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 alternate format information, contact the Forms Management Unit at (916) 445-1233, TTY 711, or write to Records and Forms Management, 1120 N Street, MS-89, Sacramento, CA 95814.

1. REPORT NUMBER	2. GOVERNMENT ASSOCIATION NUMBER	3. RECIPIENT'S CATALOG NUMBER
CA16-2874		

4. TITLE AND SUBTITLE

Combining California Household Travel Survey data with harvested social media information to form a self-validating statewide origin-destination travel prediction method

5. REPORT DATE

April 18, 2016

6. PERFORMING ORGANIZATION CODE

7. AUTHOR

Principal Investigator: Konstadinos Goulias Lead Researcher: Jae Hyun Lee

8. PERFORMING ORGANIZATION REPORT NO.

N/A

9. PERFORMING ORGANIZATION NAME AND ADDRESS

University of California, Santa Barbara
552 University Road.
Santa Barbara, CA 93106.

10. WORK UNIT NUMBER

11. CONTRACT OR GRANT NUMBER

Contract 65A0529 Task 026

12. SPONSORING AGENCY AND ADDRESS

Division of Research, Innovation and System Information
P.O. Box 942873, MS-83
Sacramento, CA 94273

13. TYPE OF REPORT AND PERIOD COVERED

Final Report, 4/15/2015 – 3/31/2016

14. SPONSORING AGENCY CODE

15. SUPPLEMENTARY NOTES

16. ABSTRACT

Longitudinal data of persons and households is the best source of travel behavior information for assessing policy changes. However, this type of data is rarely available and difficult to collect due to administrative barriers and technical issues in survey design. Another empirical option which would allow estimation of induced demand is tested in this project. Data from multiple sources are used to produce a statewide inventory of travel patterns and an observatory to do this repeatedly for many years in the future. In order to combine social media harvested data with the California Household Travel Survey and data in the statewide travel model, we developed a step-wise conversion procedure including a Twitter trip extraction algorithm, a spatial aggregation technique, and statistical models to study the correlation among different databases. As a result, we were able to reproduce a list of Twitter trips, a trip generation table at the block group level, and an Origin-Destination matrix. We compared the list of Twitter trips with California Household Travel Survey records (CHTS), the trip generation table with synthetically generated population, and the OD matrix with California Statewide Travel Demand Model output. Twitter trips have longer distances and durations than CHTS trips, and there are not significant differences between weekday and weekend, and weekday and Thanksgiving day. In the comparison with synthetic population, we found positive correlation between Twitter trips and walking, bicycling, and single occupancy vehicle trips in both the total number of trips and sum of the trip lengths in block groups.

17. KEY WORDS	18. DISTRIBUTION STATEMENT	
California Household Travel Survey records (CHTS) California Statewide Travel Demand Model (CSTDM)	No restrictions. This document is available to the public through the National Technical Information Service, Springfield, VA 22161	
19. SECURITY CLASSIFICATION *(of this report)*	**20. NUMBER OF PAGES**	**21. COST OF REPORT CHARGED**
	76	

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.

A self-validating statewide origin-destination travel prediction method using the combination of California Household Travel Survey data with harvested social media information

FINAL REPORT

Jae Hyun Lee, Adam W. Davis, Elizabeth McBride, and Konstadinos G. Goulias

University of California Santa Barbara
Department of Geography
Geotrans Laboratory

Contract Number: 65A0529
Task Number: 026
Principal Investigator: Konstadinos G. Goulias
Lead Researcher: Jae Hyun Lee
Support Researchers: Adam W. Davis & Elizabeth McBride

April 18, 2016
Santa Barbara, CA

Disclaimer Statement

The contents of this report reflect the views of the author(s) who is (are) responsible for the facts and the 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 report does not constitute a standard, specification, or regulation.

Table of Contents

Abstract

Longitudinal data of persons and households is the best source of travel behavior information for assessing policy changes. However, this type of data is rarely available and difficult to collect due to administrative barriers and technical issues in survey design. Another empirical option which would allow estimation of induced demand is tested in this project. Data from multiple sources are used to produce a statewide inventory of travel patterns and an observatory to do this repeatedly for many years in the future. In order to combine social media harvested data with the California Household Travel Survey and data in the statewide travel model, we developed a step-wise conversion procedure including a Twitter trip extraction algorithm, a spatial aggregation technique, and statistical models to study the correlation among different databases. As a result, we were able to reproduce a list of Twitter trips, a trip generation table at the block group level, and an Origin-Destination matrix at the Public Use Microdata Area level from the social media data. We compared the list of Twitter trips with California Household Travel Survey records (CHTS), the trip generation table with synthetically generated population, and the OD matrix with California Statewide Travel Demand Model (CSTDM) output. Twitter trips have longer distances and durations than CHTS trips, and there are not significant differences between weekday and weekend, and weekday and Thanksgiving day. In the comparison with synthetic population, we found positive correlation between Twitter trips and walking, bicycling, and single occupancy vehicle trips in both the total number of trips and sum of the trip lengths in block groups. Lastly, we used a spatial lag Tobit model and latent class regression models to compare OD matrices from different sources to take censored distribution of trips and spatial heterogeneity into account. The single unit-contribution of Twitter trip in explaining CSTDM output was estimated with the former model, but four different unit-contributions depending on spatial structures is shown by the latter model.

Executive Summary

Longitudinal data is the best source of travel behavior information to assess policy changes because it has both “before” and “after” travel information from the same samples. However, longitudinal data collection faces many administrative barriers and technically complex survey design issues (Golob et al., 1997). Alternatively, one can design cross-sectional surveys as a before-after study and infer demand elasticity by examining differences in behavior between the before and after sample. This method is rarely seen in transportation applications for large geographical areas and in periods that are shorter than ten years. In fact, the California Household Travel Survey gave us this opportunity but we did not design the "after" survey that would allow estimation of induced demand based on current policies. There is, however, a third empirical option that we test in this project.

In this research project we use multiple data sources to produce a statewide inventory of travel patterns and an observatory to do this repeatedly for many years in the future. In this way we can develop a baseline short travel inventory that includes statewide vehicle miles traveled (VMT). In order to combine social media harvested data with the California Household Travel Survey and data in the statewide travel model, we developed a step-wise conversion procedure.

Firstly, a Twitter data harvester was developed to convert 8 million tweets with geographic locations to trips. To do this, we developed three different Twitter trip extraction algorithms using distance, time route distance, and inferred trip duration. We then compared this information with the statewide travel demand model output and identified the best rules for the extraction. The list of Twitter trips was then transformed into a trip generation table and an OD matrix with spatial aggregation. We also compared the list of Twitter trips with California Household Travel Survey (CHTS) records and a trip generation table with synthetically generated population that was estimated in “Task 2644: Spatial Transferability Using Synthetic Population Generation Methods.” We did the same with the OD matrix of the California Statewide Travel Demand Model (CSTDM) output.

As a result, we found positive correlation between Twitter trips with walking, bicycling, and single occupancy vehicle trips from the synthetic population in both total number of trips and sum of

the trip lengths in block groups. Twitter data was not able to capture the differences between weekday and weekend, and weekday and Thanksgiving day. In terms of trip lengths, Twitter trip data has similar distributions to the California household travel survey data, but their trip durations were slightly different. In addition, Twitter data produce a smoother distribution than the other because it was computed using Google API. We compared the Twitter based OD matrix with a recent OD matrix (CSTDM output) given by the California Department of Transportation. We used a Spatial Lag Tobit model (a model that accounts for zero trips at geographic units and takes into account spatial relationships among zones) to develop an unbiased conversion method between Twitter trips and Travel Demand Model output. We also used Latent Class Regression models to incorporate spatial differences among zones. In these models, we also include land use and demographic characteristics so that CSTDM trips are adjusted by land use and demographic variables.

One objective is to convert Twitter estimated trip making into similar trip making estimated by other sources. In this context, a unit contribution is the multiplier we apply to a Twitter trip to estimate trips made by California residents. The spatial lag Tobit model produced a single unit-contribution (33.021) of Twitter trips in explaining the number of trips from CSTDM, but four different unit-contributions depending on spatial structures were obtained with Latent Class Regression model (class 1: 2.1279, class 2: 16.0510, class 3: 183.3002, class 4: 32.6550). Although the largest proportion of the sample (OD pairs) was found in the first class (88 %), followed by the second, third and fourth class (6%, 4%, and 2%, respectively), in terms of CSTDM OD trips, by far the largest proportion of OD trips (67.3%) were found in the fourth class followed by the third, second, and first class (26.6%, 4.6%, and 1.5%, respectively). In terms of spatial distributions of the OD pairs of latent classes, those are distributed differently across California. The first class seems to represent all of the long distance OD pairs, from the second class to the fourth class, the spatial distributions of the OD pairs tend to be much shorter than the first class. Although the second and third classes cover some inter-regional OD pairs between zones, the fourth latent class seems to cover inner zone trips as well as the shortest OD pairs. Finally, Tobit models for four MPO areas produced different unit-contributions of Twitter trips, and Southern California

Association of Governments area has the lowest conversion coefficient and Sacramento Area Council of Governments has the highest one.

With the comparison of Twitter trips and synthetic population, we found walking trips are strongly related to Twitter trips, so our immediate next step is to perform in-depth analysis of Twitter trips and their relationship with walking trips. In addition, a longer than 6-months observation period would be best to collect data of this type. For this reason, we recommend the creation of a multi-year observatory. Data and the findings from this project will also be used in another Caltrans project (Task Order- 65A0529 TO 047: Long Distance Travel in the California Household Travel Survey (CHTS) and Social Media Augmentation).

Although we focused on using geographic information of tweets to extract Twitter trips, using all other information provided by Twitter would be very helpful. For example, we may be able to identify trip information from text mining of tweets, and potential home locations of heavy Twitter service users with night time tweets' location. Moreover, we can also impute missing trips when two tweets' time difference is much longer than estimated travel time based on each user's home locations and major tweeting locations. In this way, we can identify social media data with high potential of complementary information to the traditional survey data. Finally, a method to extract data from group quarters for which travel behavior surveys are usually not available may be another potential next step.

1. Introduction

Longitudinal data is the best source of travel behavior information to assess policy changes. When we interview people repeatedly over time and preferably before and after a policy takes place we observe possible trend in behavior, change in behavior, and compute elasticity to change. This is not an easy data collection setting because longitudinal data collection faces many administrative barriers and technically complex survey design issues (Golob et al., 1997). Alternatively, one can design cross-sectional surveys as a before-after study and infer demand elasticity by examining differences in behavior between the before and after sample. This method, although feasible, is rarely seen in transportation applications for large geographical areas and in periods that are shorter than ten years. In fact, the California Household Travel Survey gave us this opportunity but we did not design the "after" survey that would allow estimation of induced demand based on current policies. There is, however, a third empirical option that we test in this project.

In this research project we use multiple sourced data to produce a statewide inventory of travel patterns and an observatory to do this repeatedly for many years in the future. In this way we can develop a baseline short travel inventory that includes statewide vehicle miles traveled (VMT). Then, a procedure is created to monitor the evolution of travel in California using data from social media adjusted by region and correlated with land uses at fine geographic areas. First, we use a synthetic inventory of travel in our State providing the reference needed to monitor if change in travel takes place due to changing land use patterns. We combine information in the California Household Travel Survey, data in the statewide travel model, and social media harvested data. Second, a conversion procedure is created to transform harvested data from the web into origin-destination travel and trip lengths to produce estimates of VMT. We modified a method created at UCSB with success with an added algorithm verification component. As we discuss later in this report, we reproduce OD matrices by converting twitter data to physical trips, but we also use the social media data as a statewide monitoring device. We test the effectiveness of methods using a variety of statistical techniques and algorithmic options. Third, an automated procedure to harvest data and convert them into travel predictions statewide is created and then

used to derive estimates of travel demand. We view this as the beginning of an observatory that we can initiate with other researchers statewide to collect data for multiple purposes. This allows us to share the data collection and model estimation costs among many agencies.

In summary the research questions of this project are:

1) Is Twitter a valid source of travel behavior data?
2) Can we compare Twitter with synthetic population travel behavior?
3) Are there population segments that are better represented by Twitter data or travel patterns?
4) Is there a way to convert Twitter data to a statewide origin destination matrix?
5) Does accounting for spatial auto-correlation influence the conversion of Tweets to statewide origin destination matrix?
6) Is there spatial heterogeneity in all of the above?

