Title
Long Distance Travel and Destination Attractiveness

Journal
ITS Reports, 2017(44)

Authors
Goulias, Konstadinos G., PhD
Davis, Adam W.
McBride, Elizabeth C.

Publication Date
2018-06-01

DOI
10.7922/G2X34VN3

UNIVERSITY OF CALIFORNIA

Long Distance Travel and Destination Attractiveness

A Research Report from the University of California Institute of Transportation Studies

Konstadinos G. Goulias, Department of Geography, University of California Santa Barbara

Adam W. Davis, Department of Geography, University of California Santa Barbara

Elizabeth C. McBride, Department of Geography, University of California Santa Barbara

June 2018

PROJECT ID: UC-ITS-2017-44 | DOI: 10.7922/G2X34VN3

TECHNICAL REPORT DOCUMENTATION PAGE

<table>
<tr><td>1. Report No.
UC-ITS-2017-44</td><td>2. Government Accession No.</td><td>3. Recipient's Catalog No.</td></tr>
<tr><td colspan="2" rowspan="2">4. Title and Subtitle
Long Distance Travel and Destination Attractiveness</td><td>5. Report Date
June 2018</td></tr>
<tr><td>6. Performing Organization Code
UC Santa Barbara</td></tr>
<tr><td colspan="2">7. Author(s)
Konstadinos G. Goulias Ph.D. https://orcid.org/0000-0002-7436-9447, Adam W. Davis, and Elizabeth C. McBride</td><td>8. Performing Organization Report No.</td></tr>
<tr><td colspan="2" rowspan="2">9. Performing Organization Name and Address
Department of Geography, University of California, Santa Barbara,
Ellison Hall, University of California Barbara, Santa Barbara, CA 93117</td><td>10. Work Unit No.</td></tr>
<tr><td>11. Contract or Grant No.
UC-ITS-2017-44</td></tr>
<tr><td colspan="2" rowspan="2">12. Sponsoring Agency Name and Address
The University of California Institute of Transportation Studies
www.ucits.org</td><td>13. Type of Report and Period Covered
Final Report (May 2017-June 2018)</td></tr>
<tr><td>14. Sponsoring Agency Code
UC ITS</td></tr>
<tr><td colspan="3">15. Supplementary Notes
DOI:10.7922/G2X34VN3</td></tr>
<tr><td colspan="3">16. Abstract
This report provides a summary of analyses using data of long distance tours by each household from an 8-week California Household Travel Survey travel log. The first analysis, uses Structural Equations Models (SEM) and a simpler variant called Path Analysis on three censored variables (tour miles by air, miles driving, and miles by public transportation) and two categorical variables (main trip tour purpose) and number of overnight stays. The second analysis, uses Latent Class Cluster Analysis (LCCA) to identify five distinct, informative patterns of long-distance travel. This analysis shows that long-distance tours for vacation, business travel, medical, and shopping are substantially distinct in terms of their travel characteristics and correspond to different combinations of other activities in the tour and they are done by different types of households. The methods used here to identify the typology of long distance travel can be easily expanded to include a variety of other explanatory variables of this type of behavior in more focused data collection settings.</td></tr>
<tr><td colspan="2">17. Key Words
Travel behavior, travel surveys, households, recreational trips, trip length</td><td>18. Distribution Statement
No restrictions.</td></tr>
<tr><td>19. Security Classif. (of this report)
Unclassified</td><td>20. Security Classif. (of this page)
Unclassified</td><td>21. No. of Pages
58   22. Price</td></tr>
</table>

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

ABOUT THE UC ITS

The University of California Institute of Transportation Studies (ITS) is a network of faculty, research and administrative staff, and students dedicated to advancing the state of the art in transportation engineering, planning, and policy for the people of California. Established by the Legislature in 1947, ITS has branches at UC Berkeley, UC Davis, UC Irvine, and UCLA.

ACKNOWLEDGEMENTS

This study was made possible through funding received by the University of California Institute of Transportation Studies from the State of California's Public Transportation Account and the California Department of Transportation. The authors would like to thank the State of California for its support of university-based research, and especially for the funding provided in support of this project.

Long Distance Travel and Destination Attractiveness

UNIVERSITY OF CALIFORNIA INSTITUTE OF TRANSPORTATION STUDIES

June 2018

Konstadinos G. Goulias, Department of Geography, University of California Santa Barbara

Adam W. Davis, Department of Geography, University of California Santa Barbara

Elizabeth C. McBride, Department of Geography, University of California Santa Barbara

[page intentionally left blank]

