrutiny.

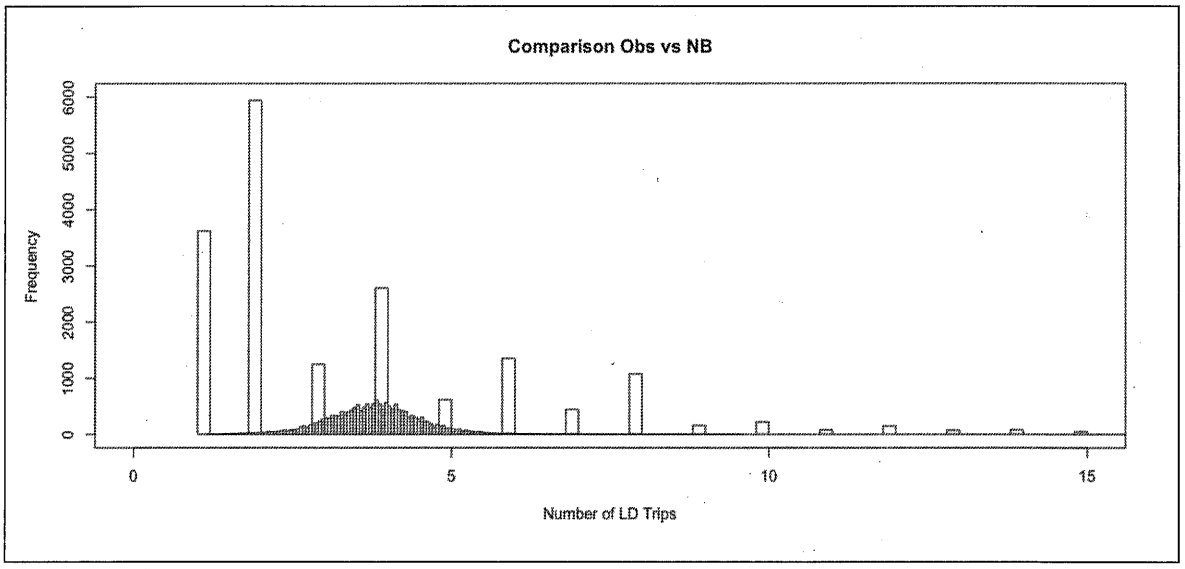


Figure 4.1 Observed versus predicted long distance trips among 18,008 households

All these are useful descriptors of traveler profiles. Table 4.2 shows we have systematic self-selection in reporting long distance travel. The same variables that influence the amount of trip making also influence the reporting decision, so a better approach would be to study the number of long distance trips and the number of long distance tours in a way that can take into account the selectivity bias by the non-response in the long distance component of CHTS.

Table 4.4 Negative Binomial estimates for long distance trips and tours

	Long Distance Trips	Long Distance Tours
Household Size	-0.096***	-0.079***
	(0.009)	(0.009)
# Workers	-0.023***	-0.016**
	(0.008)	(0.008)
# Students	0.001	0.009
	(0.010)	(0.010)
# Cars	0.062***	0.059***
	(0.007)	(0.007)
# Trips in Daily Diary	0.019***	0.015***
	(0.001)	(0.001)
Interview in English	0.226***	-0.063
	(0.049)	(0.047)
Hispanic Household	0.117***	0.069***
	(0.016)	(0.017)
Does Not Own Home	-0.061***	-0.053***
	(0.019)	(0.020)
Lives in Single family Home	0.080***	0.065***
	(0.018)	(0.018)
Household Income is		
$10,000 to $24,999	0.112**	0.054
	(0.053)	(0.054)
$25,000 to $34,999	0.167***	0.125**
	(0.053)	(0.055)
$35,000 to $49,999	0.195***	0.134**
	(0.051)	(0.053)
$50,000 to $74,999	0.269***	0.206***
	(0.050)	(0.051)
$75,000 to $99,999	0.327***	0.272***
	(0.050)	(0.052)
$100,000 to $149,999	0.346***	0.301***
	(0.050)	(0.052
$150,000 to $199,999	0.364***	0.322***
	(0.053)	(0.054)
$200,000 to $249,999	0.403***	0.346***
	(0.057)	(0.058)
$250,000 or more	0.474***	0.409***
	(0.056)	(0.057)
Income Don't Know	0.149**	0.108
	(0.065)	(0.067)
Income Refused	0.339***	0.289***
	(0.054)	(0.055)
Constant	0.632***	0.498***
	(0.067)	(0.066)
Observations	18,008	18,008
Log Likelihood	-41,464.540	-32,829.190
theta	3.033*** (0.053)	5.818*** (0.177)
Akaike Inf. Crit.	82,971.070	65,700.370

Note: *p<0.1; **p<0.05; ***p<0.01

Accounting for Zero Inflation

One method that addresses the inherent bias in the number of long distance trips and tours due to the selective non-response is to estimate a regression model that is able to correct for the self-selection using information from the sample available. In essence this method estimates a probability of participation in the long distance log (as the model in Table 4.2 does) and estimates the number of long distance trips that we should have observed if self-selection bias did not take place. In this section we report on an experiment we did with this type of model.

First we estimate a Poisson regression model that performs a similar task to the models in Table 4.4 and we estimate the expected number of long distance trips. Then, we estimate a zero inflated Poisson regression. This has two components: one estimates the probability of having zero long distance trips and the second estimates the number of long distance trips. The variables used to explain both events (reporting zero trips and the number of trips) are the same as in Tables 4.2 and 4.4 and offer similar indications about the correlation of household characteristics and participation in the long distance log.

We will refer to the long distance trips model that does not account explicitly for self-selection (probability of participation in long distance log) as the Poisson regression model and the model with the two-component structure as the Zero Inflated Poisson (ZIP) model. The mean prediction for the number of long distance trips from the Poisson model is 1.607 trips in 8-weeks and replicates the observed average. The maximum value of the predicted long distance trips in the Poisson model is 16.54 (it is not an integer because it is an expected value in the real line) while the maximum observed number of long distance trips in 50. Figure 4.2 top third shows the distribution of the predicted values of the number of long distance trips from the Poisson model in comparison to the observed values. In essence the Poisson predicted average number of long distance trips reduces substantially the zeros and "spreads" them over the sample in accordance to the household characteristics. In this way wealthier households that show

zero have now a predicted number of long distance trips that is greater than zero. The model however reproduces the observed average long distance trips in the sample that we know is an under-report of the actual long distance trips because many households that have in the database zero long distance are similar to the household that make many long distance trips. This is partially taken into account by estimating a ZIP model that jointly determines the probability of zero trips (making and reporting zero long distance trips) and the expected number of long distance trips. Prediction of the number of long distance trips from this model is an average of 1.797 long distance trips per household in the 8-weeks. The model is reproducing the probability of a household being classified in the zero trip reporting group as 0.5756, which is exactly the number of observed households with no long distance trips records ((42,431-18,008)/42,431). This indicates we have a good model to explain differences between households that have long distance records and those that do not. The middle third of Figure 4.2 shows the predicted distribution of trips in this model. The last third of the figure compares the distribution of long distance trips from the Poisson and the ZIP. The maximum predicted by the ZIP model in 7.02. Figures 4.2, 4.3, 4.4, and 4.5 show how the two models handle the zeros (in similar ways) but also how both models decrease the value of the positive extremes. This happens because the high number of long distance trips (e.g., habitual long distance commutes) were underreported in the 8-week log and the regression model "treat" as extreme outliers. In the next tasks we will develop methods to account for this.

It is possible with the models we estimated using all the 43,431 household data that the large amount of zeros in the number of trips to have been mistakenly taken as legitimate zeros (i.e., households actually did not make any long distance trips). To circumvent this possibility we can use the Negative Binomial model of Table 4.4 and predict the number of long distance trips of the households that did not report a long distance log and then compare this to the observed. Table 4.5 below shows the average and variance number of long distance trips by all three models and the observed. None of the models succeeds in predicting the average number of trips observed and they all under-estimate the variance substantially with the worst being the negative binomial model that is estimated on the 18,008 observations that provide long distance data and then used to predict the

values for the rest of the sample of CHTS. Interestingly the ZIP model is moving towards the right direction of replicating observed behavior but cannot reach the high values of long distance trips that as we discussed earlier is due to structural issues of underreporting in the 8-week diary that cannot be undone with modeling.

Table 4.5 Comparison among the three models and observed data

Long Distance Participant	Negative Binomial 18008	Poisson	Zero Inflated Poisson	Observed
Yes (mean)	1.31	1.85	1.95	3.79
Yes (variance)	0.0458	0.658	0.251	13.8
No (mean)	1.18	1.43	1.69	0
No (variance)	0.0697	0.598	0.249	0

The Negative Binomial predictions using the model estimated on the 18,008 observations that reported long distance trips predicts a narrower range of the number of long distance trips than the other two models. All models, however, as shown by the variance and the spread of values in Figures 4.2, 4.3, 4.4, and 4.5 shrink the variability of the observed counts. This leads us to believe that a wider set of explanatory variables are needed to explain this variation and the exercise here needs to be repeated using this wider set of explanatory variables that capture household context and circumstances that motivate households to make many long distance trips in a 8-week period. As discussed earlier we also have under-reporting of habitual long distance trips such as commuting. For this reason we are exploring methods to differentiate between commuting long distance trips and tours and non-commuting with a parallel analysis of the trips reported in the daily diary.

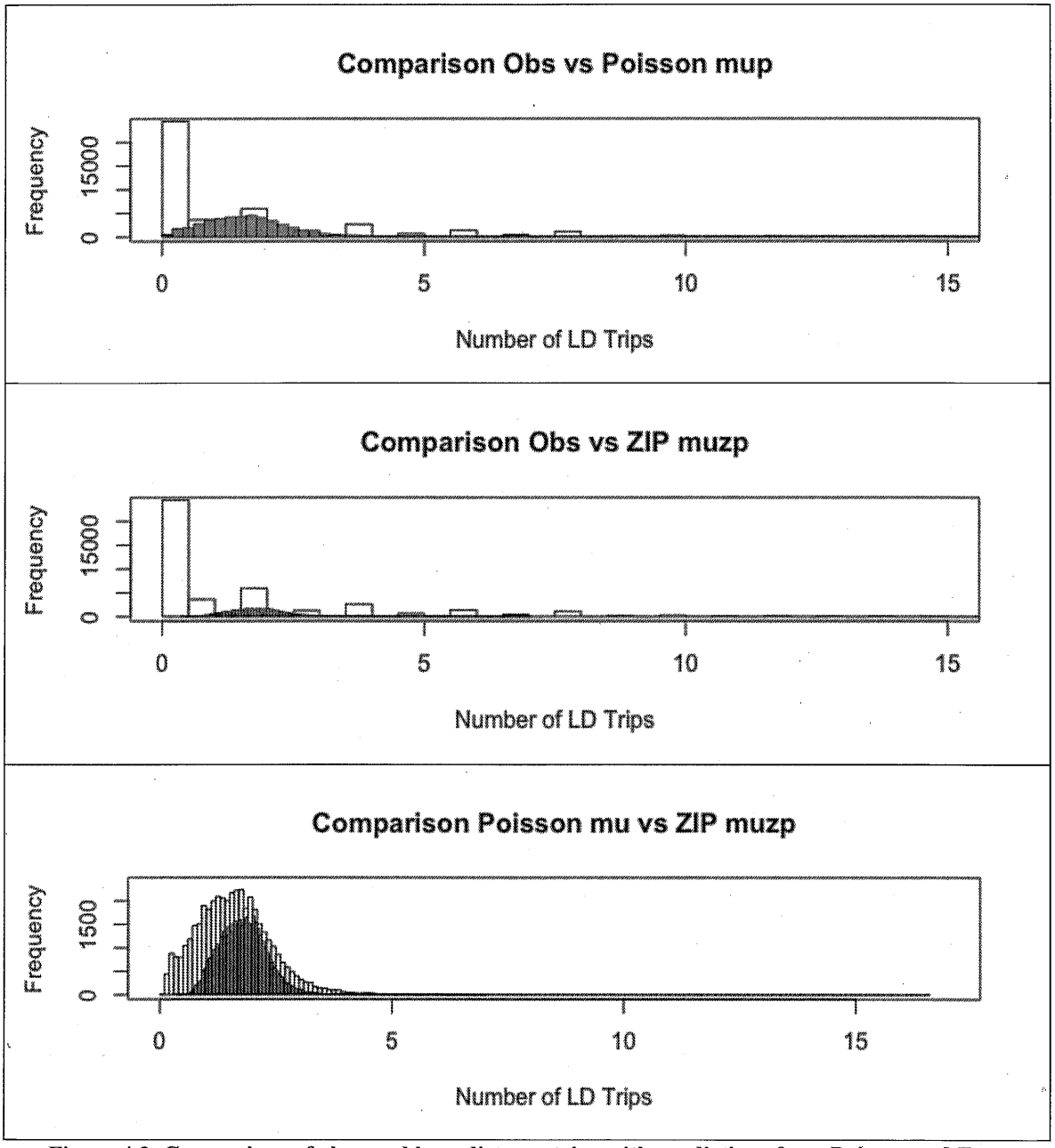


Figure 4.2 Comparison of observed long distance trips with predictions from Poisson and Zero Inflated Poisson

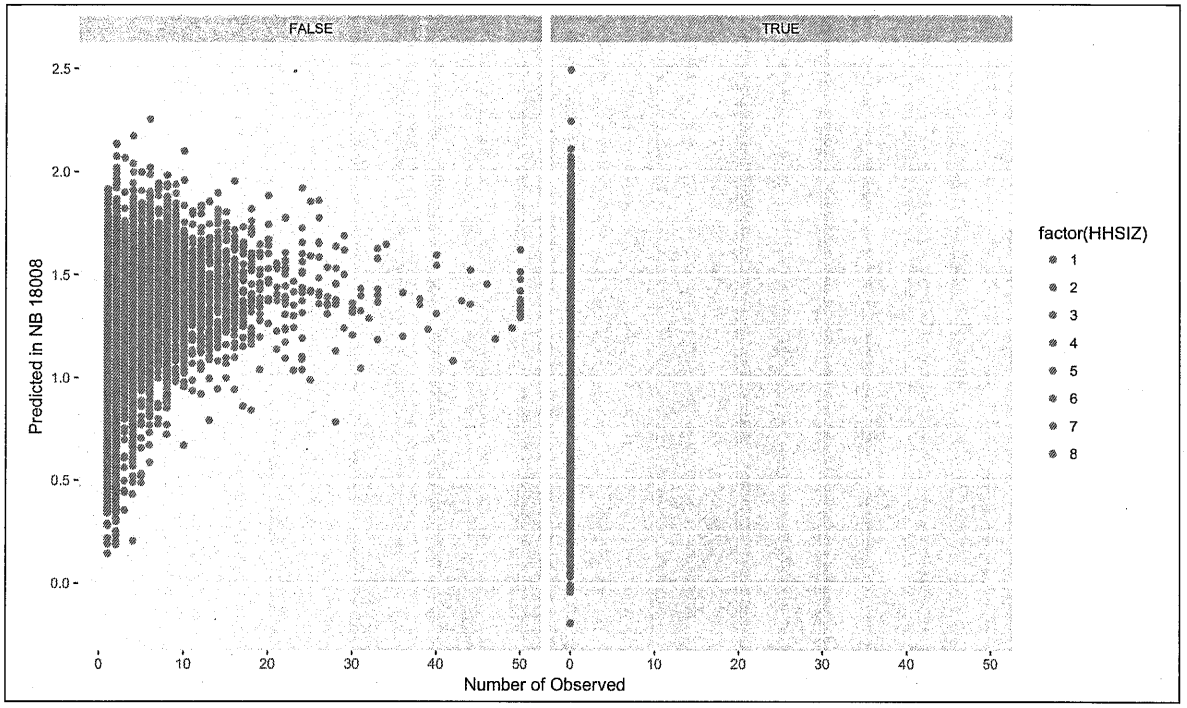


Figure 4.3 Observed versus predicted number of long distance trips in the Negatie Binomial model

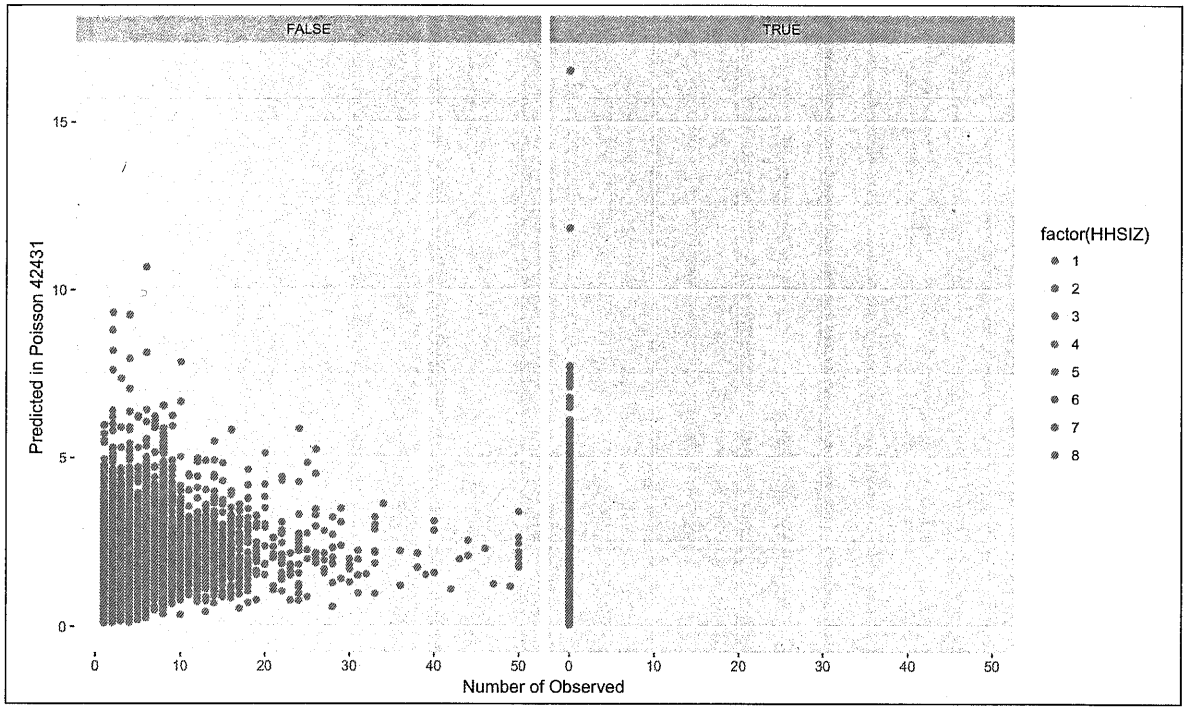


Figure 4.4 Observed versus predicted number of long distance trips in the Poisson model