2. Literature Review

Household travel survey data have been playing the most significant role in travel behavior research for many years. Although information extracted from this source allows governments to develop their transportation plans, the cost of data collection increase over time (Leiman, Bengelsdorf, & Faussett, 2006). Furthermore, recently developed modeling and simulation approaches for travel demand forecasting require even more detailed information than many travel surveys provide (Goulias et al., 2012). Respondent burden is a factor for increased costs and we see many attempts to reduce the respondent burden with a variety of new technologies including Global Positioning System loggers, computer-based survey systems, smartphone applications, personal digital assistants, and car navigation systems (Auld, Williams, Mohammadian, & Nelson, 2009; Cottrill et al., 2013; Fan, Chen, Liao, & Douma, 2012; Nitsche, Widhalm, Breuss, & Maurer, 2012; Shirima et al., 2007; Turner, 1996).

More recently, online social media services (e.g. Facebook, Foursquare, and Twitter) have received attention from a range of social scientists. Although the use of social media services is heavier in people under the age of 30, about 23% of the American internet-using population in 2014 use this service (Duggan et al., 2015), data derived from social media sources has become attractive to many researchers because of their unique advantages. The data is free to use and it provides a detailed temporal record of users' locations as well as textual information about users' activities and their emotional status. Yang & Mu (2015) and Yang, Mu, & Shen (2015) used text mining algorithms for twitter data to detect depressed users, their socio-economic characteristics, and climate. Health of food activities were also explored with a text mining algorithm called sentiment analysis (Widener & Li, 2014). On the other hand, the geographical and temporal details of twitter data were found to be very useful in identifying individuals' food environment (X. Chen & Yang, 2014); their activity space (Lampoltshammer, Kounadi, Sitko, & Hawelka, 2014; Lee, Davis, Yoon, & Goulias, 2015); and enhancing areal interpolations of resident population (Lin & Cromley, 2015).

Transportation researchers have also begun to consider social media data in travel behavior analysis. Collins et al. (2013) used about 500 twitter texts to evaluate transit riders' satisfaction with a Sentiment Strength Detection Algorithm. Cebelak (2013) focused on Foursquare, and she was able to reconstruct a zonal Origin-Destination matrix based on the check-in counts for Austin, Texas. Another travel analysis example with Foursquare data was given by Hasan & Ukkusuri (2014). They employed activity pattern model to infer the latent pattern of weekly activities with geo-location data. Chen, Frei, & Mahmassani (2015) also use this dataset to explore activity and destination choice behavior. Coffey & Pozdnoukhov (2013) used a different algorithm to study temporal and spatial patterns of activity participation using tweets.

The algorithm used in this paper was first used to extract Origin-Destination pairs from Twitter data for the Greater Los Angeles metropolitan area (Gao et al., 2014). This algorithm contains two steps: individual-based trajectory detection and place-based trip aggregation. In essence, if a person posted tweets in two different geographic zones within 4 hours, they are assumed to have made an OD-trip between these zones. The extracted OD-trips were aggregated into 30-minute intervals. Then, these trips were compared with the commuting data in the American Community Survey (ACS) for validation. Among the Weekday, Weekend, and Christmas datasets, Weekday data has the most similar temporal distribution of trips (Pearson correlation coefficient=0.91, p=0.002). This high correlation is misleading due to extreme spatial heterogeneity that we investigate here in more detail using latent variable regression methods.

3. California Household Travel Survey

The California Household Travel Survey (CHTS) is designed to support California's new transportation policy framework, building an inventory of travel behavior and taking into account possible use of new mobile technologies, as its Steering Committee clearly defined in the following paragraph during the inception of a partnership to build a consortium of agencies supporting CHTS.

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

One objective for the data collected in this household travel survey is to develop a variety of travel demand forecasting systems throughout the State and integrate land use policies with transportation policies (CALTRANS, 2016). Very important for regional agencies is the provision of suitable data that inform a variety of new model developments including activity-based models (ABM) and their integration with land use models at the state level and for each of the four major Metropolitan Planning Organizations (MPO) that surround Sacramento, San Francisco, San Diego, and Los Angeles. It is also the source of data for the many refinements of older four-

step models and activity-based models in smaller MPOs and serves as the main source of data for behavioral model building, estimation of modules in other sustainability assessment tools, and the creation of simplified land use transportation models. Moreover, added details about a variety of choice contexts of households such as car ownership and car type are collected to develop a new set of prediction models to more accurately estimate emissions of pollutants at unprecedented levels of temporal and spatial resolutions.

CHTS meets the data needs criteria for a main core survey with satellite in-depth survey components that is similar in design to the ideal travel survey described at an international travel survey methods conference recently (Goulias et al., 2013). Figure 1 provides a pictorial representation of the CHTS survey design and its components. The CHTS databases include data collected by one contractor (NUSTATS) for the entire State of California and an added sample and supplement collected by another contractor (Abt-SRBI) for Southern California Association of Governments - SCAG. The databases include information about the household composition and facilities available, person characteristics of household members, and a single day place-based activity and travel diary. There are two stages in data collection with the first stage called the recruitment and the second called the retrieval. Sample selection was done using residential addresses and stratification to populate the final database with households that live in lower density environments. Additional efforts were made to identify areas where response rate was expected to be low and intensify efforts to recruit residents in those areas. Details about the sampling method and efforts to make the resulting sample representative of the population in California can be found in Nustats, 2013. CHTS data were collected using paper and pencil as mail-in and mail-back survey, telephone (Computer Aided Telephone Interview, CATI), and the Internet using an interactive survey interface. CHTS also includes GPS data collection and a component administered by a different consultant for the California Energy Commission (CEC). All recruits were invited by an initial letter, TV videos (https://www.youtube.com/watch?v=h1KjCZQaDJ8). An effort was also made to contact community leaders and increase awareness of the public about the survey. A variety of monetary incentives were also used for different parts of the survey depending on the amount of time

people needed to dedicate to record their responses. During the design and pretesting phases of the project a high degree of harmonization among the three instruments of data collection was achieved using national guidelines (Goulias and Morrison, 2010, NUSTATS, 2013).

The CHTS (NUSTATS and Abt-SRBI) sample selection is a combination of exogenously stratified random and convenience sampling scheme (see NUSTATS Final Report, 2013). The final delivered databases for the statewide databases include slightly over 42,000 households (approximately 109,000 persons) for the core survey with most of their information complete. The core statewide CHTS travel days reported by respondents started on February 1, 2012 and ended in January 31, 2013 and include weekdays and weekends and spanned 58 counties of California and covering 366 days. CHTS is a joint effort among agencies to procure data collected using the same standards and funding was provided by Caltrans ($4,221,000), Strategic Growth Council ($2,028,000), Metropolitan Transportation Commission ($1,515,000), Southern California Association of Governments ($1,415,834), Council of Fresno County Governments ($49,500), Kern Council of Governments ($118,000), Association of Monterey Bay Area Governments ($183,810), San Joaquin Valley Air Pollution Control District ($150,000), Santa Barbara County Association of Governments ($33,000), Tulare County Association of Governments, ($49,500), and California Energy Commission ($250,000). This leads to approximately $240 per complete household record. One way that decreased the costs of this survey and increased its response rate was to select a subset from the core of households to participate in different added components (the satellites of Figure 1).

The long distance travel component in CHTS is very important for statewide and regional applications to capture what is called interregional travel and long-distance travel. A long distance log (for trips longer than 50 miles) extending for up to 8 weeks before the diary day was also designed and administered. Many of the trips in this class are commute trips, business related, leisure related, visits to friends and family or simply long commutes. Regional forecasting applications need data to estimate this type of trip making, but also need data to correlate long distance travel and short distance travel. This data component aims to accomplish exactly this objective, and to enable the study of trade-offs people make when they engage in travel that, for

example, requires an overnight stay outside the home base. In addition, it is also desirable to study the relationship between land use and the propensity to make long distance travel. The separate long distance log instrument was designed to minimize burden to the respondents and maximize data yields and it is a retrospective survey of 8 weeks before the assigned travel date of the diary. This component will be used in a new project during the fiscal year 2016-17.

The GPS (person wearable and vehicle) data collection enables analysis of trip reporting and comparison with other media to detect any under-reporting, detailed description of route choice and time-of-day, day to day and geographic variations that other surveys do not capture. It also includes a small sample that is asked to use an on-Board Diagnostics device (OBD) to record data about car use with the GPS traces. For a review of GPS data and an example using CHTS GPS data see RSG et al. (2015).

The California Energy Commission component expands the car ownership questions of CHTS for a small subsample to include added questions about fuel types and vehicle technologies, measurement of refueling station availability, added questions on solar energy access plans, questions to differentiate between leased and purchased vehicles.

For the SCAG region, Abt-SRBI designed and administered a core survey for additional households and a supplemental satellite survey containing questions that are specifically needed for model building at SCAG. The SCAG supplement includes additional questions of behaviors that are related to travel and energy use and includes questions about residence characteristics, energy use at home, smart phone and internet use by respondents, added questions on bike use and parking availability at work and school, a series of attitudinal questions about tolls, and questions about cessation of driving.

Details about sample selection and contents of different CHTS components are available in NUSTATS (2013), CALTRANS (2013), and NREL (2016). Some early analysis of the data can also be found in Goulias et al. (2014) and MTC (2016).

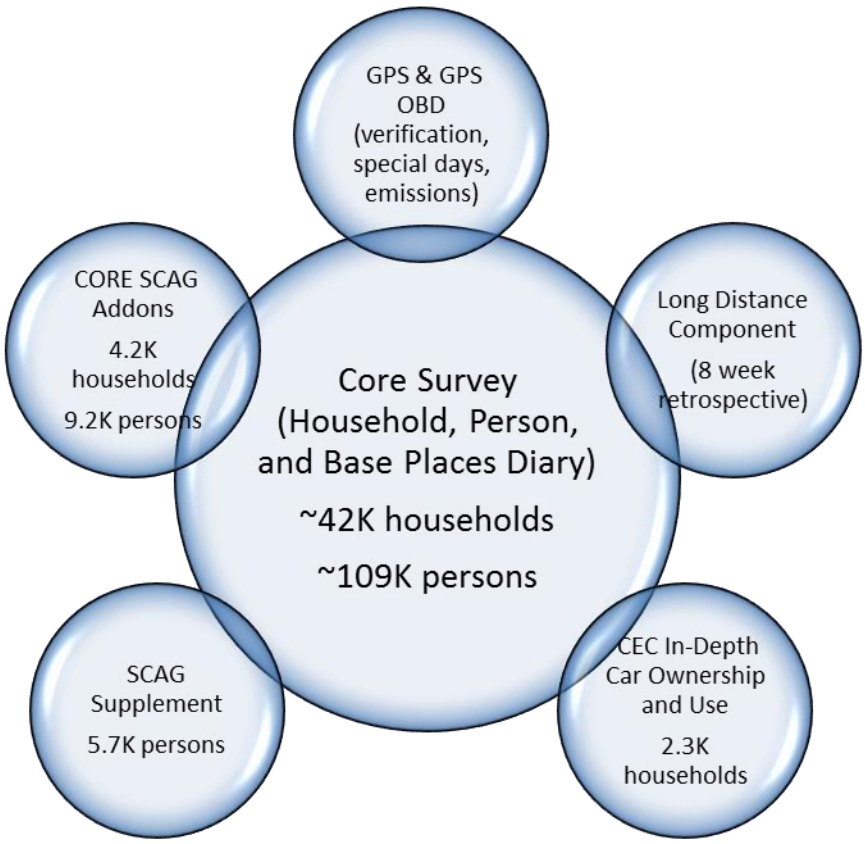


Figure 1. California Household Travel Survey Components

4. Synthetic Population Generation

The emergence of individual-based travel behavior models, including discrete choice models, and activity-based microsimulation models have ushered in a new era in travel demand forecasting. These models operate at the level of the individual traveler and the household within which this individual lives. These models are regression-like equations with dependent variables the behavior we are trying to predict (e.g., number of trips per day) and explanatory variables person and household characteristics such as age, employment, education, household size, and so forth. Therefore we need household and person attribute information to inform these models and then use them for the entire regional population to predict changes in behavior. However, such detailed information is virtually never available at the disaggregate level for an entire region. Public Use Microdata Sample (PUMS) and travel surveys provide this detail but they either do not include fine geographic levels or are not representing the study region. PUMS is particularly useful because it provides 1% and 5% samples from the US Census and in this way offers dependable joint distributions of multiple variables at the level of the most elementary unit of analysis. In this way, we can replicate multiple times observed multidimensional relationships among variables and generate a synthetic (virtual) population with comprehensive data on attributes of interest.