TABLE OF CONTENTS

Executive Summary

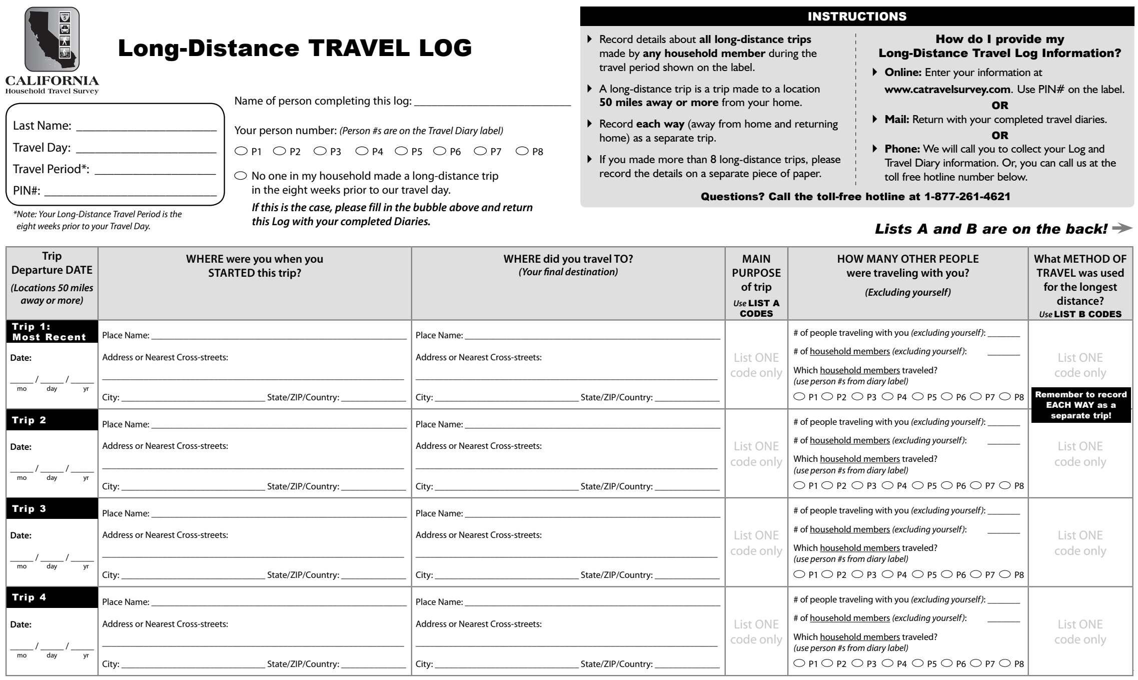
This report provides a summary of analyses using data of long distance tours by each household from an 8-week California Household Travel Survey travel log. Each tour record contains summary data from the 8-week log, a single day diary, household sociodemographic information, place of residence characteristics, and destination attractiveness. Each tour contains a main trip, selected tours with a main trip that is not a commute trip, and added destination descriptors from Foursquare. The first analysis, uses *Structural Equations Models* (SEM) and a simpler variant called *Path Analysis* on three censored variables (tour miles by air, miles driving, and miles by public transportation) and two categorical variables (main trip tour purpose) and number of overnight stays. The second analysis, uses *Latent Class Cluster Analysis* (LCCA) to identify five distinct, informative patterns of long-distance travel. This analysis shows that long-distance tours for vacation, business travel, medical, and shopping are substantially distinct in terms of their travel characteristics and correspond to different combinations of other activities in the tour and they are done by different types of households.

We find annual household income to be a major determinant of air travel and multiple overnight stays. Moreover, travel by air is more likely to be in large dense destinations with high attractiveness ratings. In contrast, car ownership is a major determinant of more mileage accrued driving a car to urban environments. In terms of main trip purpose, work and shopping are not associated with overnight stays and vacation/sightseeing is the purpose associated with longer overnight stays. Moreover, California destinations are more likely to be in tours completed within a day. We also find that vacation, sightseeing, and leisure trips are also more likely to be combined with other vacation and sightseeing trips. In contrast, business and combined business and leisure trips are less likely to be combined with vacation trips. We also find substantial and significant differences in long distance tour behavior among residents of urban versus rural environments. The methods used here to identify the typology of long distance travel can be easily expanded to include a variety of other explanatory variables of this type of behavior in more focused data collection settings.

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 a 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 1.1. 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.

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).


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)

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 simila 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 Figures 2.3, 2.4, and 2.5. The next section provides an overview of the analysis of trips and tours to understand the data we have and to guide subsequent work. Additional detailed trip by trip analysis of the data we use here can be found in Goulias et al. (2017).

Figure 2.1 One-Day Journey

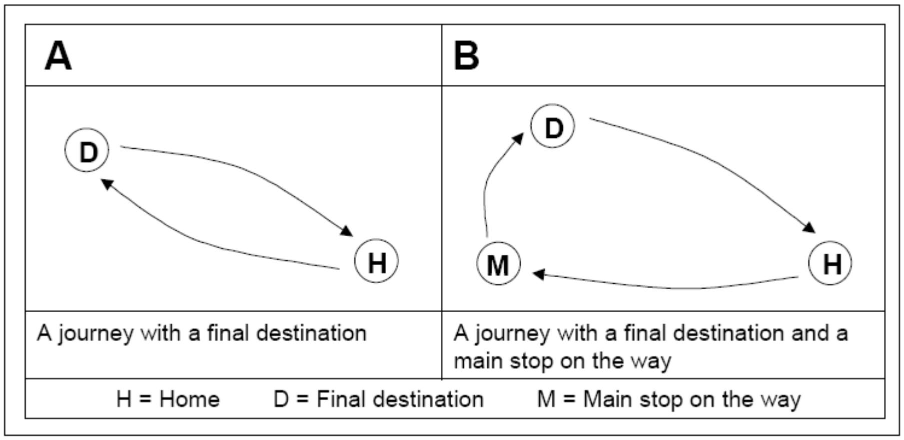
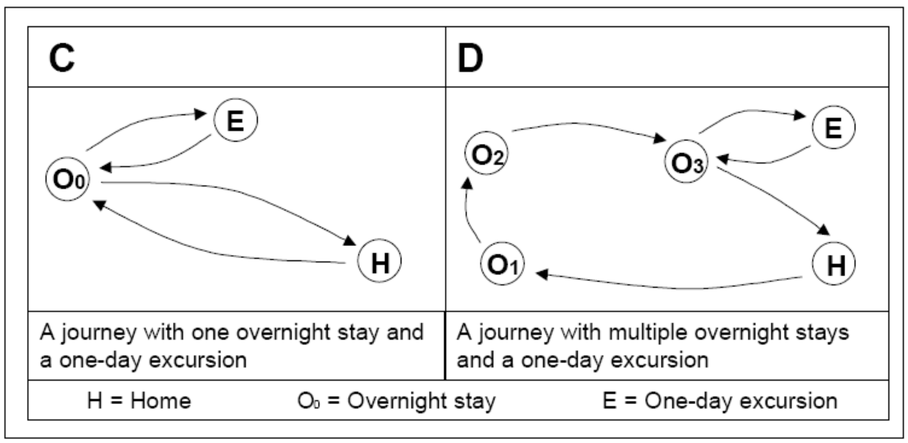