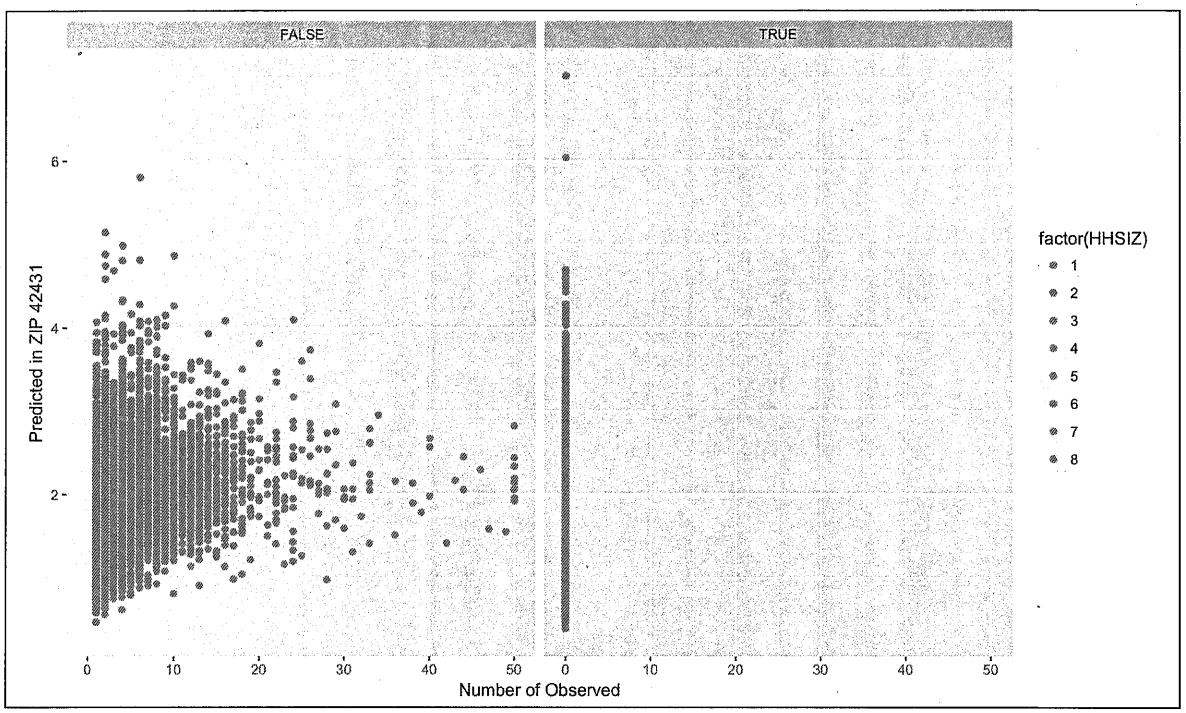


Figure 4.5 Observed versus predicted number of long distance trips in the Zero Inflated Poisson model

A Regression Approach to Self-Selection Bias

The pattern of participation in the CHTS long distance component suffers from self-selection bias. In this section, we present a model that analyzes participation in the long-distance component and attempt to answer the following questions:

1) What household characteristics predict long distance travel survey completeness?
2) How do the characteristics linked to full reporting of long distance travel differ from those that predict partial completeness? and
3) Do these models work equally well everywhere in California?

In this model, we select a 36,925-household subset of the CHTS with complete records for all relevant household variables and use household characteristics to predict whether they reported a long-distance trip and the level of detail of their record. The response variable takes three levels: no long-distance trips (21,107 households), single unattached trips (5,270 households), and complete or partially complete tours (973 households with partial tours only, 9,575 with at least one complete tour).

We use a multinomial logit model to investigate the different factors that predict which category a household belongs to, and all coefficients are reported in reference to the no-LDT group (the reference category). Variables that correspond to making of any long-distance trips will have significant coefficients that are similar between the two outcomes. Variables that have a significant coefficient for tours but not for single trips are useful for distinguishing between households with high-quality complete records and those that did not fully complete the long-distance travel log. Table 4.6 shows the model estimation results.

Table 4.6 Multinomial logit model for self-selection in reporting LDT

	Single Trip Only			Tour		
Estimate	Coeff	SE	P	Coeff	SE	P
(intercept)	-2.5580	0.0689	*0.0000*	-2.0127	0.0544	*0.0000*
Household Size	0.0348	0.0121	*0.0042*	-0.1085	0.0104	*0.0000*
Income	0.1480	0.0082	*0.0000*	0.2179	0.0065	*0.0000*
Homeowner	-0.0298	0.0437	*0.4959*	0.1819	0.0320	*0.0000*
Number of Cars	0.0820	0.0190	*0.0000*	0.0964	0.0153	*0.0000*
Home – Other	-0.2824	0.0981	*0.0040*	-0.3219	0.0779	*0.0000*
Home – Apartment	-0.0214	0.0561	0.7025	-0.1260	0.0449	*0.0050*
High Suburb / Exurb	0.1423	0.0437	*0.0011*	0.0453	0.0341	0.1842
Low S/E or High Rural	0.2861	0.0460	*0.0000*	0.1822	0.0361	*0.0000*
Low Rural	0.4607	0.0484	*0.0000*	0.4145	0.0379	*0.0000*

This model shows that larger households may be more likely to make long-distance trips, but they are less likely to fill out the survey completely, possibly because filling out a large survey for several people requires more effort. Wealthier people (higher incomes and homeowners) are more likely to travel long-distance and more likely to provide complete reports of these trips if they do. People with more household vehicles engage in more long-distance travel, but there is not a significant difference in the coefficients for number of cars between single trips and full tours, so car ownership should not be assumed to affect long distance travel reporting rates. People who live in single family homes are more likely to make and report long-distance trips than people with apartments and other types of dwellings. Residents of all density/centrality categories are more likely to make long-distance trips than residents of city centers (the reference category for this set of categorical variables), and the rate of long distance travel increases the further people live from city centers, since the number of opportunities that can be reached within 50 miles is much higher in urban areas.

The largest source of self-selection bias highlighted by this model surrounds wealthier urbanites. Notably, coefficients on the land use density categories are smaller for full tours than they are for single trips. This difference and the large difference in the two

coefficients on income suggests that the types of long distance travel made by wealthier urban residents may be overrepresented in tour-based analysis of this dataset. The results of our latent-class clustering and the structural equations (discussed in section 7) indicate that these types of people are much more likely to make business trips by plane than people with lower income and residents of less dense areas, and so trips like these are likely to make up a larger share of the travel reported in the CHTS long distance travel supplement than they do in the population. It should also be noted we used the same categories in synthetic population generation confirming these simple indicators of place of residence are an appropriate way to account for diversity in living environments.

Mapping MNL Model Residuals

To determine whether this model fits well for the entire state or whether there are spatial patterns that it misses, we map the model's overall local accuracy across the state.

For each household, the model provides three predicted probabilities for each of the model's outcomes: respectively no long-distance trips, only single trips, and complete trips. We take these three choice probabilities and a record of the household's actual choice (coded as 1 for the choice they made and 0 for the other two choices) and georeference them to the household's home location. We then produce a set of six raster images, corresponding to the three predicted and three observed variables, in which each raster cell takes the sum of the values for a predicted or observed variable of all households within 20 km. The resulting images can be interpreted as a set of contingency tables containing the number of predicted and observed households with each outcome in the surrounding area. We then use Equation 4.1 to calculate the absolute difference between the predicted (P_{cxy}) and observed (O_{cxy}) counts for each outcome at each cell and use these values to extract a total classification error for the area surrounding each cell. This method does not consider the accuracy of any single prediction, but instead determines the accuracy of the predicted totals for each outcome in the area. The 20 km search radius for pixel values was chosen to highlight regional differences, and a smaller

radius might be more useful for examining differences in accuracy within a region of California.

$$E_{xy} = \frac{\sum_c abs(O_{cxy} - P_{cxy})}{2 * \sum_c O_{cxy}} \mid c \in \{NoLD, Single, Tour\} \qquad \text{(Eq. 4.1)}$$

For the map (Figure 4.6), we also consider the sign on the model's prediction error for the NoLD option. Areas in the map shaded dark red represent spatial regions where the model performed poorly and predicted fewer long-distance trips (O-P<0 for NoLD) than were reported. Areas shaded dark blue show the model performed poorly and predicted more long-distance trips than were reported. The model performed relatively well in areas with lighter shades, particularly the cream-colored cells where its overall predictions missed by less than 2%. Many of the areas that appear to have the largest errors (particularly on the east side of the Sierra Nevada and on the Northern California coast) consider the errors for only a few people, and are likelier to have relatively large error percentages as a result.

By this measure the model appears to fit relatively well in most of the cities in the Central Valley and somewhat well in the Bay Area, Los Angeles, and San Diego. The largest errors in somewhat densely populated regions occur in the distant parts of the Bay Area (Marin/Sonoma County and Santa Cruz), the eastern exurbs of the Los Angeles area, and the Santa Barbara area where errors are often higher than 10% and many more households made long distance trips than the model predicted. While these areas are relatively dense and present far more opportunities than most rural areas, they are also located between 50 and 100 miles from areas with many more opportunities available, which provides people a strong incentive to make (relatively short) long-distance trips for shopping and entertainment. These areas belong mainly to the middle two land-use density categories, which have more to do with local land use density than proximity to

urban centers, which suggests that local land use measures are not sufficient for predicting long-distance travel.

This map also shows a regional difference between the Bay Area and Los Angeles. While the model fits relatively well in the dense centers of both regions, residents of the Bay Area are more likely to engage in long-distance travel than the model predicts and residents of LA tend to make less than predicted. This may reflect differences in their regional geographies and transportation systems (a strip from San Francisco to southern San Jose is just over 50 miles whereas one from downtown LA to Irvine is just under), or it may reflect broader cultural or economic differences between the regions that this model is not able to take into account.

In general, this map shows that the simple land use density measures that were sufficient to improve the statewide population synthesis for everyday travel were not sufficient to account for travel error or response bias in long-distance travel. This spatial error and the various forms of response bias identified in this section (and in this model) show that much more care must be taken when extrapolating long-distance travel totals from a limited eight-week survey.

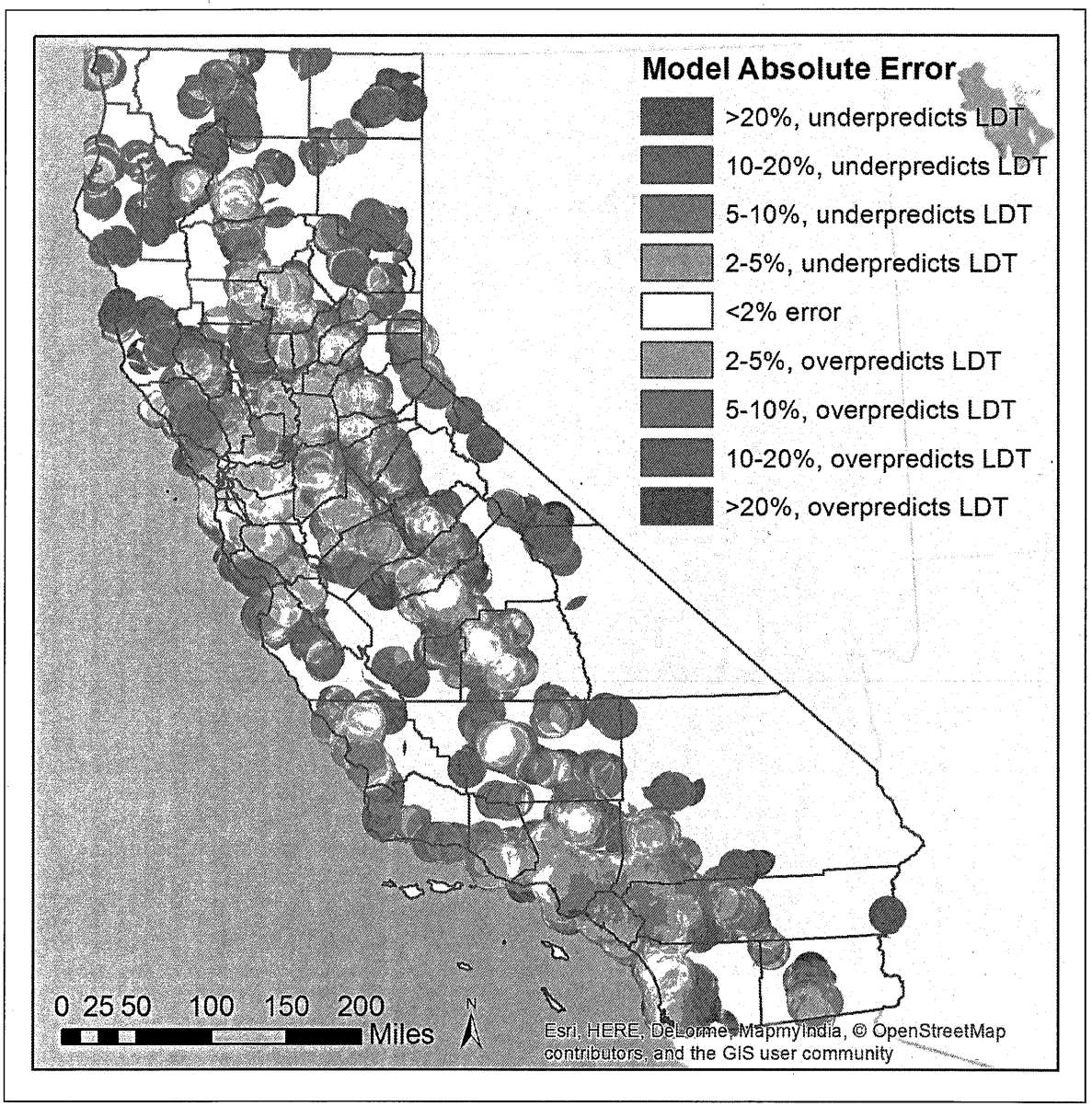


Figure 4.6 Self-selection model error by region in California

5. Statewide Synthetic Population Analysis

The first major objective of this project is to provide a method to compute VMT contributed by long distance travel in California using synthetic population generation techniques. There are a few major advantages of our envisioned method: a) it covers the entire State of California; b) it can be used to extract estimates of long distance travel by specific segments of the population; and c) it can be used by jurisdictions in smaller geographies and for corridors in the State.

To do this, we need to be able to generate a synthetic population at high resolution. From a previous Caltrans project we know that including land use as one of the key variables in developing synthetic population enhances our ability to transfer data from a survey used as seed to the overall population. In this project, we take that analysis one step further by increasing the spatial resolution of the synthetic population generation, used a new version of PopGen (version 2.0) that is more efficient and scalable, and used as control total for each geographic subdivision US Census 2010 data and when needed ACS data. Below is a review of the need to include land use and an illustration of the results and findings.

Description of three synthetic population methods

Program (Software and Method)

The program used to generate all the synthetic populations is called PopGen. This program generates a synthetic population using both household- and person- level characteristics (Bar-Gera et al., 2009, Konduri et al., 2016, MARG, 2016, Ye et al., 2009). It takes variables for which there are known distributions in the areas of interest (e.g., number of 1-person households in a block group), and uses these distributions as the basis upon which respondents from a provided survey are drawn and placed in the areas of interest. It uses an iterative process to replicate the distributions of all the given variables as closely as possible (Ye et al., 2009).

The version of the program used to generate each synthetic population mentioned above differs. PopGen 1.1 was used to create the population that did not use land use and the one that included a coarse land use categorization. PopGen 2.0 was used to create the population with a finer-grained land use categorization for this project. The most important difference between versions 1.1 and 2.0 is that version 2.0 now allows for multiple spatial resolutions for marginal inputs. This means that if some variables of interest are at a coarser spatial resolution than others, it is no longer necessary to default to the coarsest scale to include all of them. Some can be at a "fine" scale, and some at a "coarse" scale. The benefit of this is that it allowed the inclusion of variables from multiple data sources for the marginal distributions: income from the American Community Scale and all other variables from the U.S. Census. The U.S. Census surveys nearly the entire population, so it is a much more reliable source of data if it is possible to use the information it contains.