Synthetic populations can be formed from a small random sample from which we extract key information about the relationships among a set of household and person variables. These relationships are the multidimensional (from multiple variables) distributions we want to replicate in the entire population (e.g., the cross classification between household size and number of employed persons). The sample that is used to create this multidimensional distribution is called the "seed" that starts a set of iterations. These iterations reconcile seed univariate (single variable) distributions with aggregate distributions of household and person attributes available through the US Census at small geographic units such as a block or block group. These univariate distributions are called the "marginal" distributions. In the US, Census Summary Files provide the marginal distributions of population characteristics and they can be

either from the Decennial Census or the American Community Survey (ACS) that replaced the older US Census long form. The joint distributions among a set of control variables are first estimated using the seed and then their values adjusted using the Iterative Proportional Fitting (IPF) procedure first presented by Deming and Stephan (1941). In this section we review briefly the methods that have been applied to population synthesis by Beckman et al. (1996) for the use in TRANSIMS, Guo and Bhat (2007) for Texas, Auld et al. (2007) for Illinois, and Ye et al. (2009) for Florida, Arizona, and California.

Most population synthesizers currently in use are based on the method developed by Beckman et al. (1996) for use in the TRANSIMS model. This procedure matches exact large area multidimensional distributions of selected variables from the PUMS files to small area marginal distributions from Census Summary files to estimate the multidimensional distributions for the small areas. The population is synthesized in two stages. First a multidimensional distribution matrix describing the joint aggregate distribution of demographic and socio-economic variables at household and/or individual levels is constructed. This stage makes use of the IPF procedure. In this procedure, the correlation structure of the large area and within it the smaller areas is assumed to be similar. In the IPF procedure, an initial seed distribution is used and fit to known marginal totals. The difference between the current total and the marginal total for each category of the variable of interest is calculated and the cells of that category are updated accordingly. This process continues for each variable until the current totals and the known marginal totals match to some level of tolerance, producing a distribution which matches the control marginal totals. In the second step, synthetic population is constructed by selecting entire population from the PUMS in proportion to the estimated probabilities given in the multidimensional matrix obtained by the IPF technique. The number of households to be generated of each demographic type is determined from each aggregate area (or large area). For a combination of demographic characteristics a set of probabilities is assigned to each household in the PUMS, where PUMS samples close to the combination of desired demographic characteristics are assigned higher probabilities. The households are then selected randomly according to their selection probabilities. These probabilities are computed by a weight based algorithm (Beckman et al., 1996).

Guo and Bhat (2007) identify two issues associated with the first generation of population synthesis using the Beckman et al. (1996) algorithm. The first issue is incorrect zero cell values: this is an issue inherent to the process of integrating aggregate data with sample data, and the problem occurs when the demographic distribution derived from the sample data is not consistent with the distribution expected in the population. A second issue arises from the fact that the approach can control for either household-level or person-level variables, but not both. If these issues are left unaddressed, may significantly diminish the representativeness of the synthesized population. Guo and Bhat (2007) propose a new population synthesizer that addresses these issues using an object-oriented programming paradigm. The issue of incorrect zero cell values is solved by providing the users capability to specify their choice of control variables and class definitions at run time. Furthermore, the synthesizer is built with an error reporting mechanism that tracks any non-convergence problem during the IPF procedure and informs the user of the location of any incorrect zero cell values. Guo and Bhat (2007) also propose a new algorithm using IPF based recursive procedure, which constructs household-level and person-level multi-way distributions for the control variables. This is achieved by the two multi-way tables for households and persons that are used to keep track of the number households and individuals belonging to each demographic group that has been selected into the target area during the iterative process. At the start of the process, the cell values in the two tables are initialized to zero to reflect the fact that no households and individuals have been created in the target area. These cells are iteratively updated as households and individuals are selected into the target area. Given the target distributions and current distributions of households, each household from PUMS is assigned a weight-based probability of selection. Based on the probabilities computed, a household is randomly drawn from the pool of sample households to be considered and added to the population for the target area. A similar idea underlines the processes developed by Pritchard and Miller, 2012, and the PopGen method we review below.

Building on the IPF procedure for population synthesis Auld et al. (2007) propose a new population synthesizer which consists of two primary stages: creation of multidimensional distribution table for each analysis area and the selection of households to be created for each

analysis area. Auld et al. (2007) adopt the same method for creating a multidimensional distribution table as in other population synthesizers (Beckman et al. 1996, Guo and Bhat, 2007). The complete distribution for all households is fit to the marginal totals through the use of IPF procedure. This creates the regional-level multi-way table that is used to seed all the zone-level distribution tables. For each zone the seed matrix cell values are adjusted so that the total matches the desired number of households to generate. The zone-level multi-way distribution is adjusted to match the zone marginal distributions by again running the IPF procedure. The selection probability of households from the multidimensional table is performed in a similar manner as that proposed by Beckman et al. (1996), which is a weight of household divided by the sum of the total weighted households for the category variable. Auld et al. (2007) argue that there exists large variation between control marginal totals and those generated by the process so the totals are matched exactly as desired. For this reason, Auld et al. (2007) add further constraints, such that the total number of households that have been generated for each category within each control variable represented by the demographic type. If any of the totals exceed the marginal values from the zone-level marginal by more than a given tolerance, the household is rejected. This procedure works well at keeping the generated marginal totals fairly close to the actual totals. However, Auld et al. (2007) identify that this method might bias the final distribution. In the population synthesis procedures, aggregating control variables within range-type control variables is primarily done to allow for the use of more control variables and to reduce the occurrence of false zero-cells. For problems with large number of control variables the size of the distribution matrix can become very large making the IPF procedure intractable. Therefore, Auld et al. (2007) introduced the category reduction option, which occurs prior to the IPF stage. The marginal values for range variables are compared to minimum allowable totals. The minimum allowable category total is defined as the total number of households in the region multiplied by a user specified percentage. The percentage forces all categories with less than the allowable number of households to be combined with neighboring categories. The category is then removed from the multidimensional distribution table. The category aggregation threshold percentage acts as a useful limiter of the total number of categories.

Ye et al. (2009) propose a similar framework by generating synthetic populations with a practical

heuristic approach while simultaneously controlling for household and person level attributes of interest. The proposed algorithm uses lessons learned from the three examples above and is also computationally efficient addressing a practical requirement for agencies. The proposed algorithm by Ye et al. (2009) is termed as Iterative Proportional Updating (IPU), it starts by assuming equal weights for all households in the sample. The algorithm then proceeds by adjusting weights for each household/person constrain in an iterative fashion until the constraints are matched as closely as possible for both household and person attributes. The weights are next updated to satisfy person constraints. The completion of all adjustment weights for one full set of constraints is defined as one iteration. The absolute value of the relative difference between weighted and the corresponding constraint may be used as goodness-of fit-measure. IPU algorithm provides a flexible mechanism for generating synthetic population where both household and person level attribute distributions can be matched very closely. The IPU algorithm works with joint distributions of households and persons derived using the IPF procedure, and then iteratively adjusts and reallocates weights across households to match closely the household and person level attributes. As mentioned in earlier works (Beckman et al. 1996; Guo and Bhat 2007; Auld et al. 2007), the problem of zero-cells is also addressed in the population synthesis by Ye et al. (2009) borrowing the prior information for the zero-cells from PUMS data for the entire region. Moreover, due to the proposition of the IPU algorithm, Ye et al. (2009) indicate that zero-marginal problem is encountered in this context. For example, it is possible to have absolutely no low-income households residing in a particular blockgroup. If so, all of the cells in the joint distribution corresponding to low income category will be eliminated and they solve this problem by adding a small positive value to the zero-marginal categories. The IPF procedure will then distribute and allocate this small value to all of the relevant cells in the joint distribution. After the weights are assigned using the IPU algorithm, households are drawn at random from PUMS (or a survey database) to generate the synthetic population. The approach Ye et al. (2009) adopt is similar to that of Beckman et al. (1996), except that the probability with which the household is drawn is dependent on its assigned weight from the IPU algorithm. This algorithm implemented in the software PopGen was refined and used in a large geographical area with 18 million residents (The Southern California Association of Governments, SCAG,

region). The application took a reasonable low number of hours to run with multiple dimensions at the household and person levels and performed very well in terms of its ability to replicate extremely different marginal distributions at the household and person levels (Pendyala et al., 2012a, 2012b).

Synthetic populations, in addition to providing the explanatory variables for individual and household behavioral equations, are also used to provide the baseline population for demographic microsimulators, and the population for urban economy simulators (see the review by Ravulaparthy and Goulias, 2011. There are also many extensions of the methods described here including a two-stage IPF to add spatial information from different sources Zhu and Ferreira (2014); a Markov Chain Monte Carlo approach by Farooq et al. (2013) to ensure uniqueness of the identified distribution, avoidance of loss of heterogeneity, and poor scalability of IPF-based methods; and extending PopGen to multiple geographical scales (Konduri et al., 2016).

In this project, we use the synthetic population and travel demand predicted in the University of California Transportation Center and CalTrans project with title “Task 2644: Spatial Transferability Using Synthetic Population Generation Methods.”. This synthetic population was generated with PopGen software and algorithms with the addition of land use characteristics in the area of each household's residence. The seed data are from the California Household Travel Survey and the land use information from NETS data. To match the CHTS with land use data we use the 2012 land use characteristics of NETS. As a result, 34,589,650 persons were generated, and each person makes 3.31 trips, and 24 miles traveled per day (McBride et al., 2016).

5. California Statewide Travel Demand Model (CSTDM)

The California Statewide Travel Demand Model was developed to forecast all of the California residents' personal travel as well as commercial vehicle travel on a typical weekday when schools are in session. The original 2010 version of the CSTDM (CSTDM v2.0) was updated in 2013-2014. Figure 2 shows CSTDM v2.0 model system operation, six types of input data, five models, and seven formats of model outputs are described.

This model uses a Traffic Analysis Zone system of 5,474 zones. The zone boundaries were adjusted to accommodate the 2010 Federal Census zone system and to accommodate areas of major growth. The road and transit networks for the base year were updated to reflect 2010 conditions. The CSTDM road network includes over 125,000 nodes and 325,000 links and its transit network was developed using Google Transit platform. Synthetic population was generated with US Census, American Community Survey and other sources of data. Total employment per place was computed using ACS Journey to Work Data, ACS Equal Employment Opportunity, and Longitudinal and household Dynamics (OnTheMap).

This model consists of five demand models including:

- A Short Distance Personal Travel Model (for intra-California trips) (SDPTM);
- A Long Distance Personal Travel Model (for intra-California trips) (LDPTM);
- A Short Distance Commercial Vehicle Model (for intra-California trips) (SDCVM);
- A Long Distance Commercial Vehicle Model (for intra-California trips) (LDCVM);
- An External Vehicle Trip Model (for trips with origin and/or destination outside California).

The first and second models forecast intra-California passenger travel demands, and synthetic population is assigned to either SDPTM or LDPTM. The SDPTM is a tour-based travel forecasting model. The 2012 CHTS and 2010 Federal Census Journey to Work Surveys were used to calibrate sub-components of this model. In summary, this model has six main components, including Long Term Decision (car ownership or driving license), Day pattern (number / purpose of tours), Primary destination (destination of primary stop on tour), Tour mode (combination of modes for trip modes), Secondary Destination, and Trip modes. The LDPTM component was developed at household level, estimated from 2012 CHTS data. This model has five sub-models: Travel Choice model (business, commute, recreation, visiting friends, and relatives),

Party Formation Models (consisting member of long distance trip), Tour property models (Tour duration, travel day status, time of travel), Destination choice models, mode choice models.