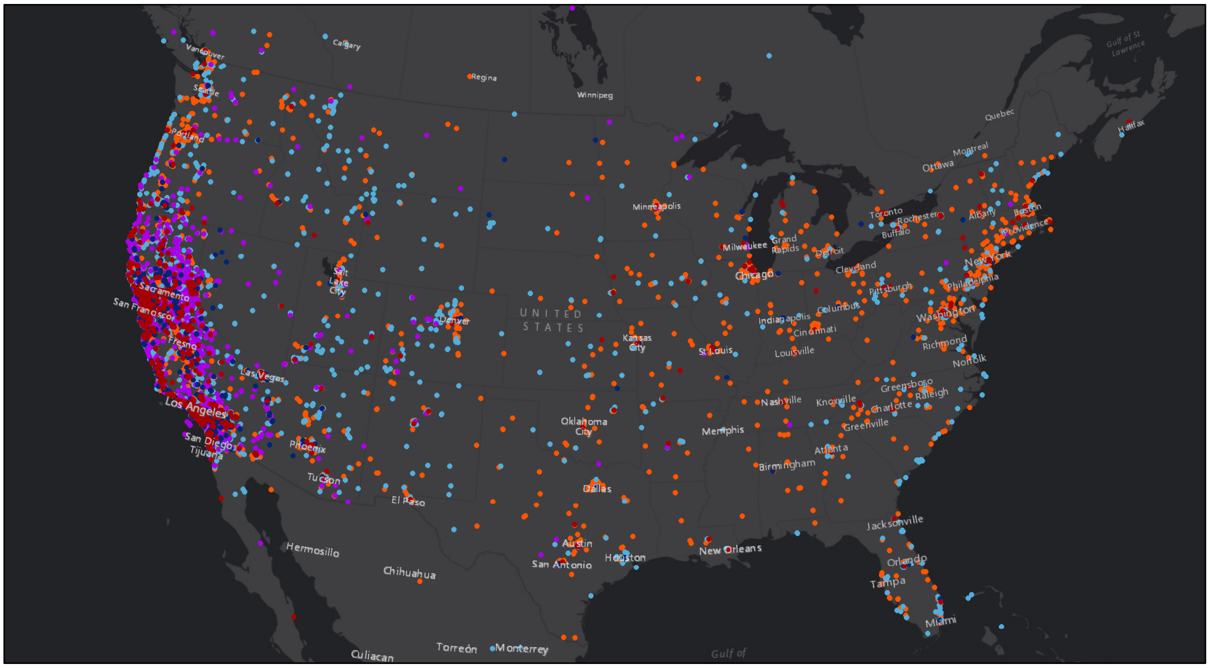
Figure 2.2 Multi-Day Journey



Figures 2.1 & 2.2 Taxonomy of long distance travel

(source DATELINE project, Broeg et al., 2003)

Long Distance Trips to US Destinations



- Combined business and pleasure
- Entertainment (theater, concert, sports event, gambling, etc.)
- Outdoor recreation (sports, fishing, hunting, camping, boating, etc.)
- Vacation/sightseeing
- Visit friends/family/relatives

0 250 500 1,000 1,500 2,000 Kilometers

Service Layer Credits: Esri, HERE, DeLorme, MapmyIndia, © OpenStreetMap contributors, and the GIS user community
Esri, HERE, MapmyIndia, © OpenStreetMap contributors, and the GIS user community