All Methods

All three synthetic populations were generated using the same set of sociodemographic characteristics as their basis. Household-level variables include *household income*, *age of householder*, *presence of children*, and *number of household members*. Person-level variables include *age* and *gender*. Every characteristic added increases the computation time significantly, so it was important to select as few characteristics as possible while also ensuring a representative population of California in terms of sociodemographics that impact travel behavior.

Although the marginal distribution data sources were slightly different for the "fine land use" population, the survey data upon which PopGen draws to create a synthetic population was the same for the three synthetic populations. The 2012-2013 California Household Travel Survey (CHTS) was used as the source of households for the program. The populations are also the same in that all three were generated at the block group level.

No Land Use

The synthetic population that did not include land use provides a baseline for comparison to the methods that do include land use. This population was generated in the way that

most synthetic populations are generated: using only sociodemographic characteristics as the basis. The variable distributions came from the 2013 American Community Survey (ACS) 5-year estimates to smooth any year to year extreme variation in the ACS sample (http://www.census.gov/programs-surveys/acs/guidance/estimates.html). This is because the ACS provides all the variables we want to use (the Census did not have all of them), and 2013 is the first year the US Census Bureau began making block group level data available.

Coarse Land Use

The population that was created with coarse land use has the same marginal specifications from the ACS as the "no land use" population. The method of including land use involved creating a land use classification scheme, dividing the areas being synthesized into groups based on the category they fall in, dividing the survey respondents based on the category their household falls in, and running the program separately for each category. This process ensures that every area is only synthesized once, and that households are only used to synthesize areas in the land use category they live in. Further details on the method can be found below.

First, we created a kernel density surface of employment density across all of California using a dataset called the 2012 National Establishment Time Series (NETS), which includes comprehensive information about the business establishments in the State. We chose *employee density* because it is a good proxy for how "urban" an area is. We created four categories from this density map by dividing the distribution of densities into quartiles. For clarity, from now on we will call these quartiles rural (low density), exurban (medium-low density), suburban (medium-high density), and urban (high density).

Next, the state was divided into PUMA's, and the average employee density in each PUMA was used to decide which urban category a PUMA would be labeled as. There are 265 PUMA's in all of California, so this classification is quite coarse. Figure 5.1 shows an example of the difference in area between PUMAs and block groups in the city of Los Angeles. The reason we used PUMAs is because PopGen 1.1 asks for PUMA-level

household locations for survey respondents to be used in the creation of the seed matrix that it uses to decide which households to select for a block group.

Finally, the households in the survey were also divided using the PUMA-level classification based on their household location, PopGen was run four times (once for each land use category), and the results were combined to get a synthetic population for the entire state.

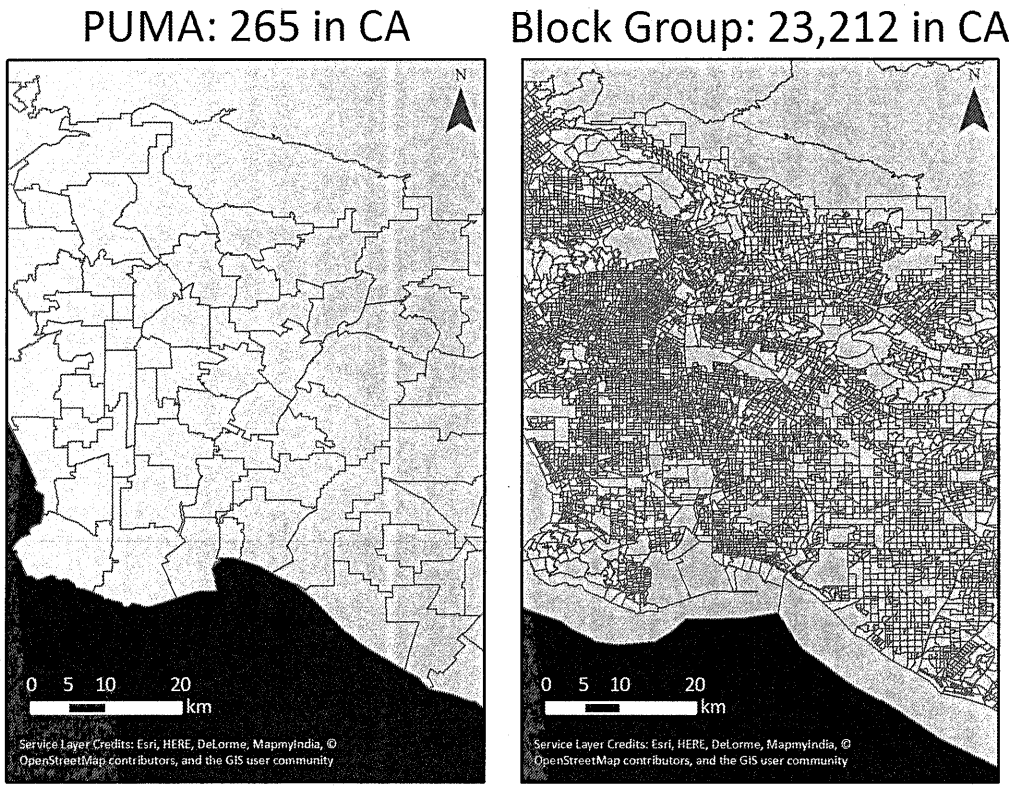


Figure 5.1 PUMA areas versus block group areas (Los Angeles)

This simple classification scheme was initially used because we want to see how coarse it can be while still showing differences in important areas. It also acts as a test of the viability of the method going forward as it gets more complex.

Finer Land Use

The method for including land use in this third population is nearly the same as for the second population, with one key difference. The land use classification is at the block

group level instead of the PUMA level. This was possible because PopGen 2.0 is much more "customizable" than version 1.1. There are 23,212 block groups in California, as opposed to 265 PUMAs. The difference in precision can be visually observed above in Figure 5.1. Aside from the block group level classification, the same method was used: the state was divided into rural, exurban, suburban, and urban areas based on the same measure of employee density, and PopGen was run four times.

For this synthetic population's marginal distributions, 2010 Census data was used for householder age, presence of children, number of household members, person age, and person gender. The 2013 American Community Survey 5-year estimates were used for household income because it is not available through the Census. As mentioned earlier, the reason we changed the data source is because PopGen 2.0 allows for multiple data sources, and since Census data is more accurate than ACS estimates (because it surveys the entire population), we used as many variables as possible from the Census.

Mapping Travel behavior

We took the synthetic populations and transferred travel traits from the CHTS back to the households. This means that every time a respondent was replicated, their travel traits are replicated along with them. From this, we calculated the total miles traveled in each block group (excluding airplane trips) and the total number of trips traveled by the household with residence in each block group. The maps show the number of miles traveled in the block group divided by the number of trips in the block group. What this gives us is the average miles per trip in a day (excluding plane trips) in each individual block group. The number of miles per trip should be higher in a rural area than in an urban area, because they live further away from areas of interest, and people will be more likely to commute farther to get to work.

Table 5.1 shows a comparison of person travel characteristics across the three synthetic population methods and the observed data we used as seed in population synthesis. In this project we aim to develop similar indicators of long distance travel using the fine

land use data population synthesis. Figure 5.2 shows the end result with a map of California. Appendix H contains zoomed in maps for major metropolitan areas.

Table 5.1 Comparison of synthetic populations with observed seed data

	No Land Use	Coarse Land Use	Fine Land Use	CHTS Seed
Number of Trips	3.25	3.30	3.25	3.25
Person Miles Traveled	25.32	24.35	25.54	26.35
Vehicle Miles Traveled	23.53	22.56	23.86	24.83
Number of Non-Motorized Trips	0.55	0.58	0.58	0.47

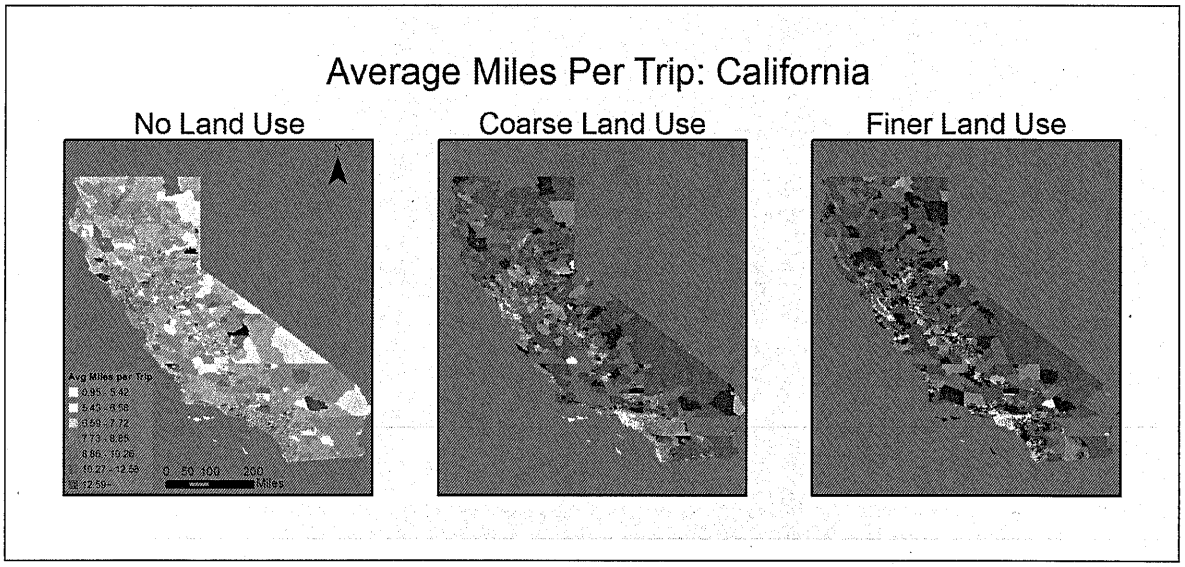


Figure 5.2 Comparison of the three synthetic population methods

Increasing the spatial resolution of the land use classification increases the fidelity of transferred travel behavior information, and this is clearly visible through the maps. In the population with no land use, there is much less of a discernible pattern in the miles per trip. As the land use classification becomes finer, the pattern we would hope to see becomes clearer. Urban areas show the fewest miles per trip (this was studied statistically in McBride et al., 2017), and the further away a block group is from an urban area, the higher the miles per trip becomes. The discernable difference between the two populations with land use bodes well for future plans to create a more complex classification system that includes more than just employee density.

6. Trips Augmentation with attractiveness indicators

The second major objective of the research project here is to complement the long distance trip records with data about the long distance tours. Figure 6.1 shows a conceptual tour of a long distance trip from the statewide model used by CALTRANS. CHTS provides data about the main (central) outbound leg and return leg but not the access and egress portion of this tour. It also does not provide information about opportunities for activity participation at the access station and egress station, home location, and primary destination. In this project we attempt to augment the CHTS trip records with information from social media and other resources that are available online (internet or otherwise). This information will then be used to explain the destination choices of travelers. After this is done we can identify determinants of each long distance behavioral facet including distance traveled and duration of each trip, destinations and modes selected, timing of trips (day of the week and time of day), party size, and tour complexity (number of legs, nodes, and sequencing of trips).

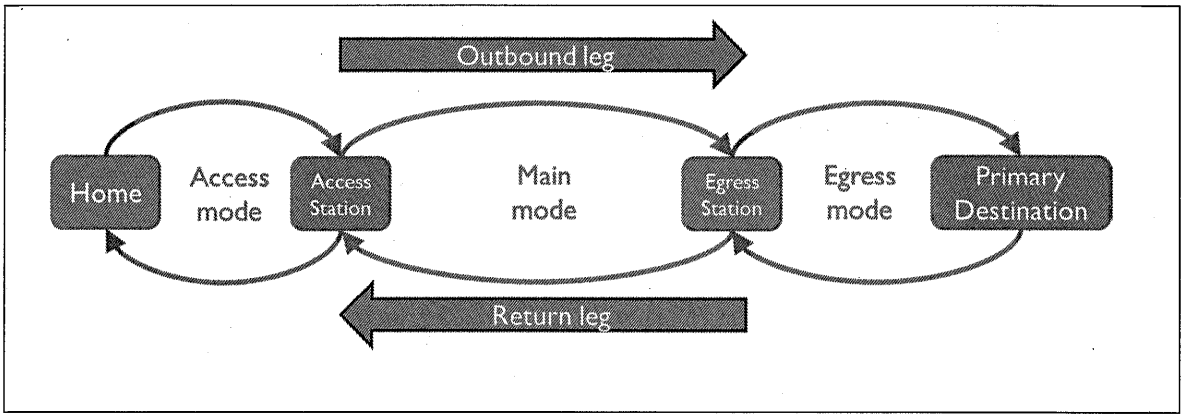


Figure 6.1 Long distance conceptual tour structure (reproduced from CSI, 2014)

The literature review shows that only accessibility and different types of level of service at the destination are used as determinants of travel in the past. These are descriptors of

locations but do not capture the meaning of place (e.g., historical land marks). Also, as noted in section 3 of this report, destinations are recorded with noise (e.g., airports of arrival at destinations instead of locales visited). In this project we explored the potential of social media data and selected to use Foursquare. A list of potential sources and content of information we considered are (also reported in first quarterly report):

- Foursquare (venues): (https://developer.foursquare.com/docs/responses/venue)
- Flicker: (https://www.flickr.com/services/api/)
- DBpedia Places
- Climate data: Daymet (Monthly and 1km) (https://daymet.ornl.gov/)
- Precipitation NOAA (Monthly and by stations): (https://www.ncdc.noaa.gov/cdo-web/datasets)
- DBpedia: (http://wiki.dbpedia.org/OnlineAccess)

Foursquare is a local search-and-discovery service that provides search results for points of interest to each user. The software application provides recommendations that are tailored to each user and until recently allowed users to check and data were collected from the users and provided to research through and API (https://developer.foursquare.com/overview/).

In this project for every destination of approximately 68,000 trips we have computed characteristics of destinations that include the density and diversity of business establishments surrounding the reported longitude and latitude of the destinations and a set of indicators from foursquare derived measures. Figure 6.2 shows the results of this trip record augmentation (in essence we compute many measures of attractiveness for each destination) with longitudes and latitudes masked for privacy reasons. Each column

of this database contains the reported characteristics of each trip and the indicators computed by our team.

Definition of the different attractiveness and statistical analysis indicated a few key variables are sufficient to describe the attractiveness of destinations worldwide as explained in the analytical sections later in this report.

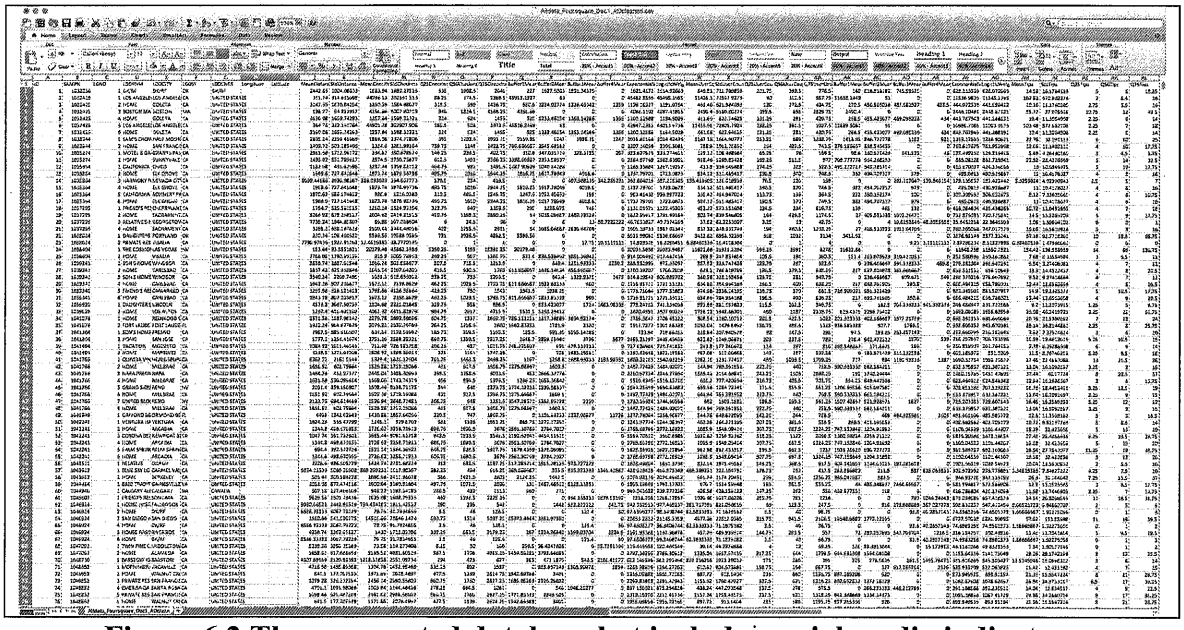


Figure 6.2 The augmented database hat includes social media indicators

In this report we use only a small number of the indicators derived from the foursquare data we assembled. We limited our analysis to indicators that could be used readily in the statistical analysis of tours. For each destination we estimated the space covered by the fifty closest reporting locations (checkins). This space is different depending on the place characteristics (e.g., low density environment) and the willingness of foursquare users to check in. The second variable is the number of individuals checking in these locations. The text of the foursquare report was also analyzed to identify unique topics characterizing each destination but not reported here because it produced a wide variety of groups of words reported by foursquare uses (these groups are called topics) and did not lend themselves to typical statistical modeling.