The third and fourth models produce the travel demand estimates of commercial vehicles. The SDCVM is a microsimulation tour-based model, originally developed by HBA Specto for Calgary and Edmonton in Canada, adapted to California. The Commodity Flow Survey data were used to calibrate this model and incudes all sectors of the economy (i.e., industrial, wholesale, retail, service, transport, and so forth). The LDCVM was originally developed for the CSTDM09, and applied in CSTDMv2.0. The PECAS (Production, Exchange, and Consumption Allocation System) modeling framework are used to develop a computer-based model of the California spatial economic system. The 2008 PECAS model output created an initial year 2008 weekday long distance commercial vehicle OD matrix at TAZ level, then scenario-based growth factors were applied for the future years. The final component is a disaggregate microsimulation model, called External Vehicle Trip Model, to forecast the trips from and to external stations. This model has 51 external stations classified into six districs: California/Oregon border, northern part of the California/Nevada border, southern part of the California/Nevada border, California/Arizona border, California/ Mexico border and the ports.

These models produce outputs that are trip lists, trip tables, loaded network (vehicles on roadways), travel times and costs, summary travel statistics, maps and graphs of different metrics. This includes mode-splits by interregional and intraregional geographies. In this project, we received daily Trip tables containing 91,077,692 trips that are distributed on 297,746,116 OD pairs between 5,454 Traffic Analysis Zones.

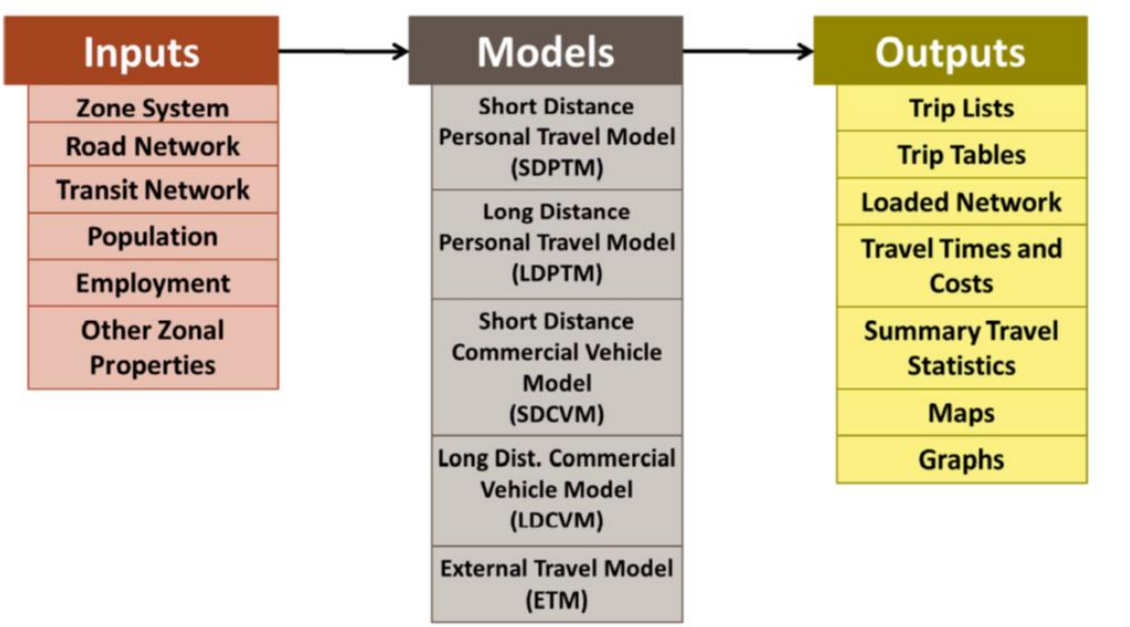


Figure 2. CSTDM Modeling Framework (reproduced from California Statewide Travel Demand Model, Version 2.0 Model, Overview p 1-1)

6. Twitter Data

Twitter data are the primary source of information for this project. Twitter users are clearly not a representative sample of the residents of California or any particular regions, mainly because it is much more popular among people under the age of 30, which account for roughly 23% of the American internet-using population in 2014. However, its usage is increasing in every age group, and it allows to collect large amount of data for much longer periods and at a lower cost than any traditional survey design could. This indicates that Twitter is not the ideal source for studying all aspects of travel behavior analysis but it can be used as a supplementary source for conventional data. This section describes the anatomy of tweets, Twitter data collection, and Twitter trip extraction methods.

6.1. Anatomy of a Tweet

Each tweet has 73 fields including ID, text, time, retweets, followers, language, place, source, and a variety of other information. Table 1 provides a detailed record of this information. In this project, user and place fields are the two important fields, because these provide key information to extract Twitter trips. The Twitter user profile contains each person's unique user ID that is used to determine if some tweets were produced from the same user. The place field contains 6-decimal geographical coordinates, and it is used to extract Twitter trips, and match it with geographical subdivisions such as Public Use Microdata Area, Traffic Analysis Zone, and Census Block Groups.

Table 1. Twitter data structure (attributes)

Field Name		Example
_id		558b5b74592a6e42948exxxx
contributors		null,
coordinates	coordinates	[-118.19xxxx,33.98xxxx],
	type	Point
created_at		"Thu Jun 25 01:37:59 +0000 2015",
entities	hashtags	[]
	symbols	[]
	trends	[]
	urls	[]
	user_mentions	[]
favorite_count		0
favorited		FALSE
filter_level		low
geo	coordinates:	[33.98xxxx,-118.19xxxx]
	type	Point
id		"61388380514995xxxx"
id_str		"61388380514995xxxx"
in_reply_to_screen_name		null
in_reply_to_status_id		null
in_reply_to_status_id_str		null
in_reply_to_user_id		null
in_reply_to_user_id_str		null
lang		en
place	attributes	
	bounding_box	coordinates: [[[-118.20xxxx,33.97xxxx],[-118.20xxxx,33.99xxxx],[-118.16xxxx,33.99xxxx],[-118.16xxxx,33.97xxxx]]],
	type	"Polygon"
	country	"United States"
	country_code	"US"
	full_name	"Maywood, CA"
	id	"8a8b2699803bxxxx"
	name	"Maywood"
	place_type	"city",
	url	https://api.twitter.com/1.1/geo/id/8a8b2699803bff27.json
possibly_sensitive		false
retweet_count		0
retweeted		false
source		"\u003ca href=\"http://twitter.com/download/android\" rel=\"nofollow\"\u003eTwitter for Android\u003c/a\u003e"
text		"My favorite pie is any pie."
timestamp_ms		"143519627xxxx"
truncated		false
user	contributors_enabled	false
	created_at	"Sun May 11 10:24:41 +0000 2014"
	default_profile	true
	default_profile_image	false
	description	null
	favourites_count	1
	follow_request_sent	null
	followers_count	4
	following	null
	friends_count	10
	geo_enabled	true
	id	"248964xxxx"

	id_str	"248964xxxx"
	is_translator	false
	lang	"en"
	listed_count	0
	location	""
	name	"feeduncensored"
	notifications	null
	profile_background_color	"C0DEED"
	profile_background_image_url	"http://abs.twimg.com/images/themes/theme1/bg.png"
	profile_background_image_url_https	"https://abs.twimg.com/images/themes/theme1/bg.png",
	profile_background_tile	false
	profile_banner_url	"https://pbs.twimg.com/profile_banners/2489646553/1434848402",
	profile_image_url	"http://pbs.twimg.com/profile_images/612423738261254144/YxKSDD6I_normal.jpg",
	profile_image_url_https	"https://pbs.twimg.com/profile_images/612423738261254144/YxKSDD6I_normal.jpg",
	profile_link_color	"0084B4"
	profile_sidebar_border_color	"C0DEED"
	profile_sidebar_fill_color	"DDEEF6"
	profile_text_color	"333333"
	profile_use_background_image	true
	protected	false
	screen_name	"feeduncensored"
	statuses_count	8
	time_zone	null
	url	null
	utc_offset	null
	verified	false

6.2. Deriving a trip from Tweets

Because Twitter data provide user ID and time, it is possible to select pairs of the consecutive tweets from the same user. For these pairs of tweets, four attributes of potential Twitter trips can be computed that are:

1) Euclidean distance;
2) Time difference between tweets;
3) Route distance; and
4) Estimated trip duration.

Figure 2 shows two tweets from the same Twitter user, and what can be computed from those tweets. This person posted a tweet at 8:03 AM October 9th 2015 in the University of California Santa Barbara, and about an hour later, he/she posted another tweet in downtown Santa Barbara.

The Euclidean distance between a pair of tweeting locations was 15.468 km, computed with GIS software, and their time difference was 63 minutes. Using Google Map API is possible to compute route distance and trip duration between those tweeting locations, 17.381 km and 18 minutes, respectively. These four trip attributes can be computed for all of the pairs of consecutive tweets (potential Twitter trips), and will play pivotal roles in extracting Twitter trips.

TWEETID	LON	LAT	TIME	COUNTRY	PLACE	PLACETYPE	USER_STRID	SOURCE
652333501165862912	-119.715xxxxx	34.422xxxxx	Fri Oct 09 04:03:21 +0000 2015	US	Isla Vista	city	id842252712	Twitter for iPhone
652349460257726465	-119.853xxxxx	34.413xxxxx	Fri Oct 09 05:06:46 +0000 2015	US	Santa Barbara	city	id842252712	Twitter for iPhone

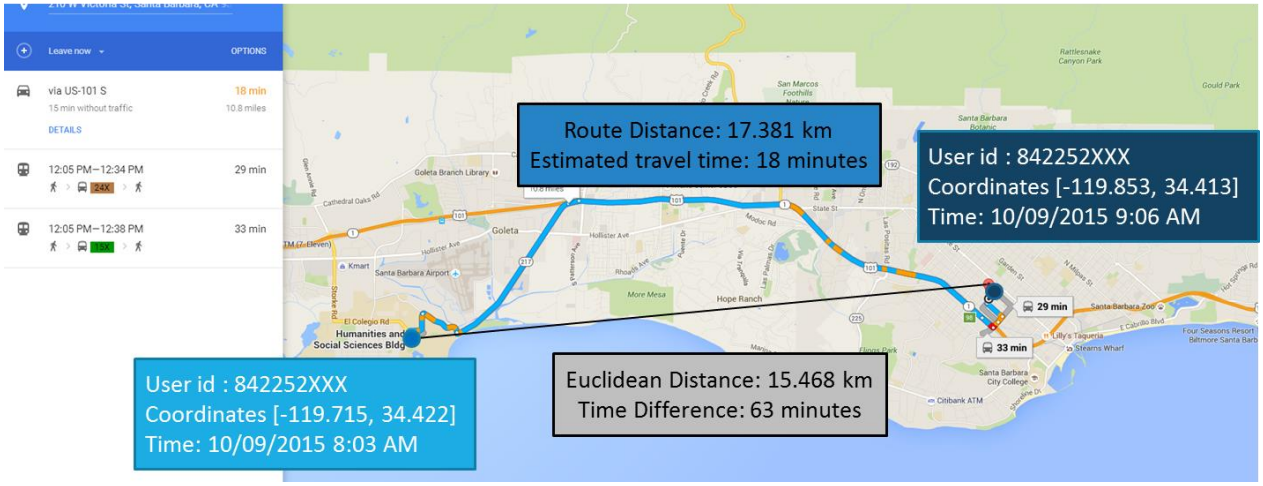


Figure 3. Key facts of Twitter data

6.3. Large-scale Data Collection

Twitter provides streaming APIs (application program interface), which allows developers to retrieve the tweets with given conditions (request parameters) including String length, language, User ID, phrase, location, followers, replies, etc. In order to retrieve the tweets needed for this project, we need to use the location parameter for obtaining the geo-tagged tweets generated within California. Therefore, we use the bounding box of [West: -124.0, South: 32.0, East: -114.0, North: 42.5], covering entire area of state of California and Nevada (Figure 4). With this API, the daily limit for Twitter data streaming is up to one percent of total number of tweets (i.e., approximately 5 million tweets per day). We developed Twitter data harvester with Python code, and made it connect to MongoDB server 3.2x because it provides the functionality needed to store and manage large volumes of data. We started our data collection on June 25th, 2015 and

finished it on December 15th, 2015. Although about 90 million geo-tagged tweets were collected in this period, only 8,285,593 tweets contain geographical coordinates (longitude, latitude). These tweets were generated from 437,095 unique users, and the average number of tweets per person was 18.96 for 6 months, and ranged from 1 to 54,469. Table 2 shows frequency tables of all geo-tagged tweets by day of week, month and sources of tweets. Although there were slightly more tweets per day on weekends, the tweets are almost equally distributed across days. By month, August seems to be the most popular period for Twitter users to generate geo-tagged tweets, followed by October and July. There were significantly lower numbers of tweets collected in June and December due to the duration of data collection. More than 40 percent of geo-tagged tweets were generated from Instragram, followed by TweetMyJOBS and smartphone applications such as iPhone and Android. Figure 5 illustrates all of the geo-tagged tweets, larger number of tweets were found in large metropolitan areas including Los Angeles and San Francisco. These tweets were matched with California Census block group zones (23,092 units), and Figures 6 and 7 describe number of geo-tagged tweets (red) and density of geo-tagged tweets (blue) in each zone, respectively. Although more tweets were collected in larger zones than the zones in metropolitan areas, the opposite pattern was found in geo-tagged tweets' density; higher tweet density is found in zones of Los Angeles, San Diego, and San Francisco metropolitan areas.