Figure 2.3 Long Distance Destinations in CHTS in the US

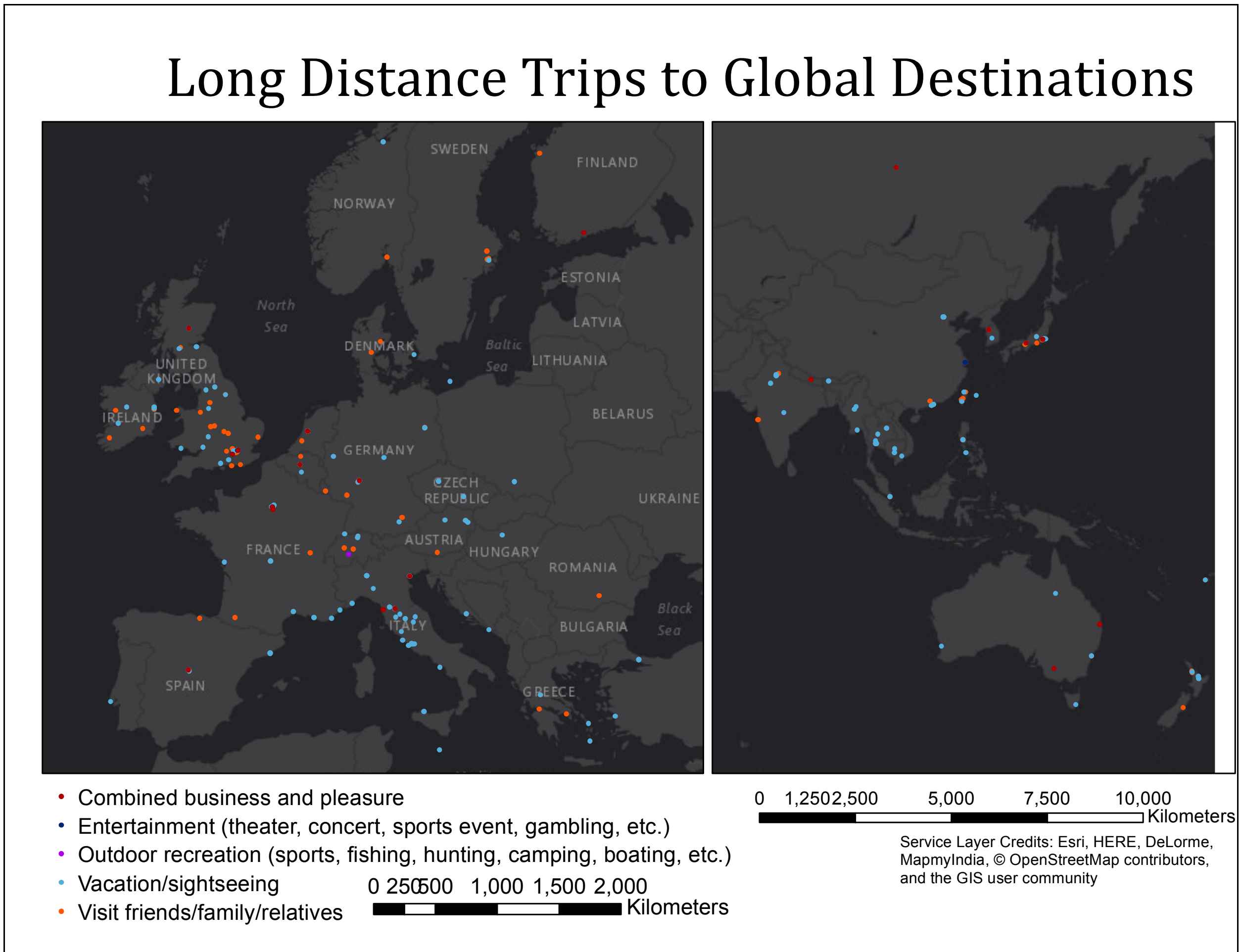


Figure 2.4 Long Distance Destinations in CHTS outside the US

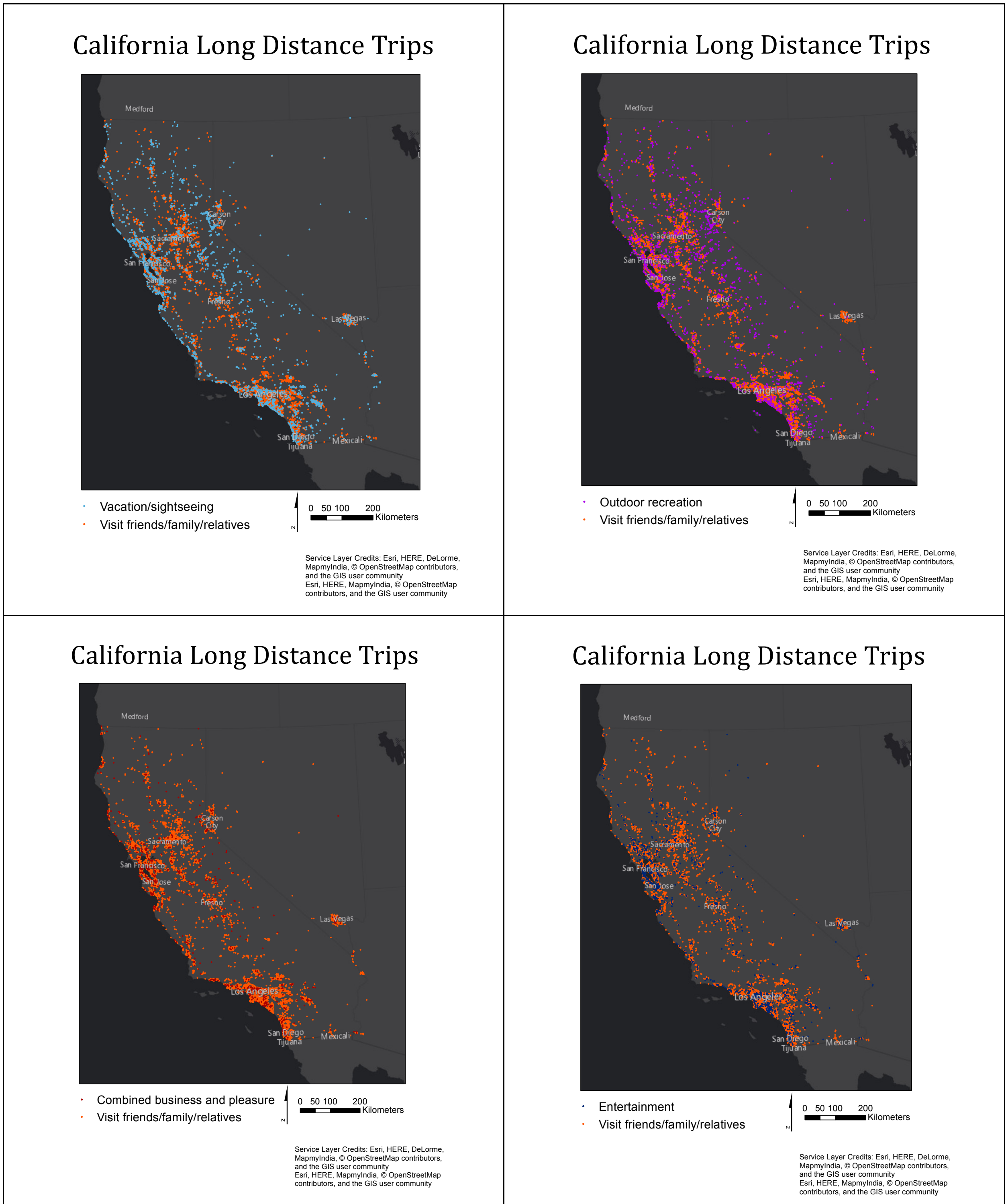


Figure 2.5 Long Distance Destinations in CHTS within California

3. 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 past Caltrans projects we know that including land use 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.

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 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.

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.