7. Structural Relations Among Behavioral Variables

In this section we review the findings of the analysis in Task 4 of the project. The first group of analytical techniques contains Structural Equations Models (SEM) that aim at understanding the relationships in the covariance of travel behavior variables with variables representing attractiveness of destinations and social and demographic characteristics of the analyzed households. A parallel approach is to examine relationships among variables and identify different groups of households for which we found the relationship to be fundamentally different. This approach is called Latent Class Analysis and is presented in the second part of this section. We use in both models the same independent (x) and dependent (y) variables.

Data Processing for Tour-Level Analysis and Variables

In this section, we investigate tour characteristics and destination choice at the level of single tours. To do this, we collapse each tour to a single record that contains a mix of household-level variables (income, household size, etc.), tour-level variables (total distance by each mode, total duration, and presence/absence of multiple purposes), and variables that pertain to the tour's primary leg (single trip purpose and measures of destination attractiveness for the primary destination). For this analysis, we consider the primary destination to be the destination at which members of the household stayed for the most time during the tour.

Because we wish to consider as wide a range of tours and destinations as possible, we start by taking a generous approach to identifying tours. Long distance tours are sequences of long distance trips made by a household with end-to-end continuity (the destination location of each trip is the origin of the next trip), beginning with a trip from home and ending with a trip back to home. We expand the pool of potential tours from what we described in the section on data quality ("Extracting Tour Characteristics") by running the tour identification process on households' eight-week long distance logs sorted both by reported long distance trip number and by date, keeping the results of

whichever method produces a larger number of complete home-based tours for a given household (trip number ordering worked better for 91% of households). This analysis also includes tours with incomplete records, but we take steps to account for uncertainty in duration and distance: in tours missing the from-home or to-home trip, the mode-specific mileage from whichever of those trips is present is double-counted in the tour total. This tour identification process leaves us with 23,511 full or partial tours that we may analyze.

We take the following steps to identify the primary leg of each tour to extract the primary purpose and destination characteristics:

- Disqualify trips ending at an airport, with their purpose coded as "Return Home", or ending within 30 miles of home. Some tours have all their destinations eliminated by this process, and 21,584 tours remain.
- 2,128 tours have multiple destinations; we select the longest-duration destination for each of these.
- 522 tours have multiple destinations with equally long durations. To break these ties, we select the destination farthest from home.
- 76 ties remain, which we resolve by selecting the first tour in order.
- Lastly, for this analysis, we eliminate all commute tours.

Tables 7.1a and 7.1b show a list of the variables and their averages. Table 7.2 shows the frequencies of the categorical variables. The total number of observations used in different models in this section is different across models because the treatment of observations with missing data differs from model to model.

Table 7.1a List of variables used in the analysis (household level)

Definition	Level	Mean
Household Size	HH	2.55
Number of Employees	HH	1.33
Number of Students	HH	0.63
Number of Licensed Drivers	HH	2.03
Number of Cars	HH	2.08
Number of Bikes	HH	1.95
Income (category, treated as numeric)	HH	6.04
Homeowner	HH	0.87
Number of trips in Daily Diary	HH	10.35
Hispanic Household	HH	14.7%
Total density (emps/km2) around home	HH	1,123.54
Agriculture density (emps/km2) around home	HH	5.03
Mining density (emps/km2) around home	HH	1.10
Utilities density (emps/km2) around home	HH	5.28
Construction density (emps/km2) around home	HH	43.93
Manufacturing density (emps/km2) around home	HH	57.11
Wholesale trade density (emps/km2) around home	HH	34.66
Retail trade density (emps/km2) around home	HH	104.96
Transportation and warehousing density (emps/km2) around home	HH	19.48
Information density (emps/km2) around home	HH	46.75
Finance, insurance, real estate and rental and leasing density (emps/km2) around home	HH	212.35
Professional services density (emps/km2) around home	HH	136.00
Educational services density (emps/km2) around home	HH	102.07
Health care density (emps/km2) around home	HH	66.81
Entertainment and food services density (emps/km2) around home	HH	105.00
Other services density (emps/km2) around home	HH	106.05
Public administration and armed force density (emps/km2) around home	HH	76.18
Distance from Household to Business Center (meters)	HH	11,771.09
Distance from home to nearest airport of any size (meters)	HH	10,368.94
Distance from home to nearest international airport (meters)	HH	80,098.20
Distance from home to nearest freeway (meters)	HH	7,377.93

Table 7.1b List of variables used in the analysis (tour and destination levels)

Definition	Level	Mean/Percentage
Total driving distance (miles)	Tour	276.46
Total passenger distance (miles)	Tour	51.50
Total flying distance (miles)	Tour	485.84
Total ground transit distance (miles)	Tour	18.22
Total other mode distance (miles)	Tour	4.17
Total unknown mode distance (miles)	Tour	5.41
Any trips in tour with purpose Business (work-related meeting/convention/seminar)	Tour	14.2%
Any trips in tour with purpose Combined business and pleasure	Tour	2.8%
Any trips in tour with purpose School-related activity	Tour	2.4%
Any trips in tour with purpose Visit friends/family/relatives	Tour	35.3%
Any trips in tour with purpose Medical	Tour	5.0%
Any trips in tour with purpose Vacation/sightseeing	Tour	18.7%
Any trips in tour with purpose Outdoor recreation (sports, fishing, hunting, camping, boating, etc)	Tour	8.0%
Any trips in tour with purpose Entertainment (theater, concert, sports event, gambling, etc)	Tour	8.3%
Any trips in tour with purpose Personal Business (e.g. shopping)	Tour	7.1%
Area of convex hull of 50 POIs around primary destination (square meters)	Dest	39,265,844.44
Log of convex hull area (ln m2)	Dest	16.47
Entropy of POI types at 50 POIs around primary destination	Dest	4.68
Median checkins at 50 POIs around primary destination	Dest	1,052.33
Median rating at 50 POIs around primary destination	Dest	7.21
Median users at 50 POIs around primary destination	Dest	499.26
Destination in California	Dest	0.79

Table 7.2 List of variables used in the analysis (categories and counts)

Variable	Level	Value	Count
Household Block Group Center Class	**HH**	Center	4,238
		Suburb	2,677
		Exurb	5,363
		Rural	5,451
Household Block Group Category	**HH**	Center	4,238
		High Density Suburb/Exurb	4,879
		Low Density Suburb/Exurb or High Density Rural	4,444
		Low Density Rural	4,168
Home Type	**HH**	SingleHome	15,756
		Apartment	1,313
		Other	660
Tour Duration	**Tour**	One overnight	2,244
		Single-day	6,382
		Two-six overnights	6,070
		Seven+ overnights	2,403
		Unknown	630
Day of the Week of Tour Start	**Tour**	Sunday	2,331
		Monday	1,746
		Tuesday	1,752
		Wednesday	1,967
		Thursday	2,282
		Friday	3,473
		Saturday	3,548
		Unknown	630
Primary Purpose of Tour	**Tour**	Business (meeting/convention/seminar)	2,370
		Combined business and pleasure	466
		School-related activity	398
		Visit friends/family/relatives	5,939
		Medical	833
		Personal Business	1,170
		Vacation/sightseeing	3,123
		Outdoor recreation	1,353
		Entertainment	1,364
		Drive someone else / DK / RF	713
Primary Purpose (simplified)	**Tour**	Business	2,836
		Personal Business	2,003
		Recreational	12,890
Season	**Tour**	Shoulder (Other)	6,250
		Summer (May 15 - Sep 15)	6,846
		Winter (Nov 15 - Mar 15)	4,633
Region in USA	**Dest**	California	13,973
		Pacific	1,908
		Southwest	267
		Plains	154
		Midwest	227
		Northeast	368
		Southeast	404
		not USA	418
		DK/RF	9
		NA	1

Structural Equations Models

In this subsection we review two types of Structural Equations Models (SEM), one with latent variables and another without latent constructs also called Path Model. In the SEM with latent variables we identify latent factors that explain the variation in observed outcomes (long distance travel behavior) and correlate them with "causes" that we think determine behavior. These latent factors represent predispositions to behave in certain ways (e.g., visiting places with specific characteristics). In Path Models we build regression equations that use as dependent variables behavioral outcomes and independent (explanatory) variables the determinants of behavior. Both models reveal different aspects in the correlation structure among observed variables and are best when complex situations need to be analyzed. Appendix I contains additional details about these two methods.

Structural Equations Model with Latent Variables

Figure 7.1 is the path diagram of the first model estimated. It contains four factors (latent variables) representing four different aspects of long distance travel of the tours analyzed here. Factor 1 (labeled "main") is the latent variable that determines the level of the miles of travel by air, driving a car, and public transportation. It represents the amount of travel a household allocates to each long distance tour. The second factor (labeled "traits") is the latent variable that determines the different combinations of choices households make in their long distance tours and explains the variation in the amount of overnight stays of the tour, the season during which the tour was made, the purpose of the trip with the longest stay and an indicator if the tour was in California. The third factor labeled "purpose" explains the variation in and the composition of the trip purposes in the tour and explains the variation in the main trip purpose and also includes the number of trips in work related business, vacation, and outdoor recreation. The last factor is reserved for the attractiveness of the main destination as represented by foursquare social media indicators and includes the logarithm of the area covered by the 50 closet checkins, the number of users in these checkins and the ratings they gave to the locations surrounding the main destination.

The main factor and the purpose factor are influenced in a significant way by all the exogenous (x) variables depicting the type of household and these include variables capturing the household's wealth such as type of dwellings unit, number of cars, number of employed persons (see also Tables 7.3 and 7.4). Also included is the household size, and an indicator if the household is Hispanic. They also include an indicator of the location in which the household resides (center city, suburb, exurb, and rural). We also include the number of trips the household made in their daily diary to allow for tradeoffs between long distance travel in the past 8-weeks and daily travel behavior.

Overall we see that in addition to the household characteristics influencing long distance travel, the place and type of residence play an important and significant role in shaping long distance travel patterns. We also see that foursquare does provide significant indicators of attractiveness of the main travel destination and shows it is worthwhile continuing the collection of data of this type via an observatory that combines behavioral data from diaries with land use data of the residence and social media data. However, models of this type can become extremely complex and estimation of their parameters tedious and often impossible. This motivates the next analysis with just observed variables with some simplification of the categorical variables used here.

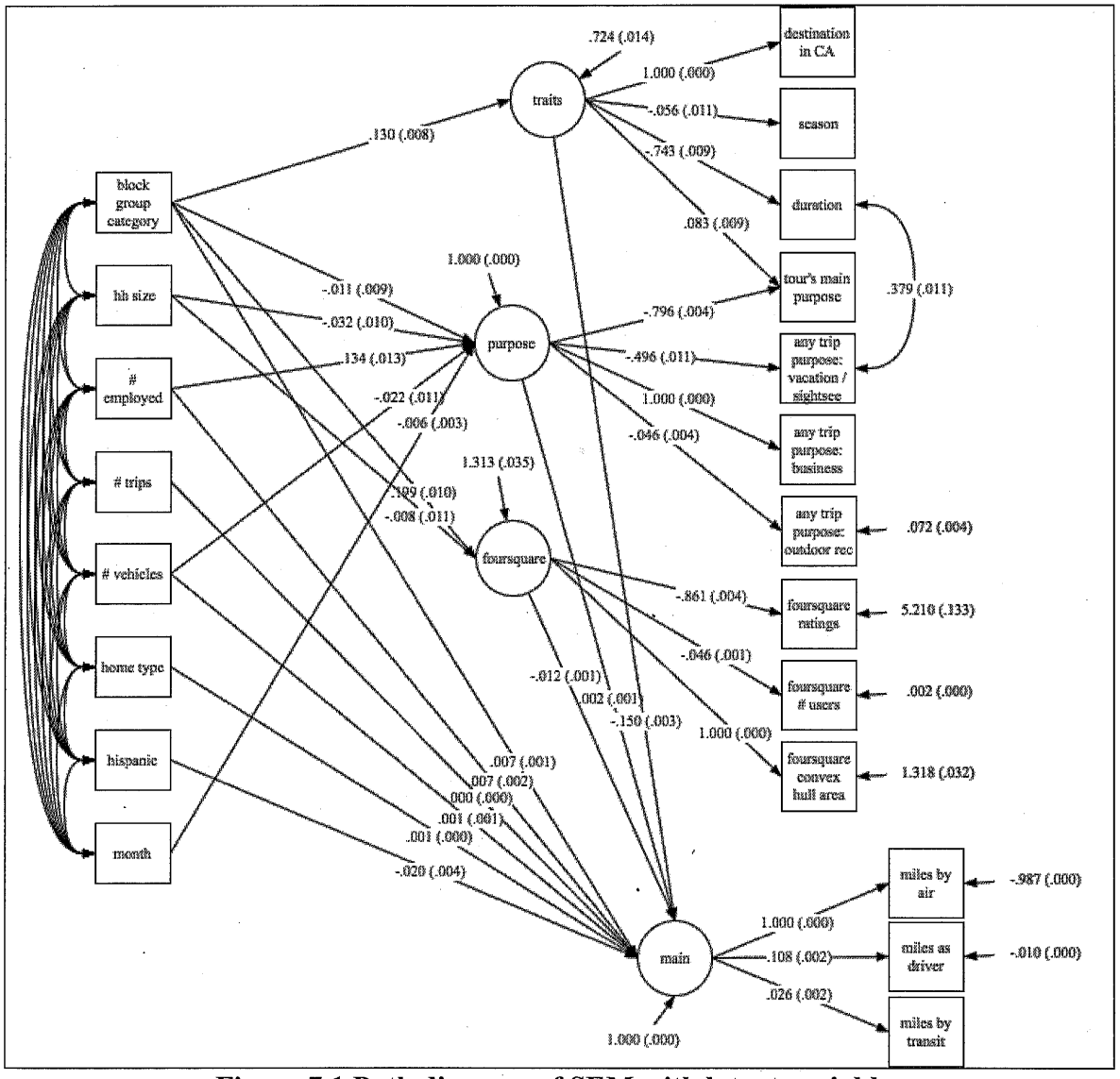


Figure 7.1 Path diagram of SEM with latent variables

Table 7.3 Latent variables in SEM *(factor loadings)*

Factor	Variable	Two-Tailed Estimate	S.E.	*p*-value
Main	Total Miles by Air in Tour	1.000	0.000	999.000
	Total Miles as Car Driver in Tour	0.108	0.002	0.000
	Total Miles by Ground Transit in Tour	0.026	0.002	0.000
Purpose	Any Trips in Tour with Purpose Business (meeting/convention/seminar)	1.000	0.000	999.000
	Any Trips in Tour with Purpose Vacation/sightseeing	-0.496	0.011	0.000
	Any Trips in Tour with Purpose Outdoor recreation	-0.046	0.004	0.000
	Long Distance Tour Main purpose (10 categories)	-0.796	0.004	0.000
Foursquare	Foursquare: Convex Hull Area of 50 nearest POIs (log km2)	1.000	0.000	999.000
	Foursquare: Median Users at 50 POIs Around Destination	-0.046	0.001	0.000
	Foursquare: Median Number of Ratings at 50 POIs Around Destination	-0.861	0.004	0.000
Traits	Destination in California	1.000	0.000	999.000
	Long distance Tour Main purpose (10 categories)	0.083	0.009	0.000
	Season of Tour	-0.056	0.011	0.000
	Duration Category	-0.743	0.009	0.000

*999 means the coefficient is fixed

Table 7.4 Regression models of the latent variables in SEM

Factor		Two-Tailed Estimate	S.E.	*p*-value
Main	Purpose (factor)	0.002	0.001	0.041
	Foursquare (factor)	-0.012	0.001	0.000
	Traits (factor)	-0.150	0.003	0.000
	Number Employed in Household	0.007	0.002	0.000
	Household Trips in Daily Diary	0.000	0.000	0.018
	Vehicles in Household	0.001	0.001	0.643
	Household Home Type	0.001	0.000	0.101
	Household is Hispanic	-0.020	0.004	0.000
	Household Block Group Category	0.007	0.001	0.000
Purpose	Household Size	-0.032	0.010	0.002
	Number Employed in Household	0.134	0.013	0.000
	Vehicles in Household	-0.022	0.011	0.055
	Household Block Group Category	-0.011	0.009	0.183
	Month of Tour	-0.006	0.003	0.039
Foursquare	Household Size	-0.008	0.011	0.502
	Household Block Group Category	0.199	0.010	0.000
Traits	Household Block Group Category	0.130	0.008	0.000

Structural Equations Model without Latent Variables (Path Analysis)

There are some important differences between the dependent variables in this analysis and the SEM with latent variables. First we use a software that allows to declare the number of miles flown, driven, and by public transportation as censored variables to account for the large number of tours that may have zero miles for each of these modes. Second we reduced the categories of the main purpose of the tour to three (3). The first category is for relatively flexible not mandatory trips to visit relatives, vacation, outdoor recreation and related, the second is for business and combined business and leisure trips, and the last category is for shopping and medical. In the Logit model that is included in the path model this is the reference category. In a categorical regression model one category is used as the reference for identification purposes and the regression coefficients should be interpreted in a relative way as we explain shortly.