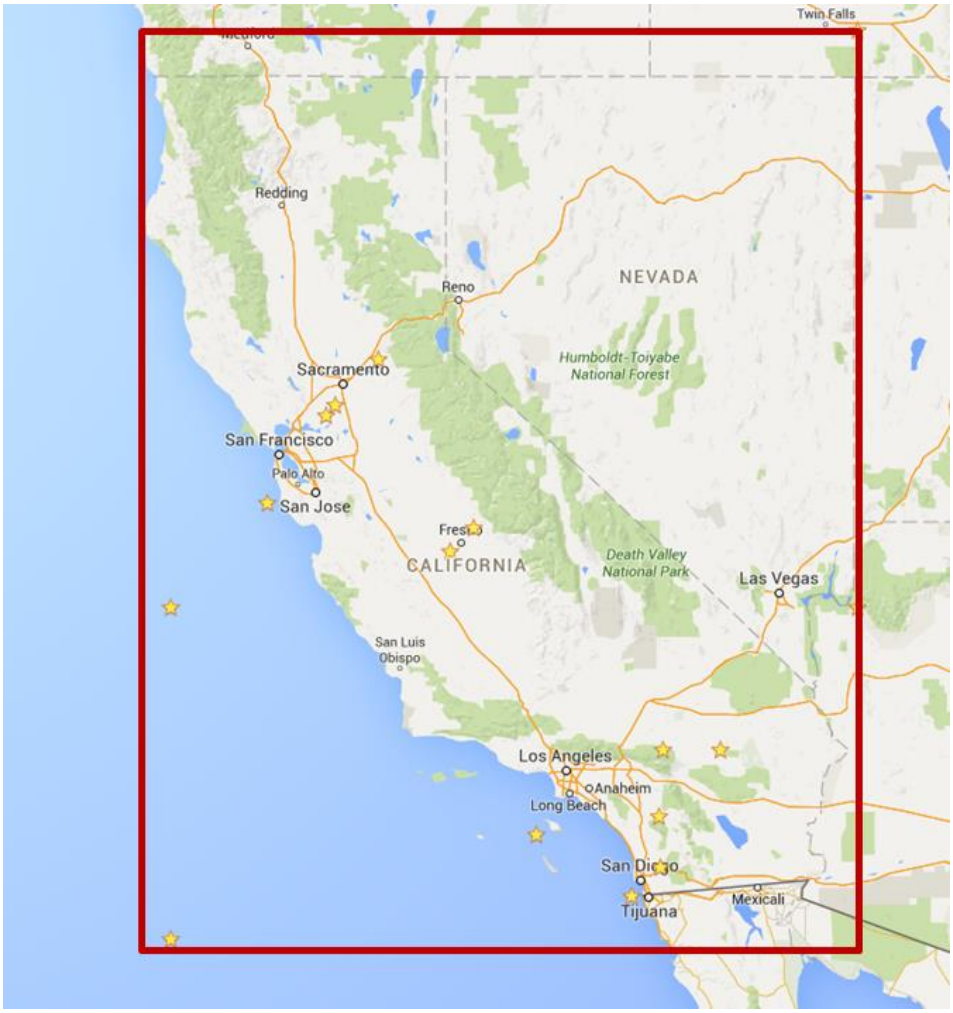


Figure 4. Map of bounding box, and number of geotagged tweets

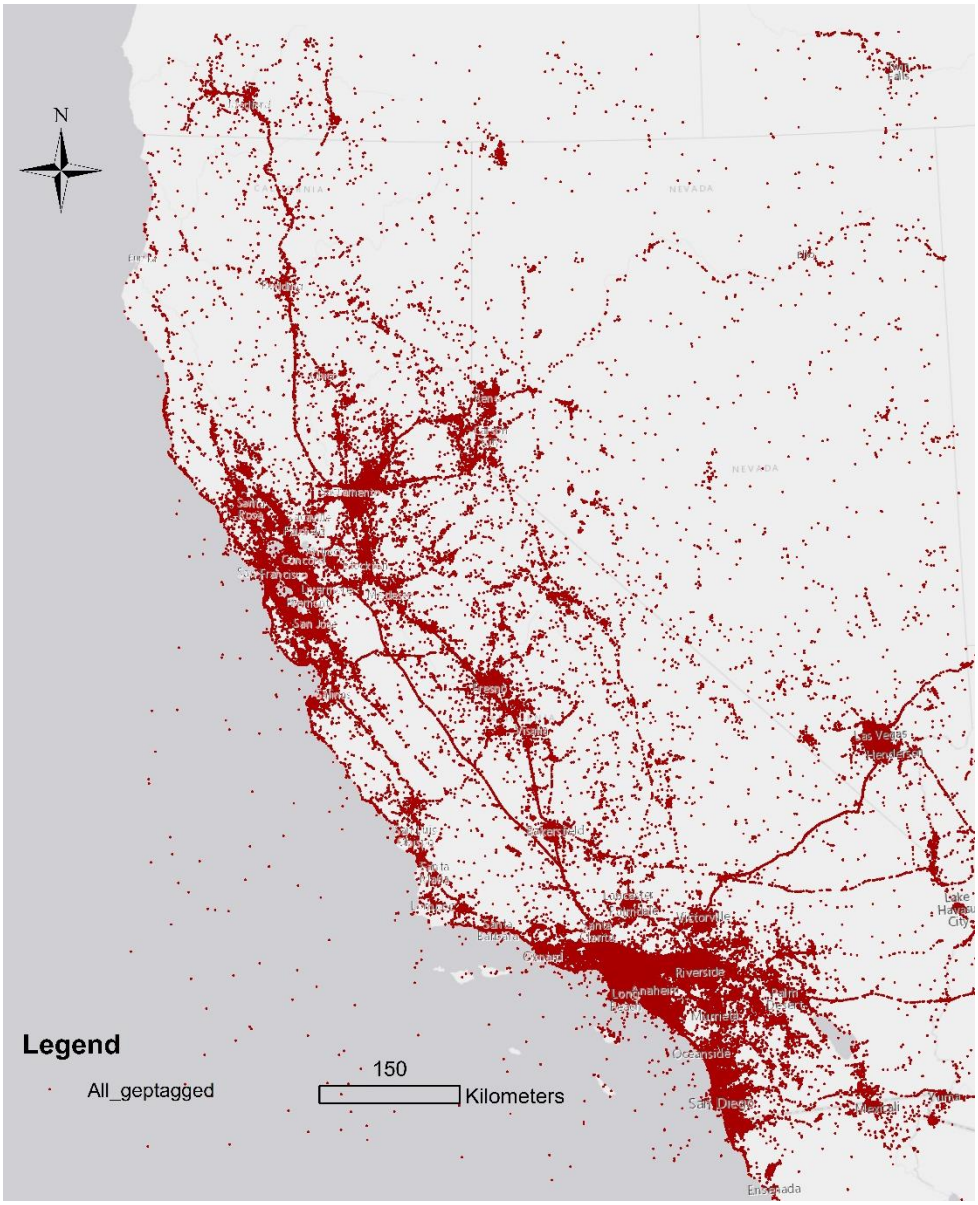


Figure 5. All Geo-tagged Tweets

Table 2. Frequency table of day of week, month, and sources

Day of Week	Frequency	Percent	Month	Frequency	Percent	Sources	Frequency	Percent
Sunday	1,288,877	15.6	June	255,293	3.1	Instagram	3,453,541	41.7
Monday	1,123,302	13.6	July	1,330,253	16.1	TweetMyJOBS	1,412,359	17.0
Tuesday	1,104,869	13.3	August	1,772,038	21.4	Twitter for iPhone	612,004	7.4
Wednesday	1,105,879	13.3	September	1,146,204	13.8	Foursquare	501,432	6.1
Thursday	1,162,168	14.0	October	1,603,761	19.4	Twitter for Android	497,640	6.0
Friday	1,199,046	14.5	November	1,489,842	18.0	SafeTweet by TweetMyJOBS	411,292	5.0
Saturday	1,301,452	15.7	December	688,202	8.3	Others	1,397,325	16.9

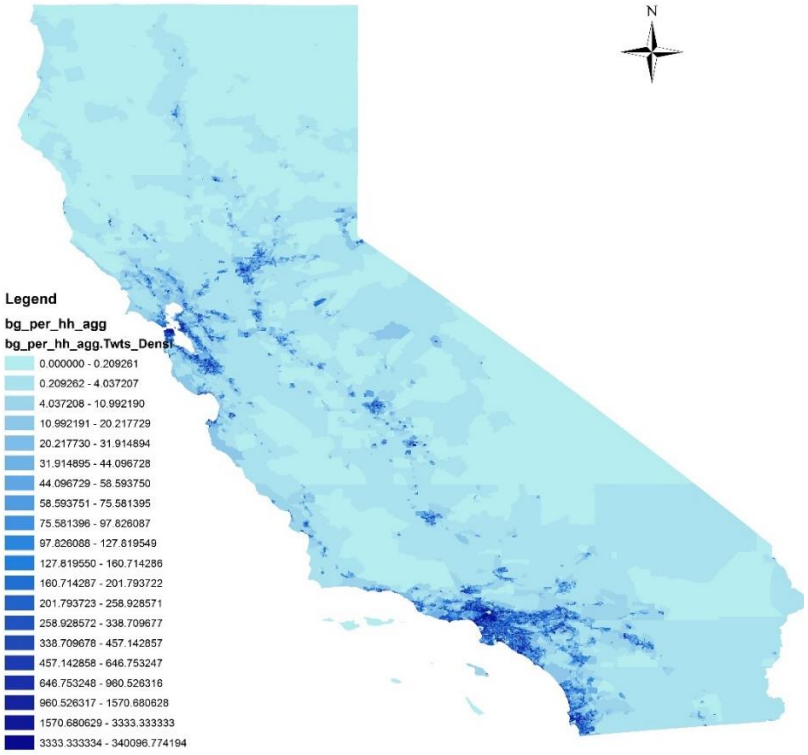


Figure 6. Number of geo-tagged Tweets in Census Block Groups

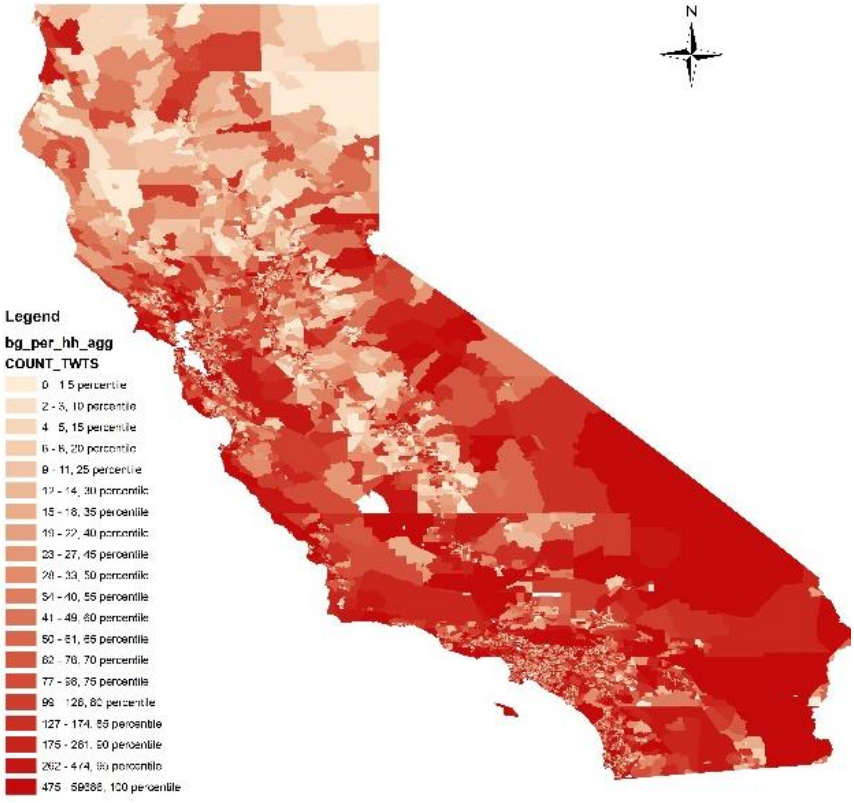


Figure 7. Geo-tagged Tweets' Density in Census Block Groups

6.4. Descriptive Analysis of All geo-tagged Tweets

An annual survey for American's internet usage partially reveals Twitter users' demographic. It does, however, reveal who uses geo-tagging functions, where they use these functions, and do not provide information about popular travel modes. In order to better understand our input data, we compare it with synthetically generated population, their travel behaviors, and land use characteristics with a statistical model.