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 in McBride et al., 2017, 2018, and 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 3.1 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.

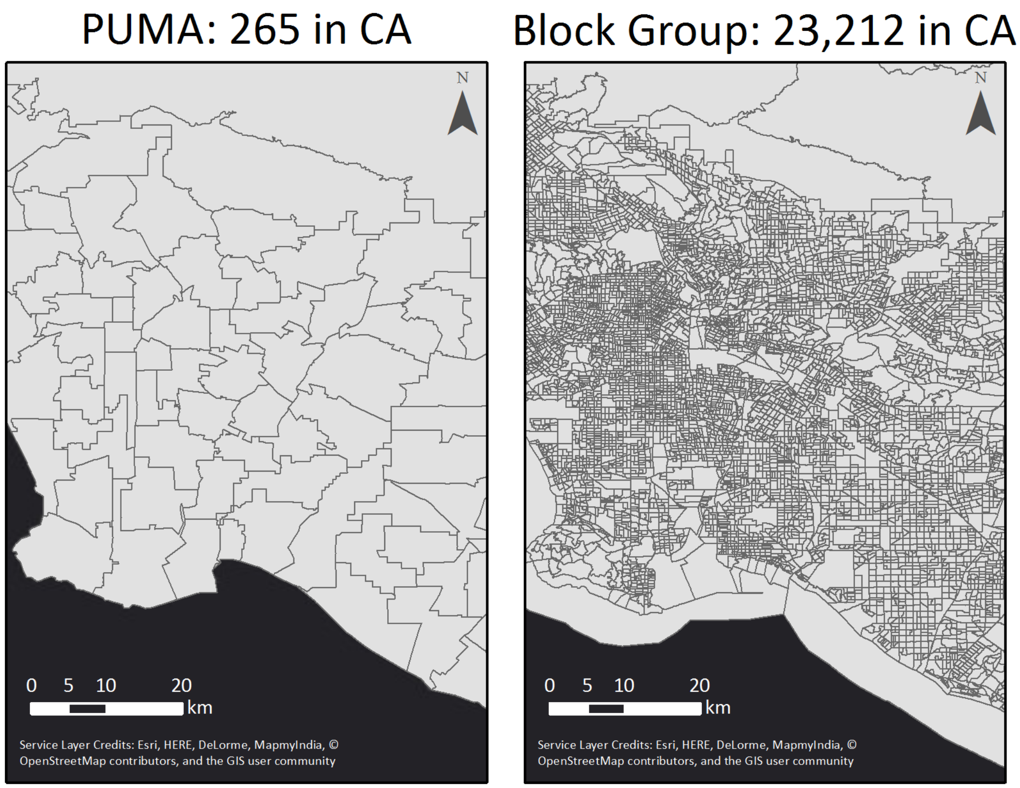


Figure 3.1 PUMA areas versus block group areas (Los Angeles)

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.

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 3.1. 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.

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 3.1 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 3.2 shows the end result with a map of California.