We also recoded the overnight stays in a way that tours with same day (no overnight stays) are used as the reference category in another Logit model used in this path analysis. Figure 7.2 shows the path diagram for the path analysis model. The left hand side variables are the exogenous variables (determinants of the travel indicators) and in this formulation are the variables that motivate households to behave in a certain way. In this case, in addition to the sociodemographic and place of residence variables, we also include the decision to stay in California for the long distance trip as determinant. We also consider in the cascade of the relationships (this model is also called recursive in which a set of variables are determined first and then another set is considered to be a function of exogenous variables and a function of the first column of dependent variables). The miles flown, driven, and by public transportation are first in the cascade followed by the overnight stays and main trip purpose of the long distance tour. The arrows in Figure 7.2 are the regression coefficients with blue the positive coefficients and red the negative. These are also shown in Tables 7.5a, 7.5b, 7.5c, 7.6a, 7.6b, 7.6c, 7.7a, 7.7b, and 7.7c with their significance.

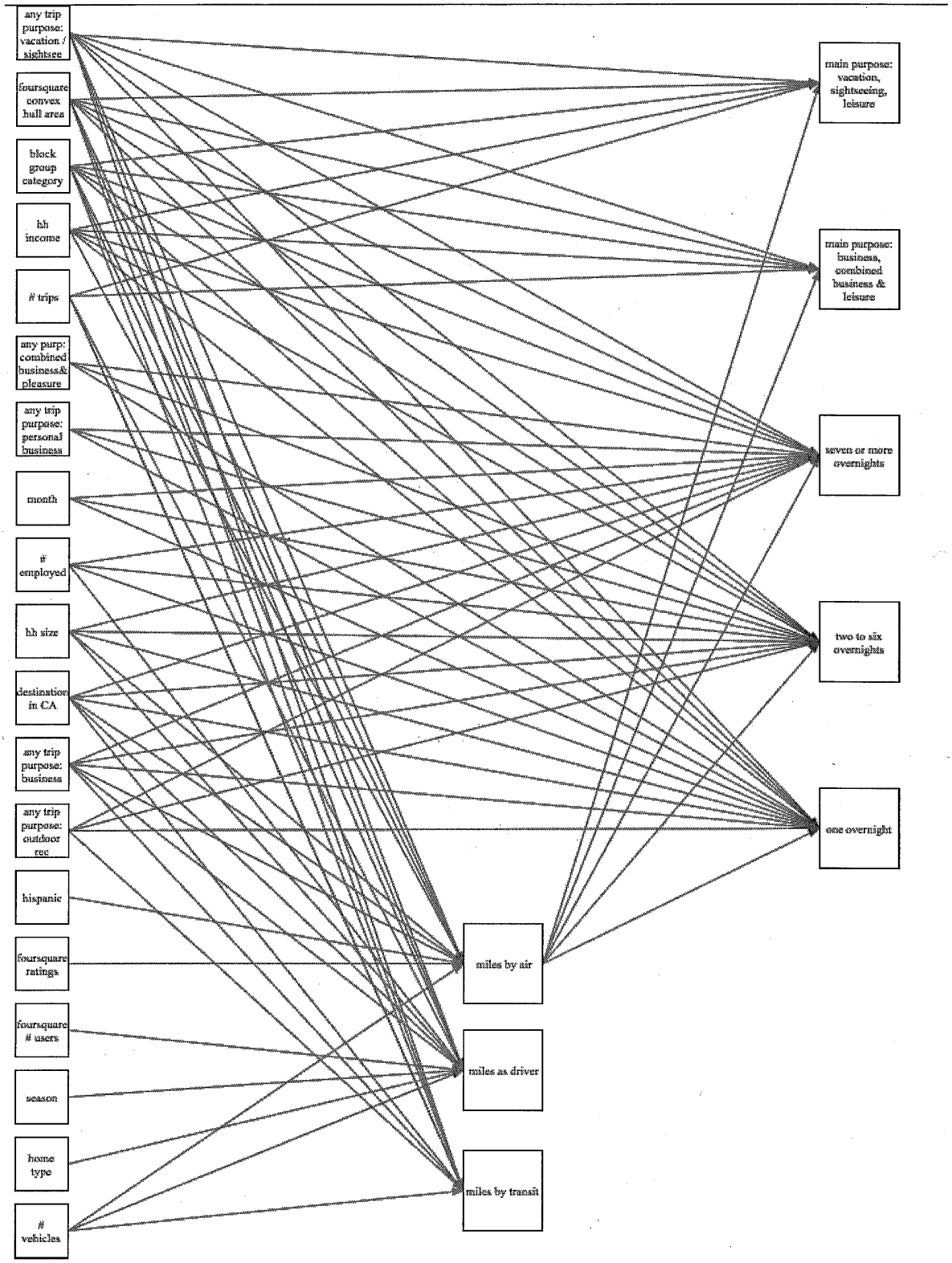


Figure 7.2 Path diagram of the path analysis model

Table 7.5a Regression of miles traveled by air

Variable	Two-Tailed Estimate	S.E.	*p*-value
Any Trips in Tour with Purpose Business (meeting/convention/seminar)	0.267	0.016	0.000
Any Trips in Tour with Purpose Vacation/sightseeing	0.073	0.016	0.000
Destination in California	-0.857	0.016	0.000
Household Size	-0.018	0.006	0.003
Vehicles in Household	-0.032	0.008	0.000
Household is Hispanic	-0.049	0.020	0.017
Household Block Group Category	-0.078	0.006	0.000
Household Annual Income	0.052	0.003	0.000
Foursquare: Convex Hull Area of 50 nearest POIs (log km2)	-0.025	0.004	0.000
Foursquare: Median Number of Ratings at 50 POIs Around Destination	0.014	0.003	0.000
Intercept	0.240	0.075	0.001

Table 7.5b Regression of miles traveled driving

Variable	Two-Tailed Estimate	S.E.	*p*-value
Any Trips in Tour with Purpose Business (meeting/convention/seminar)	-0.010	0.001	0.000
Any Trips in Tour with Purpose Vacation/sightseeing	0.010	0.001	0.000
Destination in California	-0.008	0.001	0.000
Season of Tour	0.002	0.001	0.000
Household Size	-0.001	0.000	0.033
Number Employed in Household	0.001	0.001	0.024
Household Trips in Daily Diary	0.000	0.000	0.000
Vehicles in Household	0.004	0.001	0.000
Household Home Type	-0.001	0.000	0.005
Household Block Group Category	0.003	0.000	0.000
Household Annual Income	-0.001	0.000	0.033
Foursquare: Convex Hull Area of 50 nearest POIs (log km2)	0.001	0.000	0.007
Foursquare: Median Users at 50 POIs Around Destination	-0.023	0.008	0.004
Foursquare: Median Number of Ratings at 50 POIs Around Destination	0.000	0.000	0.063
Intercept	0.003	0.006	0.682

Table 7.5c Regression of miles traveled by transit

Variable	Two-Tailed Estimate	S.E.	*p*-value
Any Trips in Tour with Purpose Business (meeting/convention/seminar)	-0.021	0.012	0.075
Any Trips in Tour with Purpose Vacation/sightseeing	0.020	0.009	0.026
Any Trips in Tour with Purpose Outdoor recreation	-0.041	0.017	0.018
Destination in California	-0.036	0.008	0.000
Household Trips in Daily Diary	0.001	0.000	0.077
Vehicles in Household	-0.020	0.005	0.000
Household Block Group Category	-0.010	0.003	0.003
Household Annual Income	-0.007	0.002	0.001
Foursquare: Convex Hull Area of 50 nearest POIs (log km2)	-0.016	0.002	0.000
Intercept	0.027	0.037	0.453

Table 7.6a Regression of *main tour trip's purpose vacation, sightseeing, leisure*

Variable	Two-Tailed Estimate	S.E.	*p*-value
Any Trips in Tour with Purpose Vacation/sightseeing	3.492	0.245	0.000
Household Trips in Daily Diary	0.018	0.004	0.000
Household Block Group Category	-0.482	0.027	0.000
Household Annual Income	0.119	0.014	0.000
Foursquare: Convex Hull Area of 50 nearest POIs (log km2)	0.157	0.016	0.000
Total Miles by Air in Tour	3.596	0.549	0.000
Intercept	-0.454	0.285	0.111

Table 7.6b Regression of *main tour trip's purpose work related business and combined business and leisure*

Variable	Two-Tailed Estimate	S.E.	*p*-value
Any Trips in Tour with Purpose Vacation/sightseeing	-0.521	0.326	0.110
Household Trips in Daily Diary	0.023	0.004	0.000
Household Block Group Category	-0.395	0.031	0.000
Household Annual Income	0.259	0.016	0.000
Foursquare: Convex Hull Area of 50 nearest POIs (log km2)	0.006	0.019	0.768
Total Miles by Air in Tour	5.339	0.554	0.000
Intercept	-0.468	0.334	0.162

Table 7.7a Regression of seven or more overnights

Variable	Two-Tailed Estimate	S.E.	*p*-value
Any Trips in Tour with Purpose Business (meeting/convention/seminar)	-0.547	0.090	0.000
Any Trips in Tour with Purpose Combined business and pleasure	0.715	0.159	0.000
Any Trips in Tour with Purpose Vacation/sightseeing	1.714	0.071	0.000
Any Trips in Tour with Purpose Outdoor recreation	0.484	0.104	0.000
Any Trips in Tour with Purpose Personal Business (e.g. shopping)	-0.636	0.114	0.000
Destination in California	-1.869	0.079	0.000
Month of Tour	-0.018	0.009	0.036
Household Size	-0.159	0.025	0.000
Number Employed in Household	-0.063	0.036	0.082
Household Block Group Category	-0.111	0.024	0.000
Household Annual Income	0.062	0.015	0.000
Foursquare: Convex Hull Area of 50 nearest POIs (log km2)	0.065	0.017	0.000
Total Miles by Air in Tour	9.189	0.668	0.000
Intercept	-0.550	0.300	0.070

Table 7.7b Regression of two to six overnights

Variable	Two-Tailed Estimate	S.E.	*p*-value
Any Trips in Tour with Purpose Business (meeting/convention/seminar)	-0.264	0.060	0.000
Any Trips in Tour with Purpose Combined business and pleasure	0.733	0.121	0.000
Any Trips in Tour with Purpose Vacation/sightseeing	1.283	0.059	0.000
Any Trips in Tour with Purpose Outdoor recreation	0.608	0.069	0.000
Any Trips in Tour with Purpose Personal Business (e.g. shopping)	-1.073	0.085	0.000
Destination in California	-1.517	0.068	0.000
Month of Tour	-0.002	0.006	0.702
Household Size	-0.130	0.018	0.000
Number Employed in Household	0.073	0.026	0.005
Household Block Group Category	-0.138	0.018	0.000
Household Annual Income	0.072	0.011	0.000
Foursquare: Convex Hull Area of 50 nearest POIs (log km2)	0.027	0.012	0.030
Total Miles by Air in Tour	7.320	0.663	0.000
Intercept	0.737	0.227	0.001

Table 7.7c Regression of one overnight

Variable	Two-Tailed Estimate	S.E.	*p*-value
Any Trips in Tour with Purpose Business (meeting/convention/seminar)	-0.095	0.072	0.187
Any Trips in Tour with Purpose Combined business and pleasure	0.394	0.157	0.012
Any Trips in Tour with Purpose Vacation/sightseeing	0.689	0.077	0.000
Any Trips in Tour with Purpose Outdoor recreation	0.119	0.096	0.212
Any Trips in Tour with Purpose Personal Business (e.g. shopping)	-0.885	0.108	0.000
Destination in California	-0.455	0.096	0.000
Month of Tour	0.007	0.008	0.395
Household Size	-0.114	0.023	0.000
Number Employed in Household	0.089	0.033	0.007
Household Block Group Category	-0.077	0.023	0.001
Household Annual Income	0.058	0.014	0.000
Foursquare: Convex Hull Area of 50 nearest POIs (log km2)	-0.010	0.016	0.538
Total Miles by Air in Tour	5.941	0.717	0.000
Intercept	-0.521	0.290	0.072

The amount of air miles flown is influenced by household characteristics, traits of the tour and foursquare indicators. In summary this model shows that households with higher income are more likely to travel by air for longer distances but larger households with many cars, living in exurbs and rural environments, and Hispanic households are less likely to travel by air long distances. They are more likely to fly long distance when one or more trips in a tour are for work related business and/or vacation. The foursquare relationships show tours with more air miles are more likely to be in denser areas (e.g., big cities) that received higher foursquare ratings. When the destination is in California the tour are more likely to have a smaller number of miles flown.

The amount of miles driven by car show that tours with a higher number of mileage are done by households with more employed persons, with higher car ownership levels, living in exurbs and rural environments. Larger households are less likely making tours with many miles driven and households that do not live in single homes are less likely to drive far. Trip purposes have similarity to the air miles for vacation but the opposite sign for work related business. The foursquare variables show that these are also tours in denser areas but not with the same higher ratings of attractiveness as the air miles.

The miles riding public transportation (presumably train) shows that households of lower income, with a lower number of cars, and living in central locations are more likely to make tours with more public transportation miles. Secondary trip purposes also play an important role similar to the miles driven with the added effect of trips to outdoor/recreation trips that are less likely to be done by traveling long distances in public transportation. These tours are also more likely at destinations in central city environments as the foursquare logarea variable shows.

Main trip purposes in this model are grouped together for the first category for School-related activity, visit friends/family/relatives, vacation/sightseeing, outdoor recreation (sports, fishing, hunting, camping, boating, etc.), entertainment (theater, concert, sports event, gambling, etc.). The estimated model for this group shows the propensity of engaging in this type of activity in the main trip is positively correlated with secondary

trips for vacation and sightseeing. In addition, tours with a higher number of miles are also having a higher propensity to be done for this type of purposes. Central city dwellers and of higher incomes are more likely to engage in this type of tour purpose. Interestingly households with more daily trips are also more likely to have this type of tour purpose. The positive coefficient of the variable logarea indicates this type of purposes is associated with less dense destinations.

The main trip purposes in the second category are for business (work-related meeting/convention/seminar) and for combined business and pleasure. The propensity of tours having this purpose in the main trip is negatively correlated with vacation/sightseeing in the secondary trip. Households do not seem to combine these purposes. The rest of the variables show many similarities with the previous propensity. However, the logarea is not significantly different than zero indicating tours of this type are at low density and high density destinations.

The last block is the analysis of the overnight stays. Recall the first category corresponds to tours that are longer than 7 nights, the second category with tours that are between 2 and 6 overnight stays, the third is for one overnight stay, and the reference category is same day long distance travel. The negative coefficients for the secondary trip purpose work related business shows tours that contains this type of purpose are more likely to be without overnight stays. The same happens for shopping. In contrast, secondary trips that combine business and pleasure, vacation/sightseeing, and outdoor recreation are more likely to have at least an overnight stay. Vacation/sightseeing is also the purpose with the largest coefficients (and therefore probability) of being in the many overnight stays. This is exactly as expected from other sources of information. California destinations are more likely to be in the same day travel. Longer overnight stays are more likely to be in the earlier parts of the year (however this needs further scrutiny).

The household characteristics show an interesting pattern. Only household income is positively associated with the propensity of longer than a week tours. Household income and the number of employed persons are positively associated with the 2 to 6 day tours

and single night tours. Household size is negatively associated with overnight stays indicating constraints in the ability of households to spend the night outside home. Rural households are less likely to have tours with overnight stays.

Latent Class Cluster Analysis

The purpose of this model is to identify categories of tours with similar characteristics and see what types of people make these tours and what types of destinations attract them (Appendix J contains the definition of this type of statistical analysis of data). This model investigates the distribution of a set of long distance tour characteristics (namely distance, purpose, duration, and destination region) and identifies five types of long distance tours. Active covariates are characteristics of households that help predict what types of tours they make. After the model has converged, we also extract mean values for each class for other household characteristics as well as some destination characteristics.

We start by choosing a set of variables to be clustered and run the latent class clustering process over a range of class numbers. Each additional class substantially improves the model's likelihood of (re)producing the distributions of and relationships between the observed indicator variables but these improvements are counterbalanced and ultimately overwhelmed by increases to the number of parameters being estimated and a decrease in the model's ability to see clean breaks between the clusters. Because we want to produce a parsimonious model that is relatively easy to interpret, we select a model with a relatively small number of classes and a low classification error that represents a substantial improvement in log likelihood and AIC/BIC over the base models (see also Appendix J). Once we have a model that works well enough for the indicator variables, we add active covariates one at a time to improve the model's ability to place observed tours into classes. Fit statistics and classification errors for models with our final specification and 1-8 clusters are shown in Table 7.8.

Table 7.8 Different latent class cluster models and their performance

	LL	BIC(LL)	PARAMETERS	CLASSIFICATION ERROR
1-CLUSTER	-227363	454941	22	0.0000
2-CLUSTER	-213929	428403	56	0.0469
3-CLUSTER	-209021	418918	90	0.0863
4-CLUSTER	-205625	412459	124	0.0737
5-CLUSTER	**-204569**	**410677**	**158**	**0.0844**
6-CLUSTER	-204104	410080	192	0.0936
7-CLUSTER	-202920	408043	226	0.1417
8-CLUSTER	-202046	406626	260	0.1379

We select a model with five latent classes, 4 sets of indicator variables, and 4 sets of active covariates. Models with more classes had better likelihood and BIC scores, but also more substantial classification errors and were less clear to interpret. The model finds a clear distinction between long distance tours made by car and those made by other modes, and makes other clear breaks based on primary purpose and duration. Shorter car tours for personal business and other purposes are separated from longer tours, which frequently involve air travel. Another key distinction is made between long business trips (e.g., flying to New York for a meeting) and vacation/multipurpose trips, which may involve more modes. Households with high incomes who live in urban areas are responsible for a much larger share of business trips, whereas rural residents are responsible for more of the single-day trips (likely since they must travel longer distances to access a wide range of opportunities). Income is also a key factor in distinguishing between general purpose short trips and longer trips made for fun (vacations, entertainment trips, and outdoor activities). Day of the week is another primary distinguishing factor for predicting class membership. Household size, number of children, and number of trips in the daily diary were not strongly associated with any of the trip classes. Trip classes identified are as follows with their respective summaries.