As discussed, synthetic population is generated with PopGen software along with Census data and California Household Travel survey data. Land use characteristics are enumerated with business establishments dataset (National Establishment Time Series, 2013). Because this dataset contains geographical coordinates of all the business establishments in California, we are able to identify business employment centers with Kernel Density estimation, and compute the distance to the closest centers from each block group and percentage of center area in each block group. Figure 8 shows the business employment center areas in California, and suburban and exurban area classified based on proximities to centers and percentage of centers in each block group (more details about this are provided in the University of California Transportation Center and Caltrans project report: Task # 2851, Title: Business Establishment Survival and Transportation System Level of Service, Davis et al., 2016)

To study the correlation between the amount of tweets and land use characteristics we also develop a regression model. The number of all geo-tagged tweets per Census block group is used as a dependent variable, and explanatory variables are the number of people in age groups, males, number of trips per person by modes, number of business establishments per block groups, distance to the closest business center, and percentiles business center area per block groups. Because there are 513 zones having zero tweets, we use Tobit model to account for the censored distribution.

As a result, we found negative partial effect at the number of people in age group 0-14, 50-64, and 65+, indicating there is a lower number of geo-tagged tweets in the area of very young and

older population (Table 3). This result corresponds to the survey based demographic characteristics of Twitter users. In addition, the areas with more males show also a larger amount of geo-tagged tweets; there is one more tweet if there are around 10 more males in a block group. In terms of travel behavior of these residents, we found more tweets in the areas where their residents' travel behaviors are oriented to walking trip, driving alone, flying, and taking other modes. Moreover, more tweets were found in block groups with service industries (such as public administration and armed force, arts, entertainment, recreation, accommodation and food services, information, wholesale and retail trade), and the area of agriculture, forestry, fishing and hunting industries. On the other hand, lower numbers of tweets are found in block groups with primary and secondary sector of the economy area (such as mining, utilities, construction, manufacturing) and the areas of several types of service industries (transportation and warehousing, health care, other services (except public administration), education, and finance, insurance, real estate and rental and leasing). Lastly, there are more tweets in block groups that are closer to the business employment center area, and lower proportion of the block group area covered by centers.

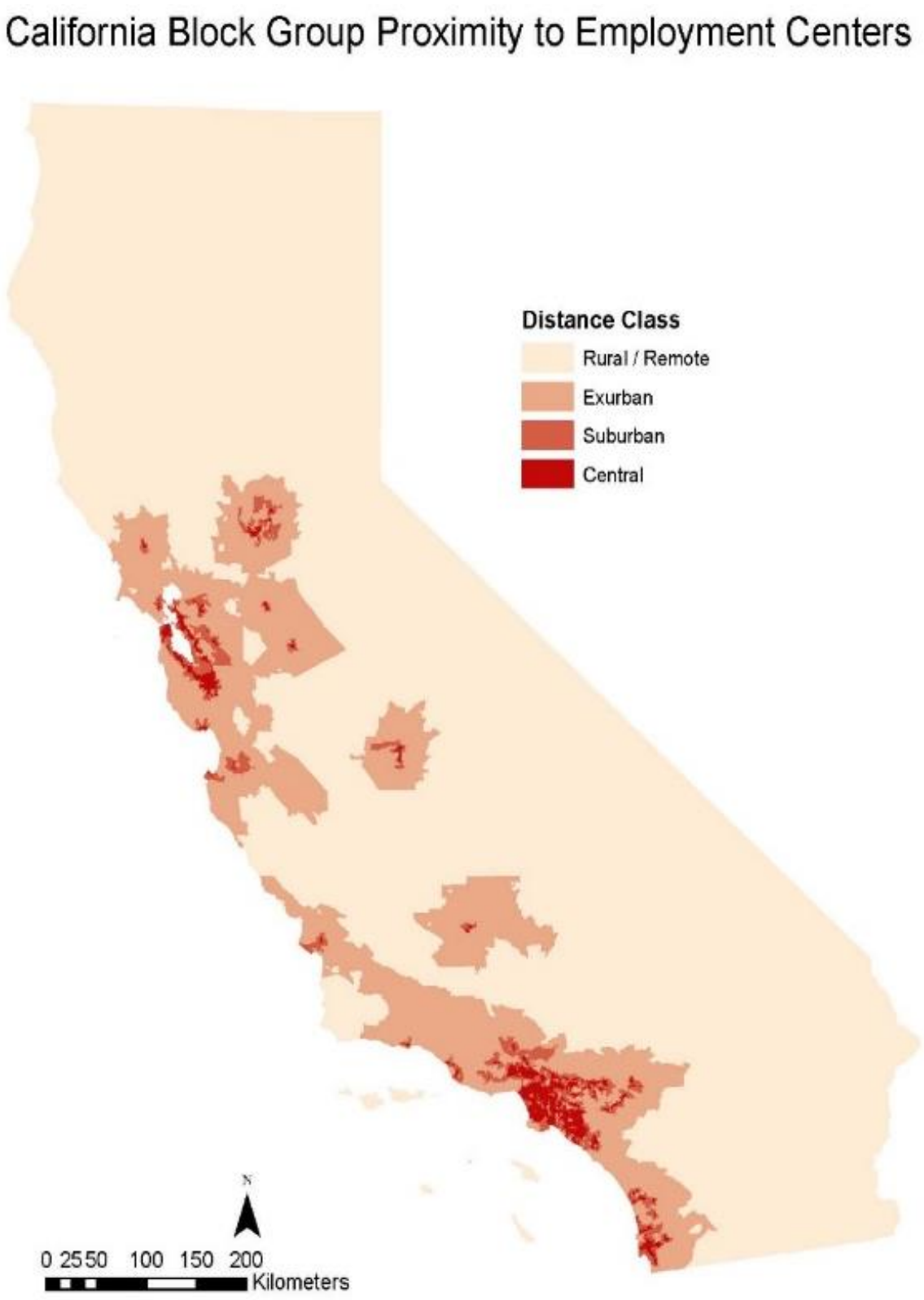


Figure 8. Business Employment Centers in California

Table 3. Descriptive Analysis Output

Independent Variables		Partial Effect		Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Number of people in age groups per block group	Age 0-14	-0.095	***	0.024	-3.900	0.000	-0.142	-0.047
	Age 15-24	0.026		0.026	1.010	0.314	-0.024	0.076
	Age 25-39	0.008		0.022	0.380	0.704	-0.034	0.051
	Age 40-49	-0.027		0.030	-0.900	0.368	-0.085	0.032
	Age 50-64	-0.077	***	0.025	-3.010	0.003	-0.126	-0.027
	Age 65+	-0.060	***	0.022	-2.740	0.006	-0.103	-0.017
Number of males in block groups		0.115	***	0.028	4.160	0.000	0.061	0.169
Number of trips per person by modes	Walking	274.331	***	26.415	10.390	0.000	222.559	326.103
	Biking	-166.159	*	88.583	-1.880	0.061	-339.778	7.460
	Driving alone	11.745		14.034	0.840	0.403	-15.762	39.251
	Driving as passenger	-89.825	***	18.786	-4.780	0.000	-126.646	-53.005
	Airplane	26.196		369.562	0.070	0.944	-698.131	750.523
	All other modes	-404.452	***	63.849	-6.330	0.000	-529.594	-279.310
Number of business establishments in block groups	agriculture, forestry, fishing and hunting	1.251	***	0.439	2.850	0.004	0.391	2.112
	mining	-13.820	***	2.720	-5.080	0.000	-19.151	-8.489
	utilities	-17.435	***	2.585	-6.740	0.000	-22.502	-12.368
	construction	-2.808	***	0.287	-9.780	0.000	-3.371	-2.245
	manufacturing	-0.776	**	0.344	-2.260	0.024	-1.449	-0.102
	wholesale trade	1.766	***	0.342	5.170	0.000	1.096	2.435
	retail trade	1.433	***	0.236	6.070	0.000	0.970	1.896
	transportation and warehousing	-2.609	***	0.394	-6.620	0.000	-3.382	-1.837
	information	4.196	***	0.487	8.620	0.000	3.241	5.150
	professional, scientific, management, administrative and wastemanagement services	0.648	***	0.134	4.830	0.000	0.385	0.910
	health care	-1.343	***	0.140	-9.600	0.000	-1.617	-1.069
	arts, entertainment, recreation, accommodation and food services	11.046	***	0.506	21.820	0.000	10.054	12.038
	other services (except public administration)	-2.633	***	0.283	-9.310	0.000	-3.188	-2.079
	finance, insurance, real estate and rental and leasing	-0.385	*	0.230	-1.670	0.094	-0.836	0.066
	public administration and armed force	23.153	***	0.593	39.060	0.000	21.991	24.315
	educational services	-1.576	*	0.916	-1.720	0.085	-3.371	0.219
Distance to the closest center area from each block group		0.000	***	0.000	-5.720	0.000	-0.001	0.000
Percentile of center area in each block group		-0.402	***	0.060	-6.710	0.000	-0.519	-0.284

Note: ***, **, * ==> Significance at 1%, 5%, 10% level.

7. Twitter Trip Extraction

We tested a few trip extraction methods in this project. Gao et al.(2014) developed a first version of the algorithm for extracting Twitter trips in Southern California. However, this algorithm is not able capture long distance travel within the whole state of California. Also, the existing algorithm did not use all of the computable trip attributes. In this project, we use this extraction algorithm and develop new extraction algorithms, then compare their outputs with the Origin-Destination matrix produced by California Statewide Travel Demand model (CSTDM). In this way we can select the best extraction algorithm to be used in this project.

7.1. Twitter Trip Extraction Rules

Figure 8 shows the three trip extraction rules tested in this project. First, we only use the pairs of the tweets from mobile devices regardless of rules. In this way, we were able to ensure that the locations of tweets reflect individuals' physical locations and avoid inclusion of social-robots' locations or default locations (hometown). In addition, we also use weekday trips because CSTDM's Origin-Destination matrix was created based on weekdays' trips.