Table 3.1 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

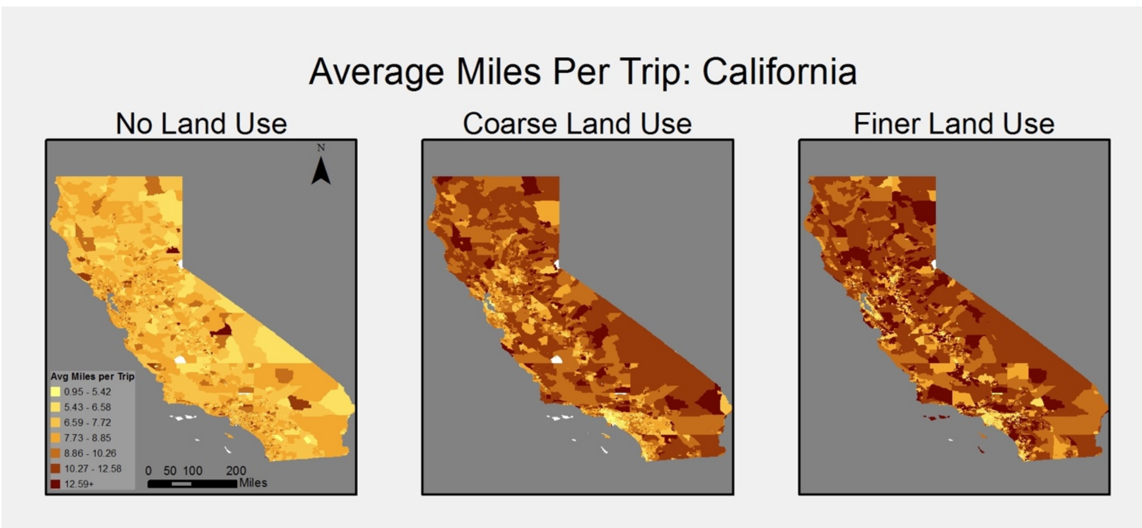


Figure 3.2 Comparison of the three synthetic population methods

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 2018), 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.

4. 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 4.1 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).

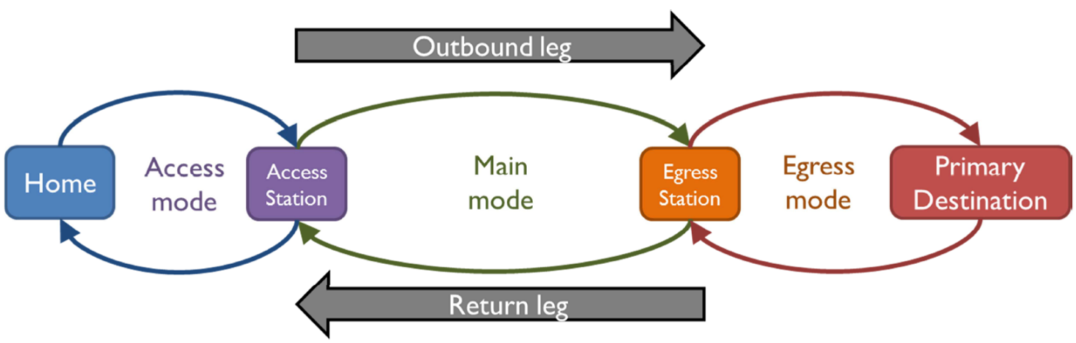


Figure 4.1 Long distance conceptual tour structure (reproduced from CSI, 2014)

The literature review shows that only accessibility and different types of level of service at the destination are used as determinants of travel. These are descriptors of locations but do not capture the meaning of place (e.g., historical land marks). Also, as noted in Goulias et al, 2017 and Davis et al. 2018a and 2018b, 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 4.2 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.

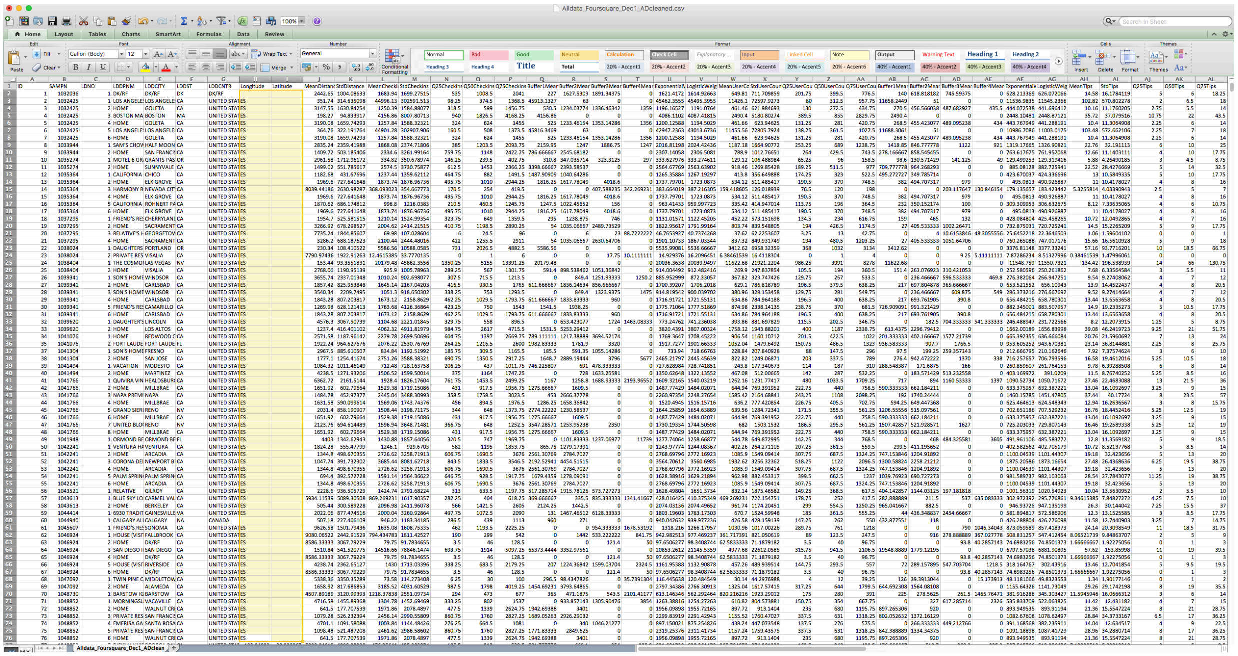


Figure 4.2 The augmented database that includes social media indicators

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.