1. **Day trips: 37.7% of tours**

The largest long-distance tour class is made up of mostly relatively short single-day tours made mostly by car. These tours are made by a wide range of households for a wide

range of purposes that range from necessary tasks like household replenishment / personal business, and medical care, to more recreational purposes like visiting friends and family, entertainment. Since relatively few of these tours are work-related, most are made late in the week and on the weekend. Since these tours are short, almost all visited destinations are within California. City-dwellers are somewhat less likely to make this sort of tour (note: all density categories have roughly the same population within the CHTS sample), but in all other ways, households who make this sort of tour are not notably distinct from the rest of the dataset. Destinations visited by these tours are similar to those visited by tours in the other general purpose class (3) and slightly higher than those visited by vacation classes 2 and 5.

2. Long weekends: 31.8% of tours

The next-largest class of tours is made up of short recreational trips made mostly by car. Unlike tours in class 1, these tend to feature a small number of overnight stays, but like class 1, most of these tours remain in California (though somewhat more of them visit neighboring states). A very large share of these tours start on Fridays. These tours represent a mix of purposes, but vacation and visiting friends/family are notably popular. Unsurprisingly, these optional trips are made by households that skew wealthier and suburban (like long vacation class 5). Outdoor recreation is a somewhat common purpose for these tours, and their destinations are generally slightly lower density.

3. Passenger trips: 13.0% of tours

This latent class classifying method separates passenger trips from all other categories. In general, these tours remain in California, and they are like tours in class 1. These tours are somewhat more likely to feature group activities (e.g., outdoor recreation) and trips for medical purposes (which may require someone else to drive).

In contrast to tour classes 1-3 that are predominantly made by car, the last two classes contain nearly all the trips by air in our dataset.

4. **Business trips: 9.6% of tours**

Class 4 is in some ways the most clearly defined of the classes, in terms of mode mix, purpose and household characteristics, since it overwhelmingly corresponds to trips by air outside of California for business purposes with a moderate number of overnight stays. These tours are generally starting on weekdays, but have a much more mixed selection of start dates than the other tours. These tours are primarily made by wealthy urban households and are made to destinations that are notably higher in density than those visited by other tour types.

5. **Vacations 8.1%**

The final class of long distance tours contains most of the longer-duration tours in our dataset, and more than half last at least 7 overnights. These tours can generally be categorized as long vacations outside of California. Unlike the other tour classes, these tours are made by a mix of modes (plane, car, and transit). Vacation and visiting friends and family are the most common primary purposes of these tours. Since these trips are likely made for personal enjoyment, they tend to skew wealthier like the long weekend trips in class 2, but are much less uniformly made by wealthy households than class 4. The people who make these tours are more likely to live in cities, but much less so than class 4. These tours visit destinations that are slightly denser than in classes 1-3, but much less dense than those in class 4.

Possible alternative specifications

This model is not the only useful way of clustering this dataset. Merging the driving and passenger trips is useful, but it greatly decreased the model's ability to see clear class distinctions, which suggests that the reported passenger trips have different characteristics from driving trips. Instead of basing the model on primary purpose (the purpose for the stop with the longest duration), we also tried a model that treated purpose as a set of overlapping (rather than mutually exclusive) options, with each tour including the purpose of each of its legs. It might also be useful to classify households by their long-distance travel totals instead of tour characteristics, but the incompleteness of the 8-week long distance log may limit the usefulness of this model.

Characteristics of the primary trip-maker would likely be useful in clustering, but this data is not consistently present in the log, and it is less clear how to incorporate the characteristics of other trip makers (who are not identified in the eight-week log). Number of people on the tour could also be highly useful, but recording quality of this variable is very inconsistent in the dataset.

Table 7.9a The five cluster solution (cluster indicators)

	Cluster Number	1	2	3	4	5
Indicator Variables	***Travel Distance***					
	Car Driver	159.0	446.4	3.7	21.1	928.2
	Car Passenger	3.3	8.4	221.2	6.3	258.0
	Ground Transit	0.6	1.9	33.7	5.8	164.8
	Air	2.7	12.8	24.0	**2,995.7**	**2,454.7**
	Tour Duration					
	Single-day	**73.7%**	9.5%	45.3%	3.6%	4.1%
	One overnight	13.3%	16.9%	13.3%	9.2%	2.0%
	Two-six overnights	7.2%	62.1%	33.4%	**59.8%**	37.5%
	Seven+ overnights	5.8%	11.5%	8.0%	27.5%	**56.4%**
	Destination in California	96.7%	87.2%	93.5%	17.3%	11.4%
	Primary Purpose					
	Business (meeting/convention/seminar)	13.9%	9.4%	9.2%	**33.4%**	8.8%
	Combined business and pleasure	2.1%	3.3%	1.7%	4.0%	2.8%
	School-related activity	2.4%	1.3%	5.2%	1.5%	1.8%
	Visit friends/family/relatives	30.6%	38.1%	26.7%	36.6%	35.9%
	Medical	**8.2%**	2.3%	**5.9%**	0.1%	0.7%
	Personal Business (e.g. Shopping)	**11.1%**	3.4%	**7.1%**	1.6%	2.7%
	Vacation/sightseeing	7.6%	**25.0%**	16.6%	17.4%	**38.9%**
	Outdoor recreation	7.3%	9.2%	13.0%	1.3%	3.2%
	Entertainment	**10.6%**	5.3%	**12.4%**	2.2%	2.7%
	Drive someone else / DK / RF	6.3%	2.8%	2.3%	1.9%	2.6%

Table 7.9b The five cluster solution (active and inactive covariates)

Active Covariates	**Day of the Week of Tour Start**					
	Mon	9.7%	9.5%	9.1%	14.5%	12.3%
	Tue	11.3%	8.2%	8.5%	14.0%	11.9%
	Wed	10.7%	10.4%	10.7%	17.1%	14.1%
	Thu	11.0%	14.9%	11.6%	17.4%	16.4%
	Fri	13.3%	**30.4%**	**22.8%**	15.0%	15.5%
	Sat	**27.5%**	15.9%	**23.0%**	10.6%	16.9%
	Sun	16.6%	10.8%	14.3%	11.3%	12.9%
	Household Trips in Daily Diary	10.1	10.1	11.5	11.5	9.5
	Household Annual Income					
	Under $50,000	23.5%	17.4%	27.3%	8.5%	19.3%
	$50,000 to $74,999	20.1%	18.6%	17.8%	10.8%	16.9%
	$75,000 to $99,999	19.1%	19.9%	17.4%	12.6%	16.6%
	$100,000 to $149,000	20.8%	23.8%	20.7%	25.0%	23.0%
	$150,000 and above	16.6%	20.4%	16.9%	43.1%	24.3%
	Household Block Group Category					
	Urban Center	19.8%	22.3%	23.1%	44.3%	27.7%
	High Density Suburb / Exurb	27.5%	27.1%	27.4%	29.5%	26.8%
	Low Density Suburb / Exurb or High Density Rural	26.1%	25.0%	25.1%	20.7%	25.1%
	Low Density Rural	26.6%	25.7%	24.4%	5.4%	20.3%
Inactive Covariates	**Household Size**	2.6	2.5	2.7	2.5	2.5
	Children in Household	0.5	0.5	0.6	0.5	0.4
	Household is Hispanic	16.5%	13.8%	16.2%	9.3%	12.3%
	Employment Density Around Household (emp/km2)	934	1,013	1,068	2,176	1,230
	Destination Characteristics (Foursquare)					
	Median Checkins	1016	966	1069	1477	969
	Median User Count	468	463	509	720	465
	Convex Hull Area of 50 nearest POIs (log km2)	2.74	2.80	2.65	1.97	2.51

Summary of Findings

In this analysis of long distance travel in California we found systematic differences among persons and households in all aspects analyzed. In earlier sections of the report and particularly in Section 4 we show the systematic self-selection biases in the long distance reporting of trips. We also account for these self-selection biases in the synthetic population generation and demonstrate that daily diary trip making offers a good representation of diversity in trip making in California. In terms of long distance, travel differences among persons and households are mainly due to social and demographic characteristics of households with primary driver the household wealth and employment. Place of residence plays a major role in explaining long distance travel and this shows a more detailed analysis of opportunities for activities around the place of residence would inform long distance VMT contribution in a substantial way. We also found attractiveness of destinations playing a major role that is captured in this analysis using social media data. This finding offers encouragement for subsequent studies to gather information about destinations regarding their attractiveness defined not only about the specific destination but also neighboring places. We also found, using a variety of data analytic methods, a need to perform grouping of trips in tours is an efficient and insightful way in studying destination. In fact, the SEM models, path models, and latent class clustering showed clearly social and demographic variables play different roles. They also demonstrate different ways one can employ to reveal different aspects of travel behavior. We also show in this report with an example in synthetic population one way of accounting for self-selection in reporting biases and create maps of long distance travel behavior. However, additional analysis and the development of a bias correcting algorithm is needed to provide statewide estimates of long distance travel by different modes that includes trips within the State and elsewhere. Very important for future studies are also the biases found in the long distance 8-week travel log and the substantial amount of missing information. In contrast, the daily diary contains details that are needed in examining trips within tours. Unfortunately the decision to design a single day diary meant missing trips with overnight stays away from home. The clear recommendation from our analysis is to design activity diaries that span multiple-days of

complete households and a satellite survey that has diaries for an 8-week travel log that has added information about travel during the 8-week period and the people with whom travel happens. In spite of all these biases, however, the latent class cluster analysis is able to differentiate among five distinguishable types of tours in a clear way thus enabling the development of a parsimonious set of types of long distance travel that can be used in subsequent modeling.

There are many nest steps for subsequent projects. In our analysis we found a significant relationship between daily travel and long distance travel. To make the analysis here tractable within the timeline of the project we limited the study to the total number of trips in the daily diary by the households that reported complete tours in the 8-week long distance log. This should be expanded to include other travel behavior variables (mode used, destinations visited, activity types, and miles traveled).

In our tour based analysis we selected the household as the unit of analysis and within each household the long distance tour with additional information about the trip purposes of trips within each tour. We envision a continuation of the study here that examines mode choice for each trip, within each tour by each person in a household that also accounts for both person and household characteristics but also individual trip destination attractiveness. In addition, human interaction within the households may play a major role in decision making about long distance travel and this aspect was not included with specific questions in CHTS. To address this we envision a study that asks how decisions about long distance travel come about within households. Of particular interest is examining time allocation to different activities during a long distance tour by different members of the household. A study of this type can be done in at least two different ways. The first is a longitudinal study (panel survey in which the same households are interviewed repeatedly) in which the participants are asked at four different times of the year to report their long distance trips for all household members and respond to added more in-depth questions. Moreover a stated choice experiment can also be created to also examine the impact of different long distance attributes (e.g., cost, time, timing,

environmental impact, and logistics arrangements) on decision making of persons and households.

Bibliography

Axhausen, K. W. (2001). Methodological research for a European survey of long-distance travel. In *Personal Travel: the Long and Short of It. Conference Proceedings June 28–July 1, 1999 Washington, DC* (pp. 321-342).

Bar-Gera, H., K. Konduri, B. Sana, X. Ye, and R.M. Pendyala (2009) Estimating Survey Weights with Multiple Constraints Using Entropy Optimization Methods. *Proceedings of 88th Annual Meeting of the Transportation Research Board,* National Research Council, Washington, D.C.

Beckman, J. D., & Goulias, K. G. (2008). Immigration, residential location, car ownership, and commuting behavior: a multivariate latent class analysis from California. *Transportation*, *35*(5), 655-671.

Bierce, E., & Kurth, D. (2014). The Use of Three Surveys for Long Distance Travel Estimates in California. In *TRB 93rd Annual Meeting Compendium of Papers* (pp. 14-5674).

Broeg W., E. Erl, G. Sammer, and B. Schulze (2003) DATELINE - Design and Application of a Travel Survey for Long-distance Trips Based on an International Network of Expertise – Concept and Methodology. Paper presented at the International Conference on Travel Behaviour Research Lucerne, 10-14. August 2003

CALTRANS (2015) Interregional Transportation Strategic Plan: Update 2015. http://www.caltrans-itsp2015.org. Accessed September 22, 2015.

Cambridge Systematics Inc. (2014) California Statewide Travel Demand Model, Version 2.0 Long Distance Personal Travel Model. Final Report. http://www.dot.ca.gov/hq/tpp/offices/omsp/statewide_modeling/Files/Documentation/FR1_CSTDMV2_Long_Distance_Personal_Travel_Model.pdf. Accessed September 22, 2015.

Chapman, L. (2007). Transport and climate change: a review. *Journal of transport geography*, *15*(5), 354-367.

de Abreu e Silva, J., Golob, T., & Goulias, K. (2006). Effects of land use characteristics on residence and employment location and travel behavior of urban adult workers. *Transportation Research Record: Journal of the Transportation Research Board*, (1977), 121-131.

de Abreu E Silva, J., Goulias, K., & Dalal, P. (2012). Structural Equations Model of Land Use Patterns, Location Choice, and Travel Behavior in Southern California. *Transportation Research Record: Journal of the Transportation Research Board*, (2323), 35-45.

Georggi, N. L., & Pendyala, R. M. (2000). *An analysis of long-distance travel behavior of the elderly and the low-income* Paper Presented at the Conference the Long and short. Transportation Research Circular E-C026—Personal Travel: The Long and Short of It, Transportation Research Board, Washington D.C.

Golob, T.F. (2003) Structural Equation Modeling. In *Transportation Systems Planning: Methods and Applications*. (ed. K.G. Goulias). CRC Press, Boca Raton, FL, pp 11.1-11.23.

Holz-Rau, C., Scheiner, J., & Sicks, K. (2014). Travel distances in daily travel and long-distance travel: what role is played by urban form?. *Environment and Planning A*, *46*(2), 488-507.

Konduri, K.C., D. You, V.M. Garikapati, and R.M. Pendyala (2016) Application of an Enhanced Population Synthesis Model that Accommodates Controls at Multiple Geographic Resolutions. *Transportation Research Record, Journal of the Transportation Research Board* (forthcoming).

LaMondia, J. J., & Bhat, C. R. (2011). A Conceptual and Methodological Framework of Daily and Long Distance Leisure Activity-Travel Behavior. In *Transportation Research Board 90th Annual Meeting* (No. 11-2867).

MARG (2016) *PopGen: Synthetic Population Generator* [online]. Mobility Analytics Research Group. Available at:http://www.mobilityanalytics.org/popgen.html. Accessed May 2017.

McBride E., A.W. Davis, J.H. Lee, and K.G. Goulias (2017) Incorporating Land Use in Synthetic Population Generation Methods and Transfer of Behavioral Data. *Transportation Research Record: Journal of the Transportation Research Board - in press.*

McCutcheon, A.L. (2002). Basic concepts and procedures in single- and multiple-group latent class analysis. In: *Applied Latent Class Analysis* (J.A. Hagenaars and A.L. McCutcheon, eds.), pp. 56-88. Cambridge University Press, Cambridge.

Mitra S. K. (2016) Land Use, Land Value, and Transportation: Essays on Accessibility, Carless Households, and Long-distance Travel. Ph.D. Dissertation, UC Irvine (Accessed from ProQuest).

NUSTATS (2013) 2010-2012 California Household Travel Survey Final Report, Austin, TX

Shahrin, N., Som, A. P. M., & Jusoh, J. (2014). Long Journey Travel to Tourist Destination: A Review Paper. In *SHS Web of Conferences* (Vol. 12, p. 01099). EDP Sciences.

Vermunt, J.K. and J. Magidson (2002). Latent class cluster analysis. In: *Applied Latent Class Analysis* (J.A. Hagenaars and A.L. McCutcheon, eds.), pp. 89-106. Cambridge University Press, Cambridge.

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

Appendix A The Long Distance Log in CHTS

CALIFORNIA
Household Travel Survey

Long-Distance TRAVEL LOG

Last Name: _____
Travel Day: _____
Travel Period*: _____
PIN#: _____

*Note: Your Long-Distance Travel Period is the eight weeks prior to your Travel Day.

Name of person completing this log: _____

Your person number: *(Person #s are on the Travel Diary label)*

○ P1 ○ P2 ○ P3 ○ P4 ○ P5 ○ P6 ○ P7 ○ P8

○ No one in my household made a long-distance trip in the eight weeks prior to our travel day.
If this is the case, please fill in the bubble above and return this Log with your completed Diaries.

INSTRUCTIONS

- Record details about all long-distance trips made by any household member during the travel period shown on the label.
- A long-distance trip is a trip made to a location 50 miles away or more from your home.
- Record each way (away from home and returning home) as a separate trip.
- If you made more than 8 long-distance trips, please record the details on a separate piece of paper.

How do I provide my Long-Distance Travel Log information?

- Online: Enter your information at www.catravelsurvey.com. Use PIN# on the label.
  OR
- Mail: Return with your completed travel diaries.
  OR
- Phone: We will call you to collect your Log and Travel Diary information. Or, you can call us at the toll free hotline number below.

Questions? Call the toll-free hotline at 1-877-261-4621

Lists A and B are on the back!