The rule #1 is adopted from previous research (Gao et al., 2014; Lee et al. 2015). In essence, if there are two consecutive geo-tagged tweets belonging to different traffic analysis zones, and their time difference is less than 4 hours, then, those pairs of tweets are considered as trips from Twitter. However, this rule cannot extract the long distance Twitter trips that would take longer than 4 hours. Because the spatial scope of this project covers entire state of California, we need to develop different rules. Our first attempt (Rule #2) to capture long distance Twitter trips is to expand the time difference limit up to 24 hours. Moreover, instead of using geographical subdivision (such as TAZs) to extract Twitter trips, Euclidean distances between pairs of consecutive tweets; if two tweets' locations are further than 300 meters (maximum GPS device error boundary), we determined those pairs as Twitter trips. Although we assume that the pairs of tweets indicate the origins and destinations of trips, it is unclear that the time-stamps of those tweets indicate actual departure and arrival time. In addition, if the time differences between the pairs of tweets are much longer than actual trip durations between two tweeting locations,

it is more plausible that the users visited some other places between those two tweeting locations. In order to filter out the pairs of tweets whose time differences are unreasonably longer than actual travel time, we developed the third rule. In Rule #3, we use the ratio between the time differences of the pairs of tweets and the trip duration computed using Google Map API. If the time difference of two consecutive tweets divided by the estimated travel time between two tweets' locations is less than 10, we extracted these pairs as Twitter trips. As a result, we were able to extract 224,603, 483,283, and 169,849 Twitter trips, respectively. The Rule #2 turned out to be the most lenient rule and the Rule #3 was the strictest rule.

7.2. Spatial Scales in Different Options

In order to directly compare the extracted Twitter trips with model outputs, we have to match those trips into geographical subdivisions that are used in final products of the models (Table 4). The California Statewide Travel Demand Model uses traffic analysis zones (5,454 zones, 297,746,116 OD pairs between zones), and Synthetic population generation uses Census block groups (23,092 zones). Moreover, the amount of Twitter trips is much lower than model outputs. This creates bias in the OD matrix comparison due to too many zero cells in OD matrix from Twitter trips. Therefore, we need to aggregate OD matrices into higher zonal systems (less number of units and larger area). There are two options for spatial aggregation: Traffic Analysis District (1,008 units, 1,016,064 OD pairs between zones) and Public Use Microdata Areas (265 units, 70,225 OD pairs between zones). This was possible because TAZs are created using the US Census Blocks, which are also used to create Traffic Analysis Districts (TADs), and Public Use Microdata Areas (PUMAs) (US Census Bureau, Figure 9).

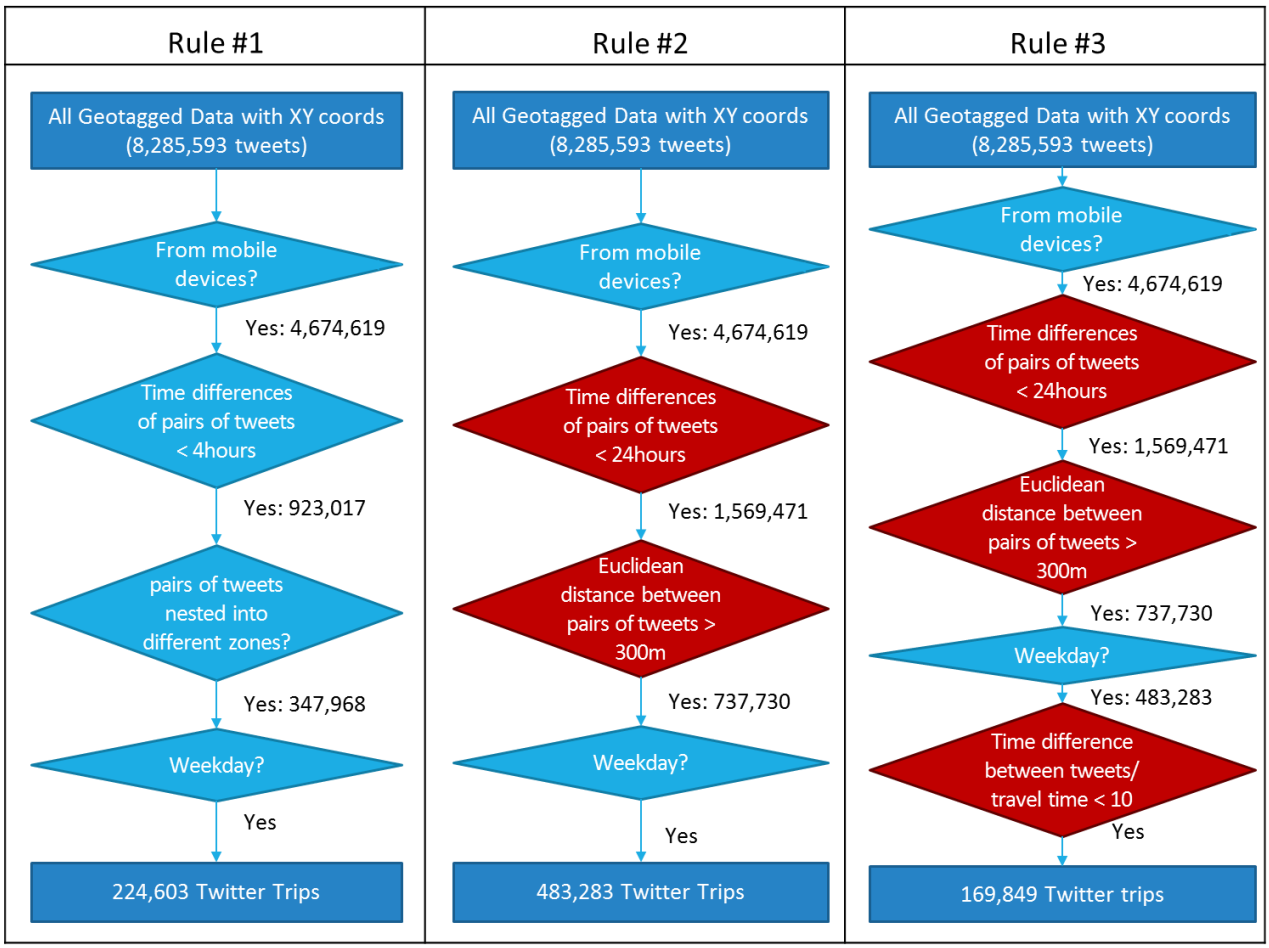


Figure 9. Three Twitter Trip Extraction Rules

Table 4. Spatial Scales in Different Options

Name of Dataset	Number of Units in California	Number of OD pairs	Our data
Census block	710,140	504,298,819,600	Census data
Census block group	23,092	533,240,464	Synthetic Population
Traffic Analysis Zones	5,454	297,746,116	CSTDM Model Output
Traffic Analysis District	1,008	1,016,064	
Public Use Microdata Area	265	70,225	

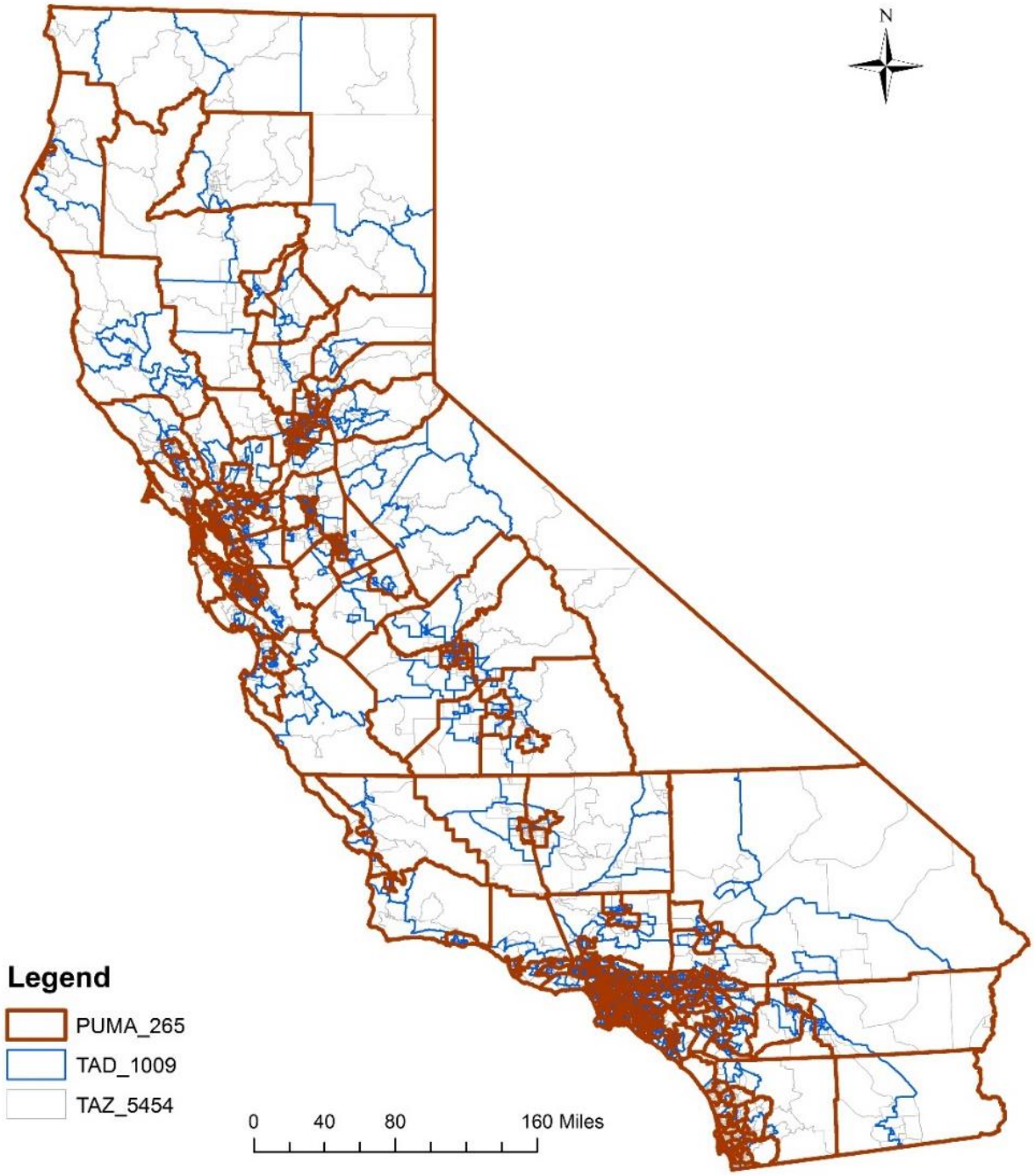


Figure 9. Spatial Zoning Systems

7.3. Comparisons of OD Matrices from Twitter and CSTDM

In order to decide the best rule to extract Twitter trips, we use Pearson correlation coefficients between the matrix from CSTDM model and the matrices from Twitter data. Because we use three spatial zoning systems, three OD matrices from CSTDM model (CSTDM TAZ, CSTDM TAD, CSTDM PUMA) are compared with the nine OD matrices from Twitter trips depending on spatial scales and Twitter trip extraction rules: Twitter TAZ from rule #1, #2, #3; Twitter TAD from rule #1, #2, #3; Twitter PUMA from rule #1, #2, #3. Table 4 shows correlation coefficients between OD matrices from CSTDM and Twitter trips. All of the coefficients were significant at the level of 0.001, the highest one was found in the relationship between CSTDM and Twitter trips extracted from Rule #3 at TAD zoning system (Table 5). However, the correlation coefficients are not very different from each other if those are compared in TAD or PUMA system. On the other hand, using TAZ system produced the lowest coefficients of correlation ranging between 0.157 and 0.159 regardless of the extraction rules used.

Table 5. Pearson Correlation Coefficients between OD matrices from CSTDM and Twitter trips

Spatial Scale (# of OD pairs)	Twitter trips from Rule #1	Twitter trips from Rule #2	Twitter trips from Rule #3
TAZ (297,746,116)	0.157***	0.159***	0.158***
TAD (1,016,064)	0.518***	0.490***	0.519***
PUMA (70,225)	0.516***	0.485***	0.503***

*** significant at the level of 0.001

Another important criterion to select the best extraction rule is producing the least number of zero-cells in OD matrix. Because the Rule #2 yielded the largest amount of Twitter trips, the OD matrices from the Twitter trips have the smallest number of zero-cells (Table 6). Unlike the correlation coefficients, number of zero-cells are very different depending on the extraction rules. Therefore, Rule #2 with PUMA system seems to create the most suitable OD matrix. Based on this test result, we use Twitter trips extracted by the Rule #2 in further analysis.