5. Structural Relations Among Behavioral Variables

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 5.1a and 5.1b show a list of the variables and their averages. Table 5.2 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 5.1a 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 5.1b 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 5.2 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.

Structural Equations Model with Latent Variables

Figure 5.1 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 5.3 and 5.4). 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.

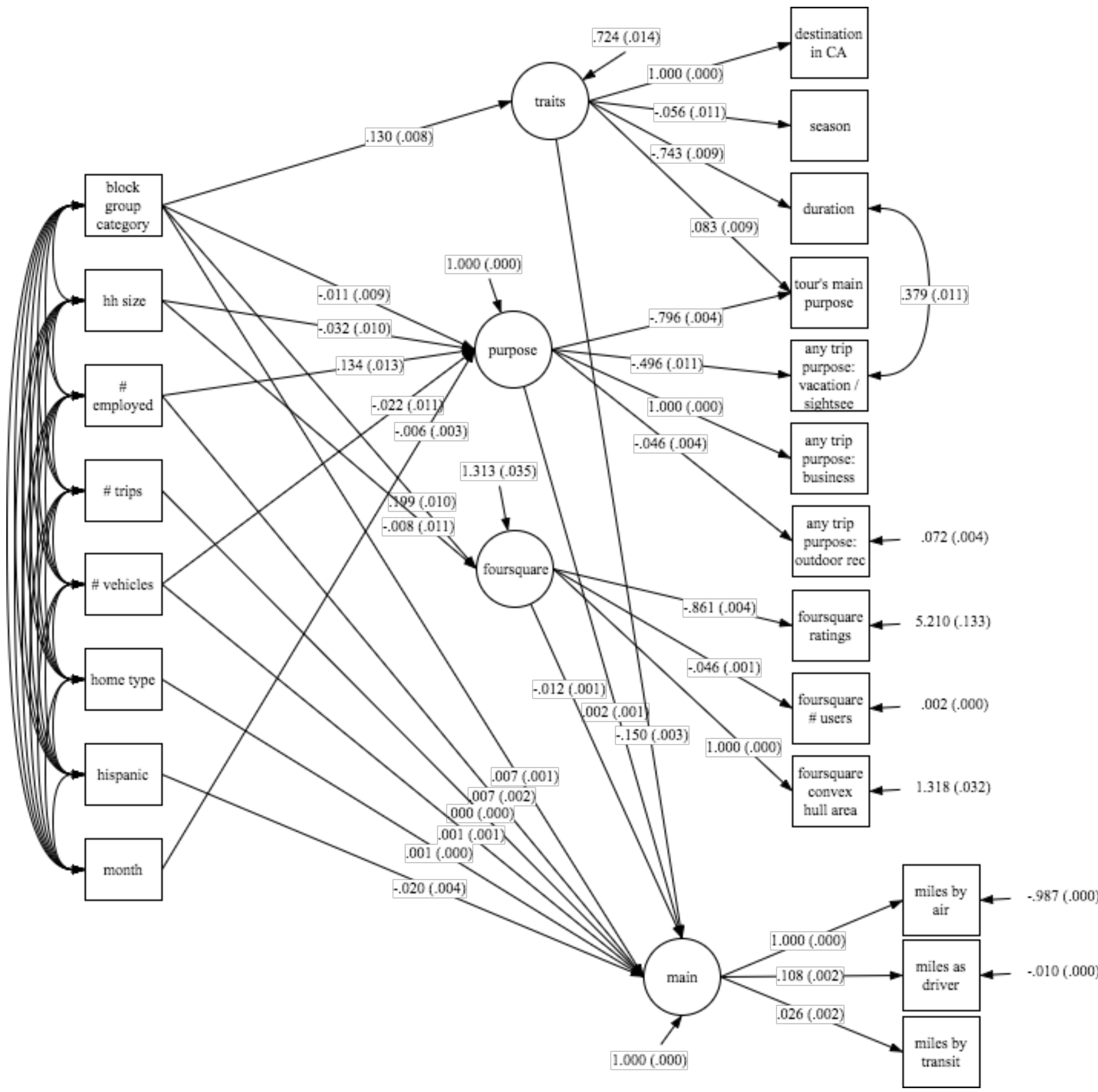


Figure 5.1 Path diagram of SEM with latent variables

Table 5.3 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 5.4 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 5.2 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 5.2 are the regression coefficients with blue the positive coefficients and red the negative. These are also shown in Tables 5.5a, 5.5b, 5.5c, 5.6a, 5.6b, 5.6c, 5.7a, 5.7b, and 5.7c with their significance.