Trip Departure DATE *(Locations 50 miles away or more)*	WHERE were you when you STARTED this trip?	WHERE did you travel TO? *(Your final destination)*	MAIN PURPOSE of trip *Use LIST A CODES*	HOW MANY OTHER PEOPLE were traveling with you? *(Excluding yourself)*	What METHOD OF TRAVEL was used for the longest distance? *Use LIST B CODES*
Trip 1: Most Recent Date: ___/___/___ mo day yr	Place Name: Address or Nearest Cross-streets: City: State/ZIP/Country:	Place Name: Address or Nearest Cross-streets: City: State/ZIP/Country:	List ONE code only	# of people traveling with you (excluding yourself): ___ # of household members (excluding yourself): ___ Which household members traveled? *(use person #s from diary label)* ○ P1 ○ P2 ○ P3 ○ P4 ○ P5 ○ P6 ○ P7 ○ P8	List ONE code only **Remember to record EACH WAY as a separate trip**
Trip 2 Date: ___/___/___ mo day yr	Place Name: Address or Nearest Cross-streets: City: State/ZIP/Country:	Place Name: Address or Nearest Cross-streets: City: State/ZIP/Country:	List ONE code only	# of people traveling with you (excluding yourself): ___ # of household members (excluding yourself): ___ Which household members traveled? *(use person #s from diary label)* ○ P1 ○ P2 ○ P3 ○ P4 ○ P5 ○ P6 ○ P7 ○ P8	List ONE code only
Trip 3 Date: ___/___/___ mo day yr	Place Name: Address or Nearest Cross-streets: City: State/ZIP/Country:	Place Name: Address or Nearest Cross-streets: City: State/ZIP/Country:	List ONE code only	# of people traveling with you (excluding yourself): ___ # of household members (excluding yourself): ___ Which household members traveled? *(use person #s from diary label)* ○ P1 ○ P2 ○ P3 ○ P4 ○ P5 ○ P6 ○ P7 ○ P8	List ONE code only
Trip 4 Date: ___/___/___ mo day yr	Place Name: Address or Nearest Cross-streets: City: State/ZIP/Country:	Place Name: Address or Nearest Cross-streets: City: State/ZIP/Country:	List ONE code only	# of people traveling with you (excluding yourself): ___ # of household members (excluding yourself): ___ Which household members traveled? *(use person #s from diary label)* ○ P1 ○ P2 ○ P3 ○ P4 ○ P5 ○ P6 ○ P7 ○ P8	List ONE code only

LIST A CODES - TRIP PURPOSE

1 Going to work
2 Business *(work-related meeting / convention / seminar)*
3 Combined business and pleasure
4 School-related activity
5 Visit friends / relatives / wedding / funeral
6 Medical
7 Vacation / Sightseeing
8 Outdoor recreation *(sports, fishing, hunting, camping, boating, etc.)*
9 Entertainment *(theater, concert, sports event, gambling, etc.)*
10 Personal business *(e.g., shopping)*
11 Drive someone else
12 Return home
97 Other *(write code 97 and specify)*

LIST B CODES - METHOD OF TRAVEL

NON-MOTORIZED:
1 Walk
2 Bike
3 Wheelchair / Mobility Scooter
4 Other Non-Motorized *(skateboard, etc.)*

PRIVATE VEHICLE:
5 Auto / Van / Truck Driver
6 Auto / Van / Truck Passenger
7 Carpool / Vanpool
8 Motorcycle / Scooter / Moped

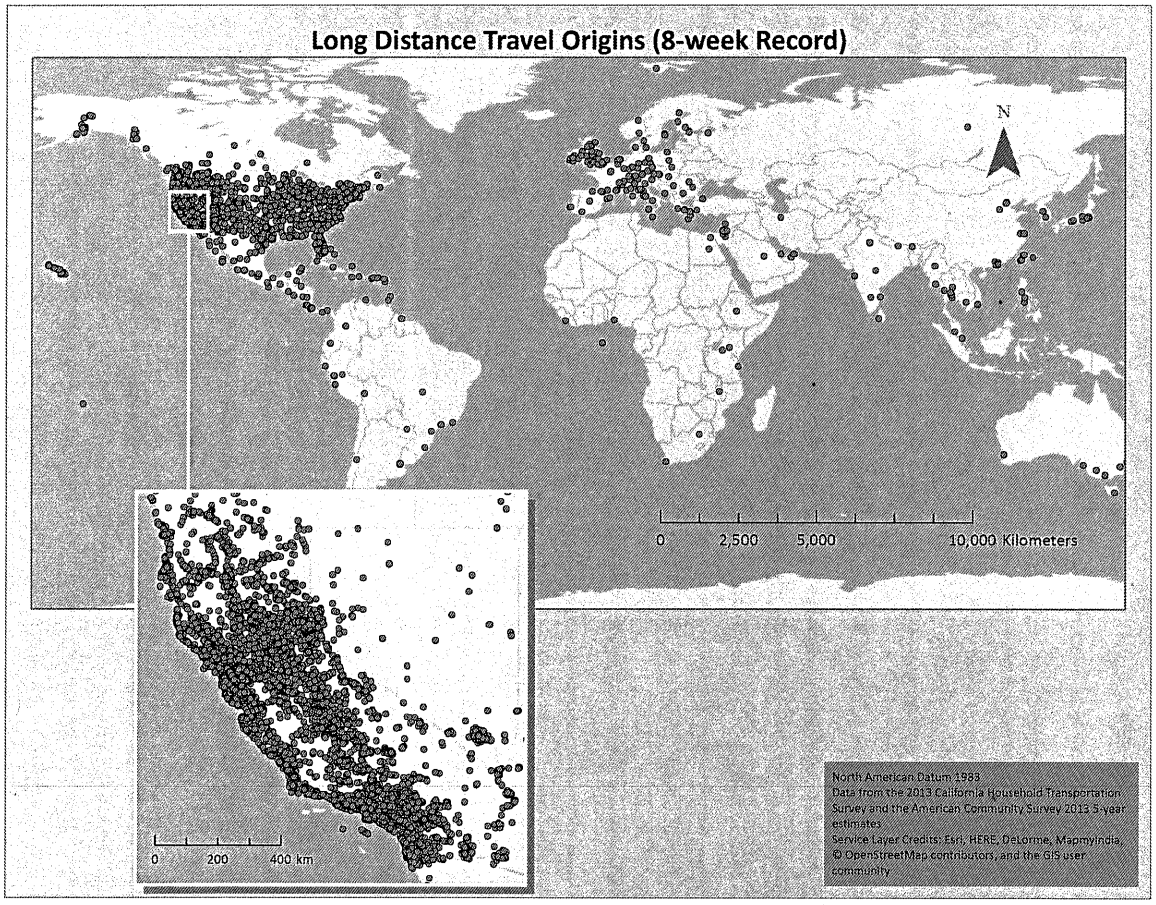
PRIVATE TRANSIT:
9 Taxi / Hired Car / Limo
10 Rental Car / Vehicle
11 Private Shuttle *(SuperShuttle, employer, hotel, etc.)*
12 Greyhound Bus
13 Airplane
14 Other Private Transit

PUBLIC TRANSIT:
Bus
15 Local Bus, Rapid Bus
16 Express Bus / Commuter Bus *(AC Transbay, Golden Gate Transit, etc.)*
17 Premium Bus *(Metro Orange / Silver Line)*
18 School Bus
19 Public Transit Shuttle *(DASH, Emery Go-Round, etc.)*
20 AirBART / LAX FlyAway
21 Dial-A-Ride / ParaTransit *(Access Services, etc.)*
22 Amtrak Bus
23 Other Bus
Rail / Subway
24 BART, Metro Red / Purple Line
25 ACE, Amtrak, Caltrain, Coaster, Metrolink
26 Metro Blue / Green / Gold Line, Muni Metro, Sacramento Light Rail, San Diego Sprinter / Trolley / Orange / Blue / Green, VTA Light Rail
27 Street Car / Cable Car
28 Other Rail
Ferry
29 Ferry / Boat

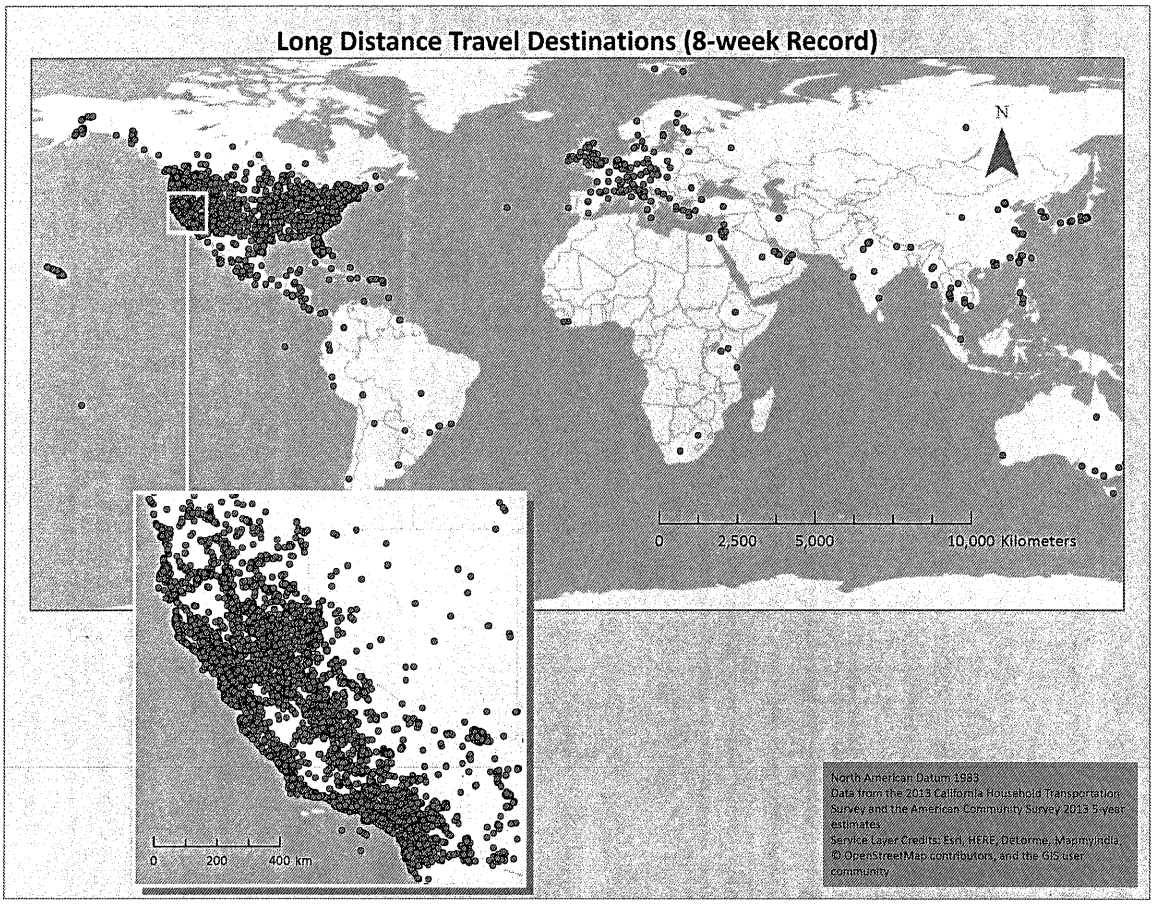
Appendix B Trip Purposes in the 8-week Travel Log 66828 Trips

Purpose	Number of T
Going to work	2745
Business (work-related meeting	6116
Combined business and pleasur	1111
School-related activity	912
Visit friends/family/relatives	12911
Medical	1847
Vacation/sightseeing	7586
Outdoor recreation (sports, fish	2732
Entertainment (theater, concer	2902
Personal Business (e.g. shoppin	2742
Driving someone else	1201
Return Home	23001
Other (specify)	745
Don't know	197
Refuse	80
Total	66828

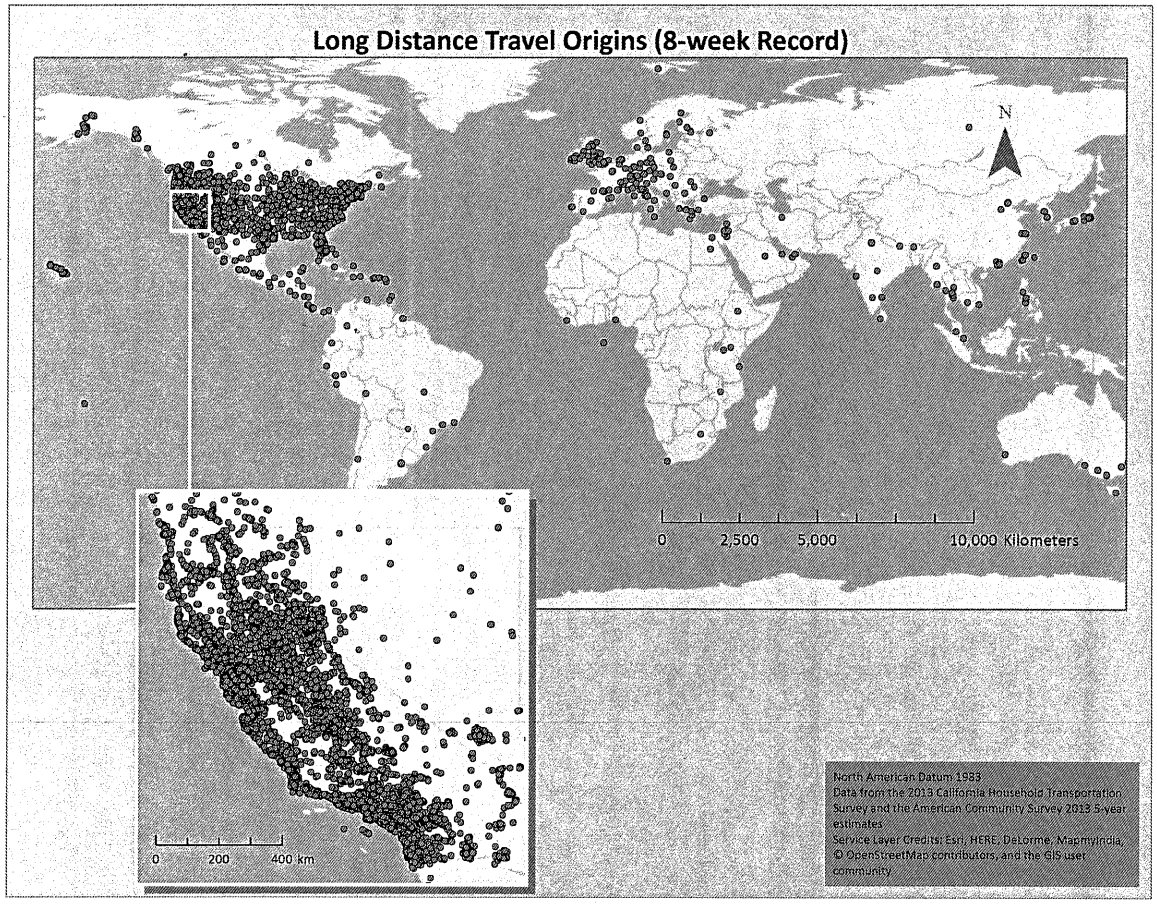
Appendix C Long Distance Travel Origins in Place Diary



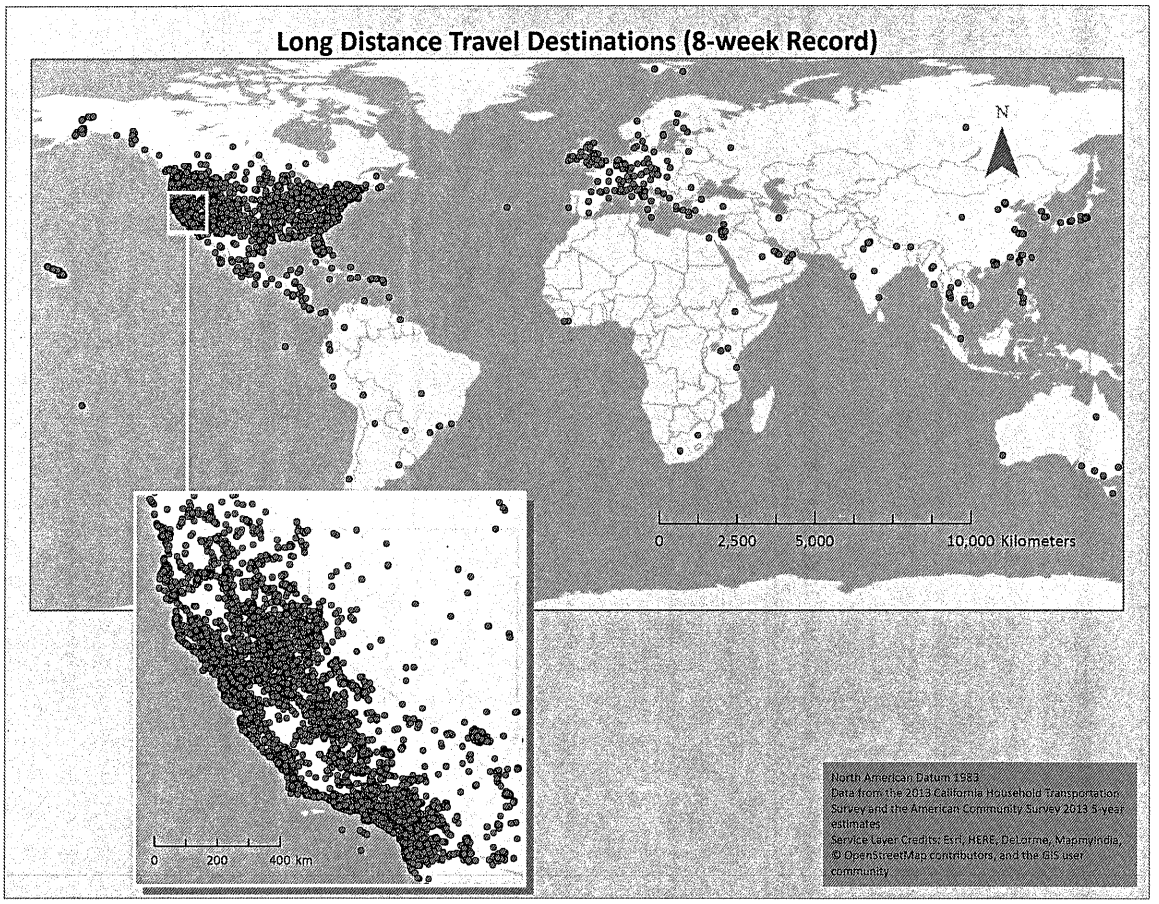
Appendix D Long Distance Travel Destinations in Place Diary



Appendix E Long Distance Travel Origins in 8-week Log



Appendix F Long Distance Travel Destinations in 8-week Log



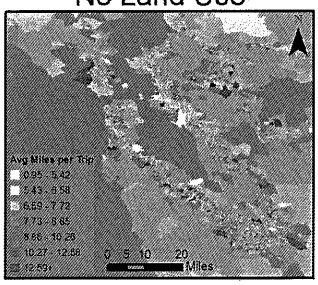
Appendix G Household file with all CHTS households

Statistic	N	Mean	St.Dev	Min	Median	Max
SAMPN	42,431	2,588,379.00	1,641,345.00	1,031,985	1,971,814	7,212,388
ILANG	42,431	1.06	0.24	1	1	2
CTFIP	42,431	6,056.12	29.5	6,001	6,059	6,115
RIBUS	42,431	1.8	0.53	1	2	9
HHVEH	42,431	1.86	1	0	2	8
HHBIC	42,431	1.58	3.79	0	1	99
VEHNEW	42,431	2.15	2.02	1	2	9
RESTY	42,431	1.91	4.16	1	1	99
OWN	42,431	1.24	0.56	1	1	9
INCOM	42,431	13.18	26.29	1	5	99
HHSIZ	42,431	2.57	1.37	1	2	8
HHEMP	42,431	1.22	0.88	0	1	6
HHSTU	42,431	0.64	1.02	0	0	8
HHLIC	42,431	1.86	0.85	0	2	8
DOW	42,431	4.02	1.99	1	4	7
HTRIPS	42,431	8.29	7.78	0	6	99
HPFlag	42,431	1.76	0.43	1	2	2
hhTours	42,431	0.97	1.82	0	0	49
hhTrips	42,431	1.61	3.06	0	0	50
onlyCA	42,431	0.28	0.45	0	0	1
onlyUSA	42,431	0.4	0.49	0	0	1
hhtWkES	41,546	0.33	0.78	0	0	16
hhtWkDS	41,546	0.62	1.39	0	0	45
hhtSngl	42,431	0.43	1.18	0	0	48
hhtPart	42,431	0.12	0.53	0	0	22
hhtGood	42,431	0.42	1.06	0	0	25
hhtOneD	41,546	0.62	1.42	0	0	47
hhtOneN	41,546	0.06	0.32	0	0	15
hhtWeek	41,546	0.16	0.52	0	0	15
hhtLong	41,546	0.1	0.36	0	0	6
noLDT	42,431	0.58	0.49	0	1	1

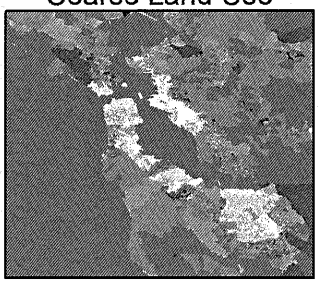
Appendix H Comparison of Population Synthesis Methods in Three Metropolitan Areas in California

Average Miles Per Trip: Bay Area

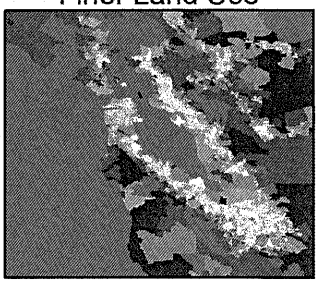
No Land Use



Coarse Land Use

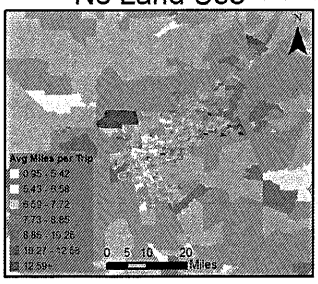


Finer Land Use

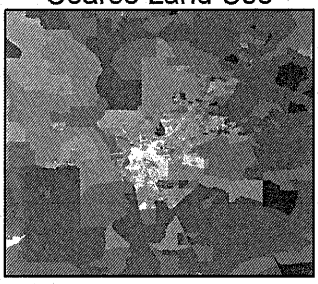


Average Miles Per Trip: Sacramento Area

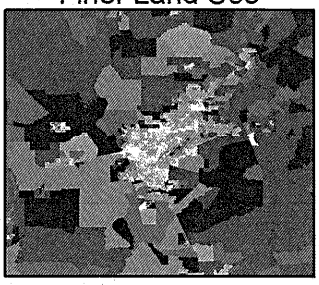
No Land Use



Coarse Land Use

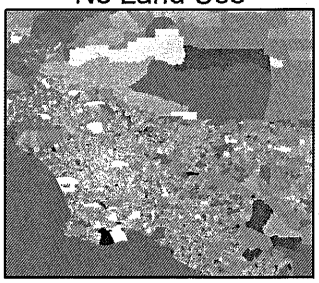


Finer Land Use

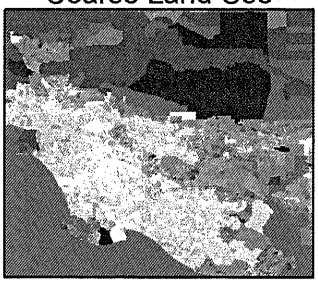


Los Angeles Area

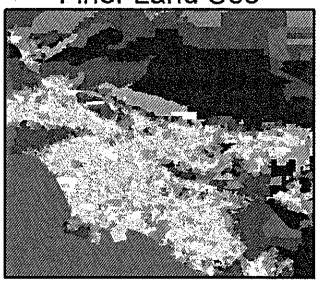
No Land Use



Coarse Land Use



Finer Land Use



Appendix I Structural Equations Models

The general Structural Equations Model (SEM) with latent variables consists of two parts: 1) measurement model and 2) structural model. The measurement model specifies how latent variables determine the observed dependent variables, while the structural model specifies the causal relationships among the latent variables and describes the causal effects of the exogenous variables on the endogenous dependent variables. The measurement model can be further classified into the measurement model for the endogenous variables (y) and the measurement model for the exogenous variables (x). The matrix formulation of the general SEM with latent variables is defined as follows:

Measurement model for y: $y = \Lambda_y \eta + \varepsilon$ (Eq I.1)

Measurement model for x: $x = \Lambda_x \xi + \delta$ (Eq I.2)

Structural model: $\eta = \mathrm{B}\eta + \Gamma\xi + \zeta$ (Eq. I.3)

where $y = p \times 1$ vector of observed endogenous variables.

$x = q \times 1$ vector of observed exogenous variables.

$\eta = m \times 1$ vector of latent endogenous variables.

$\xi = n \times 1$ vector of latent exogenous variables.

$\varepsilon = p \times 1$ vector of measurement errors in y.

$\delta = q \times 1$ vector of measurement errors in x.

$\Lambda_y = p \times m$ matrix of coefficients of the regression of y on η.

$\Lambda_x = q \times n$ matrix of coefficients of the regression of x on ξ.

$\mathrm{B} = m \times m$ matrix of coefficients of the η-variables in the structural relationships.

$\Gamma = m \times n$ matrix of coefficients of the ξ-variables in the structural relationships.

$\zeta = m \times 1$ vector of equation errors in the structural relationships.

Given the complexity and operational difficulties in estimation of a full SEM, it is rarely found in practice (Golob, 2003). However, in this report we use different types of SEM

to better understand the correlation among the many different types of relationship of long distnace travel choices. A SEM with latent variables allow us to represent different apsects of travel and destination attractiveness is well defined groups and we can study their correlation with travel behavior facets. This complicates the analysis and creates a variety of numerical issues. To simplify the study and to obtain a clearer model to understand tradeoffs households make in deciding how to travel we also show a Path analysis in which a SEM is estimated with observed variables only. This is in essence a set of regression equations in which some endogenous variables are entered as determinants of other endogenous variables. Since no latent variables are involved in the SEM, the measurement models (the ones containing latent variables) for x and y are dropped. Structural equations models with observed variables are therefore reduced to the following form:

$$y = \mathrm{B}y + \Gamma x + \zeta \qquad \text{(Eq. I.4)}$$

where $y = p \times 1$ vector of observed endogenous variables.

$x = q \times 1$ vector of observed exogenous variables.

$\mathrm{B} = p \times p$ matrix of coefficients of the y-variables.

$\Gamma = p \times q$ matrix of coefficients of the x-variables.

$\zeta = p \times 1$ vector of equation errors.

In the SEM with observed variables, y and x are assumed to exactly represent the latent η and ξ, respectively. So the number of y variables equals the number of η variables (p=m) and the number of x variables equals the number of ξ variables (q=n).

SEM is a covariance–based model, because structural equations systems are estimated by covariance analysis. In the procedure, the difference between the sample covariances and the covariances predicted by the model is minimized, instead of minimizing the difference between observed and predicted individual values. The underlying theory of this estimation procedure is that the population covariance matrix of the observed variables (Σ) is a function of a set of parameters:

$$\Sigma = \Sigma(\theta) = \begin{bmatrix} \text{covariance matrix of y} & \text{covariance matrix of y and x} \\ \text{covariance matrix of x and y} & \text{covariance matrix of x} \end{bmatrix}$$

$$= \begin{bmatrix} (I-\mathrm{B})^{-1}(\Gamma\Phi\Gamma'+\Psi)[(I-\mathrm{B})^{-1}]' & (I-\mathrm{B})^{-1}\Gamma\Phi \\ \Phi\Gamma'[(I-\mathrm{B})^{-1}]' & \Phi \end{bmatrix} \qquad \text{(Eq. I.5)}$$

where Φ = covariance matrix of x.

Ψ = covariance matrix of ζ.

The matrix $\Sigma(\theta)$ consists of three matrices. The unkown parameters $\mathrm{B}, \Gamma, \Phi,$ and Ψ are estimated by finding the parameters such that the covariance matrix ($\hat{\Sigma}$) implied by the model is as close as possible to the sample covariance matrix (S). To know when the estimates are as close as possible, a fitting fuction that is to be minimized is defined. All this discussion pertains to relationships among observed variables (x,y) and different ways to study how one variable relates to another.

Appendix J Latent Class Cluster Analysis

One technique selected to identify groups of homogeneous patterns of activity and travel behaviour in the CHTS long distance survey data is *latent class cluster analysis*. This technique includes a *J*-category latent variable with each category representing a cluster; uses many "dependent" or clustering variables (named *criteria* variables herein); uses a mixture of multiple types of criteria variables (e.g., continuous, categorical, ordered, count); uses and tests the effect of covariates of many different types; is more flexible than many other clustering algorithms; is a model-based clustering approach and it provides probabilistic membership of observations in clusters; and provides convenient interpretable output.

In this chapter, we use notation and model formulation similar to Vermunt and Magidson (2002). Assume there is one latent variable, X representing long distance travel. Different categories of this variable X denote different types of activity-travel behavior and the probability of belonging to each category of variable X represents the proportion of persons that choose that specific type of time allocation. Using observed data we would like to identify how many distinct groups we have and find the proportion of persons in each group. For each person in our sample we observe M measures (indicators) of activity and travel behavior indicated by the symbol Y that can be used to infer membership in the categories of the latent variable X. A third set of variables, which are not included as criteria variables in the clusters, are used as explanatory variables and for this indicated with the symbol Z. The probability density of the Ys given a set of Z values is:

$$f(Y \mid Z) = \sum_{x} \pi(X \mid Z) f(Y \mid X, Z) \qquad \text{(Eq. J.1)}$$

where $\pi(X \mid Z)$ is the probability of belonging to a certain latent class given a set of covariate values. Lower case x in the Sum symbol denotes the categories of the variable

X. If the Y variables belonging to different clusters (categories of variable X) are assumed mutually independent given the latent class and the covariates, we obtain:

$$f(Y \mid Z) = \sum_{x} \pi(X \mid Z) \prod_{m=1}^{M} f(Y_m \mid X, Z) \qquad \text{(Eq. J.2)}$$

Since the scores on the latent variable given the covariates are assumed to come from a multinomial distribution, the probability of belonging to a given latent class can be calculated as follows:

$$\pi(X \mid Z) = \frac{e^{\eta_{X|Z}}}{\sum_{X} e^{\eta_{X|Z}}} \qquad \text{(Eq. J.3)}$$

where the term h is a linear combination of the main effects of the latent variable (γ_{x_j}) and the covariate effects on the latent variable ($\gamma_{z_l x_j}$) defined as:

$$\eta_{X|Z} = \sum_{j=1}^{J} \gamma_{x_j} + \sum_{l=1}^{L} \sum_{j=1}^{J} \gamma_{z_l x_j} \qquad \text{(Eq. J.4)}$$

One way to visualize this model is to consider a cross-classification table underlying the model in which latent and observed variables are included. This table has dimensions equal to the categories of all the variables when all variables are categorical. The cell values of this table are the entities we are trying to estimate using formulations as in Equation 4. As in many latent class models the likelihood function takes the familiar form shown below where q denotes the unobserved parameters to be estimated.

$$LogL = \sum_{i} n_i \log f(Y_i \mid Z_i, \theta) \qquad \text{(Eq. J.5)}$$

The parameters in equation 5 can be estimated by the Expectation Maximization (EM) algorithm, which produces maximum likelihood estimates under specific conditions. In the examples here, we use the Vermunt and Magidson (2002) method, which is a combination of EM with Newton-Raphson. Standard errors for the parameter estimates are computed using the Hessian matrix (matrix of the second order derivatives of the estimating equation). As the number of parameters to estimate increases, the degrees of freedom decrease rapidly, resulting in a variety of operational problems such as identification (inability to compute a parameter) or lack of convergence (subsequent estimation step parameters are not close enough). Most latent class models are also sensitive to local maxima of the likelihood function used in estimation, which can be circumvented by testing multiple models using different initial trial values for the parameters. Estimation of models of this type is a hierarchical, iterative process in which we start with a one-cluster assumption and estimate a simple model. Then, experimentation proceeds by increasing the number of clusters until identification is no longer possible. For some parameters, the cluster sizes become too small to be meaningful, and/or the difference in goodness of fit between successive models is not significant. At this point, we select one or more models that appear to be a reasonable description of the observed data. We define alternate modeling options, such as correlations among criteria variables and variances within each cluster, and start another iterative cycle. This process continues until the addition of a more complex structure no longer yields a significant improvement (for nested models we can use a formal statistical likelihood-based step as a stop criterion).

Within these three steps, we also have two additional "mini-steps." For each model, we first develop starting values for the unknown parameters we are estimating that are drawn from a distribution of randomly selected moments. For a given set of starting values, we perform maximum likelihood iterations first using the EM algorithm until the values of subsequent iterations reach a predefined difference (or the total number of EM iterations reaches a maximum number). Then, the algorithm switches to a Newton-Raphson algorithm until a predetermined convergence criterion value is reached or the maximum number of iterations is reached. In this way, we can exploit advantages of both

algorithms, i.e., the stability of EM when far away from the optimum and the speed of Newton-Raphson when close to the optimum (Vermunt and Magidson, 2002).

Statistical goodness-of-fit measures for latent class cluster models are the typical chi-square statistics in the cross-categorical data analysis. The first measure is the likelihood ratio chi-square, G^2 or L^2. It has a chi-square distribution with degrees of freedom given by the number of "free" parameters (total number of different response patterns - the number of estimated model parameters – 1 if there are no covariates in the model). It represents the opposite of an R^2 in regression because it is the amount of unexplained associations among the criteria variables by the model. Therefore, higher values indicate models that do not fit the data well and lower values represent better fitting models. When two models are nested (i.e., they differ only in the number of estimated parameters), we could create the difference between the G^2 of these two models. This difference is chi-square distributed and can be used for hypotheses testing. A test of this type cannot be utilized between models that differ in the number of clusters because they are not nested. The L^2, the Bayes Information Criterion (BIC), Akaike Information Criterion (AIC) and the Consistent Akaike Information Criterion (CAIC) are computed to measure goodness of fit and to take into account model parsimony, penalizing models with many parameters. The lower the BIC, AIC or CAIC values, the better the model we estimate (McCutcheon, 2002).

There are many advantages using this method for clustering. First, the latent class cluster method for identifying clusters is designed for combinations of continuous and discrete criteria variables, while the k-means method is defined for continuous variables only. Second, the method used here allows for probabilistic membership of each observation in each cluster. This provides flexibility in observation classification that the k-means does not. Third, post-processing of the cluster data using regression is not required because the method used allows the inclusion of covariates. There are other advantages of latent class methods in general and the specific implementation used here as illustrated in Vermunt and Magidson (2002).