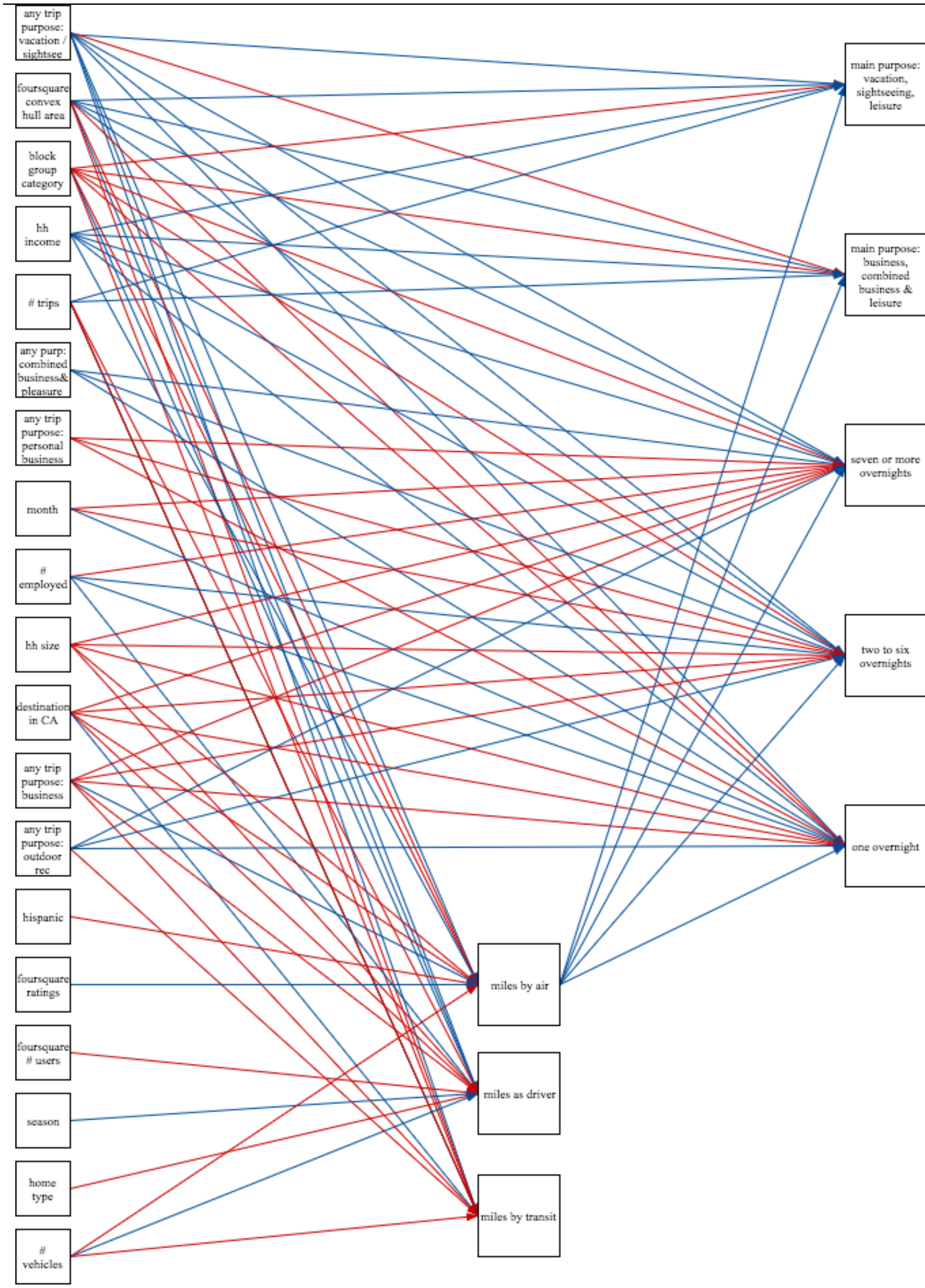


Figure 5.2 Path diagram of the path analysis model

Table 5.5a 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 5.5b 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 5.5c 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 5.6a 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 5.6b 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 5.7a 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 5.7b 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 5.7c 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 Davis et al., 2018). 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 5.8.

Table 5.8 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.

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, however recording quality of this variable is very inconsistent in the dataset.

Table 5.9a 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 5.9b 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

6. 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 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 California 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. We envision future studies on 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.

References

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.

Davis A.W., E.C. McBride, and K.G. Goulias (2018) A Latent Class Pattern Recognition and Data Quality Assessment of Non-commute Long-distance Travel in California. Paper (18-04945) accepted for presentation at the 97th Annual Transportation Research Board Meeting, January 7-11, 2018, Washington D.C. and inclusion in the TRB proceedings.

Davis A.W., E.C. McBride, K. Janowicz, R. Zhu, and K.G. Goulias (2018) A Tour-based Path Analysis of Long Distance Non-commute Travel Behavior in California. Paper (#18-00160) accepted for presentation at the 97th Annual Transportation Research Board Meeting, January 7-11, 2018, Washington D.C. and inclusion in the TRB proceedings.

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.

Goulias K.G., A.W.Davis, E. C. McBride, K. Janowicz, and R. Zhu (2017) Long Distance Travel in CHTS and Social Media Augmentation. Final Report. University of California Santa Barbara. Department of Geography, GeoTrans and STKO Laboratories, Contract Number: 65A0529, Task Order 047. Submitted to the California Department of Transportation, May 2017.

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 - No. 2668, pp. 11-20. DOI 10.3141/2668-02.*

McBride E.C., A.W. Davis, and K.G. Goulias (2018) A Spatial Latent Profile Analysis to Classify Land Uses for Population Synthesis Methods in Travel Demand Forecasting. Paper (18-06631) accepted for presentation at the 97th Annual Transportation Research Board Meeting, January 7-11, 2018, Washington D.C., inclusion in the TRB proceedings

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.