Table 6. Percentage of zero cells in CSTDM OD matrix and Twitter OD matrix

Spatial Scale	CSTDM	Twitter trips from Rule #1	Twitter trips from Rule #2	Twitter trips from Rule #3
TAZ (297,746,116)	86.8%	99.7%	99.2%	99.6%
TAD (1,016,064)	50.9%	95.9%	91.5%	94.8%
PUMA (70,225)	17.5%	80.6%	65.8%	72.3%

8. Twitter Trips vs CHTS and Synthetic population

In this chapter, we compare the extracted Twitter trips with California Household Travel Survey trip records and synthetic population generated in UCTC/Caltrans Task #2644: Spatial Transferability Using Synthetic Population Generation Methods. Although we only use weekday Twitter trips to select the best rule when we compare it with OD matrices from CSTDM output, in this chapter, we also included weekend Twitter trips because CHTS and Synthetic population data have weekend trips as well. In total, we use 737,016 Twitter trips in this analysis (254,142 Twitter weekend trips are added).

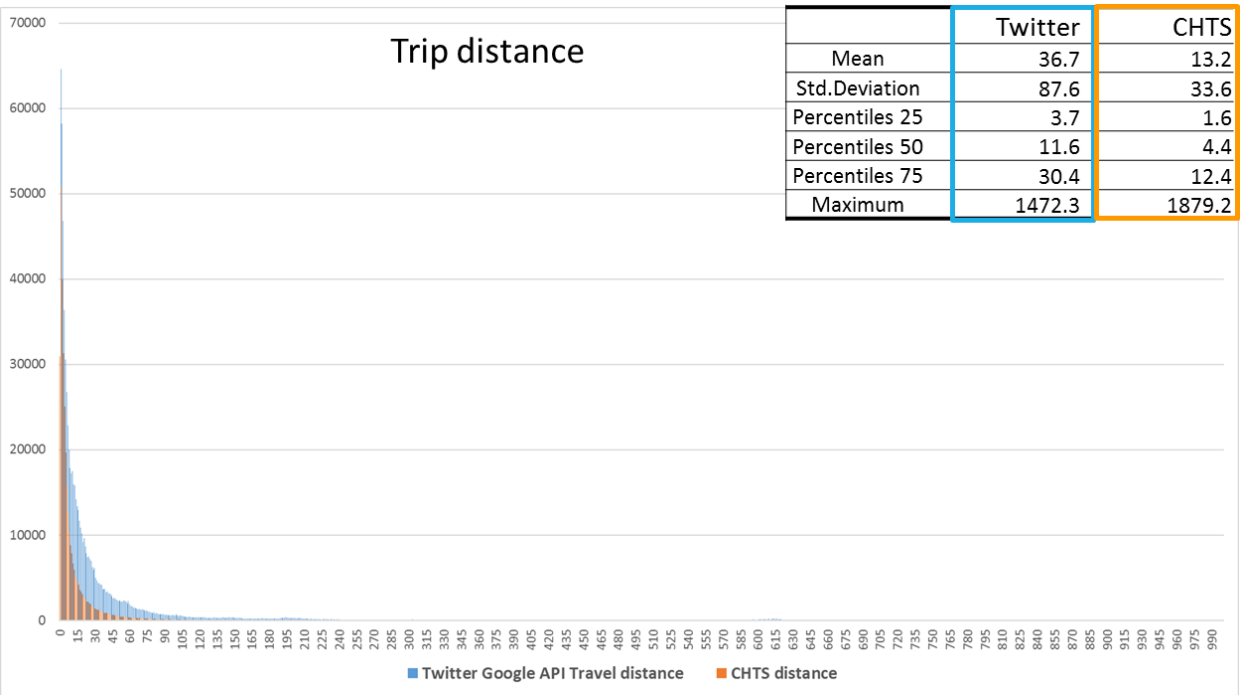
8.1. Twitter Trips and CHTS

Twitter trips have limited travel information; there is neither trip purpose nor activity type, companions, and travel modes. Therefore, the comparison between Twitter trips and CHTS trip records is also limited. However, trip distances and durations are available for both Twitter trips and CHTS data.

Figure 10 shows the descriptive statistics and histograms of both Twitter and CHTS trip distances. Twitter trips has much longer mean trip distances compared to CHTS trip records (23.5 km), and its standard deviation is also much higher than CHTS data. Overall, there are more Twitter trips than CHTS regardless of trip distance and this is due to the different size of trip records; Twitter trip records are twice larger than CHTS data. In order to avoid this bias, we created probability mass functions for both trip records, and overlaid it (Figure 11). It seems that survey methods are better to observe the short distance trips, but after 10 km, Twitter data provide higher chances to collect longer distance trips. This is presumably because there are missing locations between Twitter trips' origins and destinations. Although probability to observe Twitter trips whose distances are less than 1 km is much lower than CHTS, this is because the Twitter trip extraction rule filtered out the trips that are shorter than 300 meters.

Figure 12 shows the histograms of trip durations for Twitter trips and CHTS and descriptive statistics of their trip records. Twitter trips seem to have on average 10 minutes longer trip

durations than CHTS. However, their medians and 25 percentile are similar (1, 2.4 minutes difference, respectively). In terms of distribution of trip duration, Twitter trip has much smoother distribution than CHTS because Twitter trips' duration is computed by Google Map API algorithm. On the other hand, CHTS data has multiple peaks at 5, 10, 15, 20, 25, 30, and etc. minutes. This is because people tend to answer their travel time approximately in 5-minute intervals. When this histogram is converted into probability mass function, the same patterns are also found (Figure 13). Moreover, we also find the multiple peaks from CHTS produce unstable distribution between the peaks.



	Twitter	CHTS
Mean	36.7	13.2
Std.Deviation	87.6	33.6
Percentiles 25	3.7	1.6
Percentiles 50	11.6	4.4
Percentiles 75	30.4	12.4
Maximum	1472.3	1879.2

Figure 10 Descriptive Statistics and Histogram of both Twitter and CHTS Trip Distance

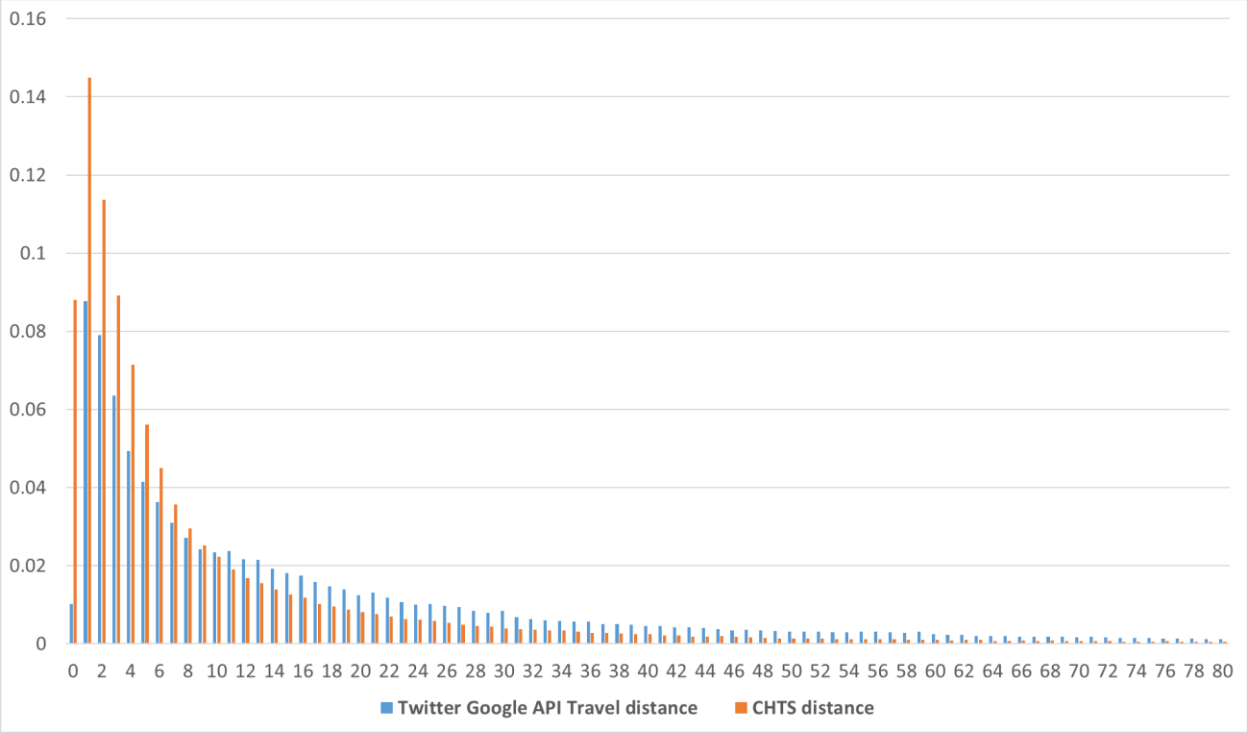
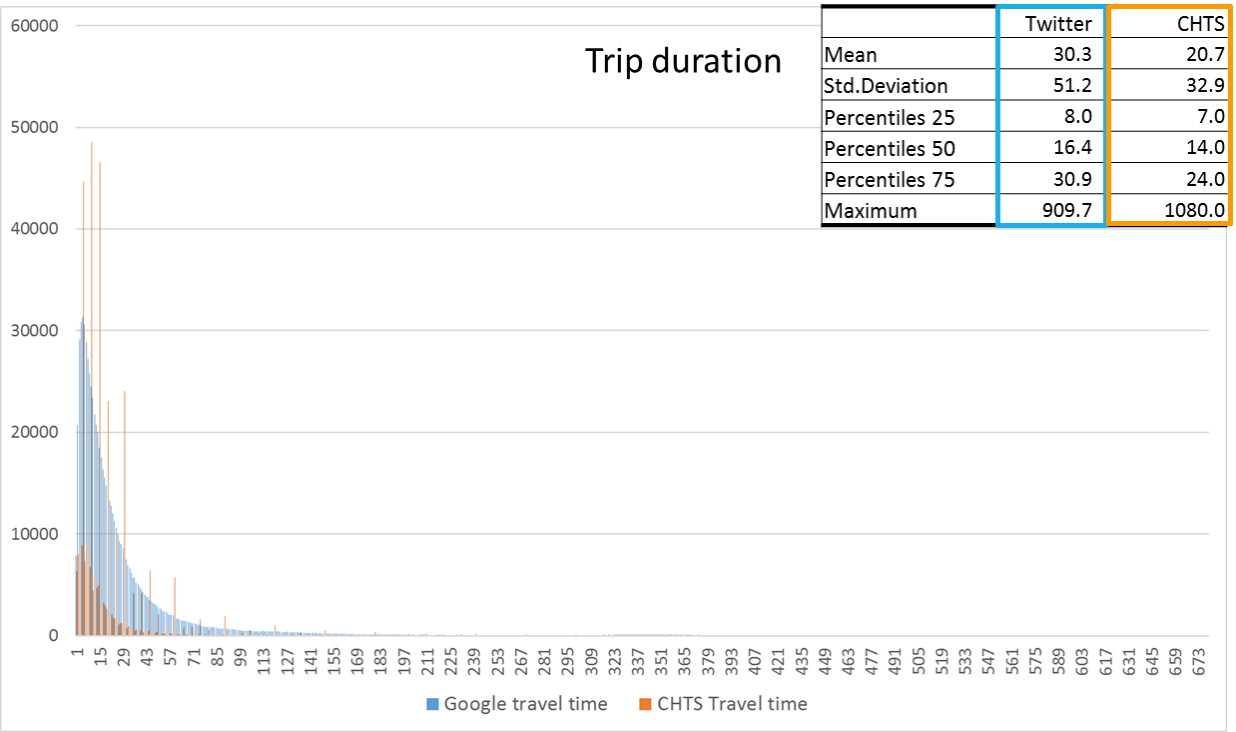


Figure 11. Probability Mass Functions for Trip distances (less than 80 km trips only)



	Twitter	CHTS
Mean	30.3	20.7
Std.Deviation	51.2	32.9
Percentiles 25	8.0	7.0
Percentiles 50	16.4	14.0
Percentiles 75	30.9	24.0
Maximum	909.7	1080.0

Figure 12 Descriptive Statistics and Histogram of both Twitter and CHTS Trip Duration

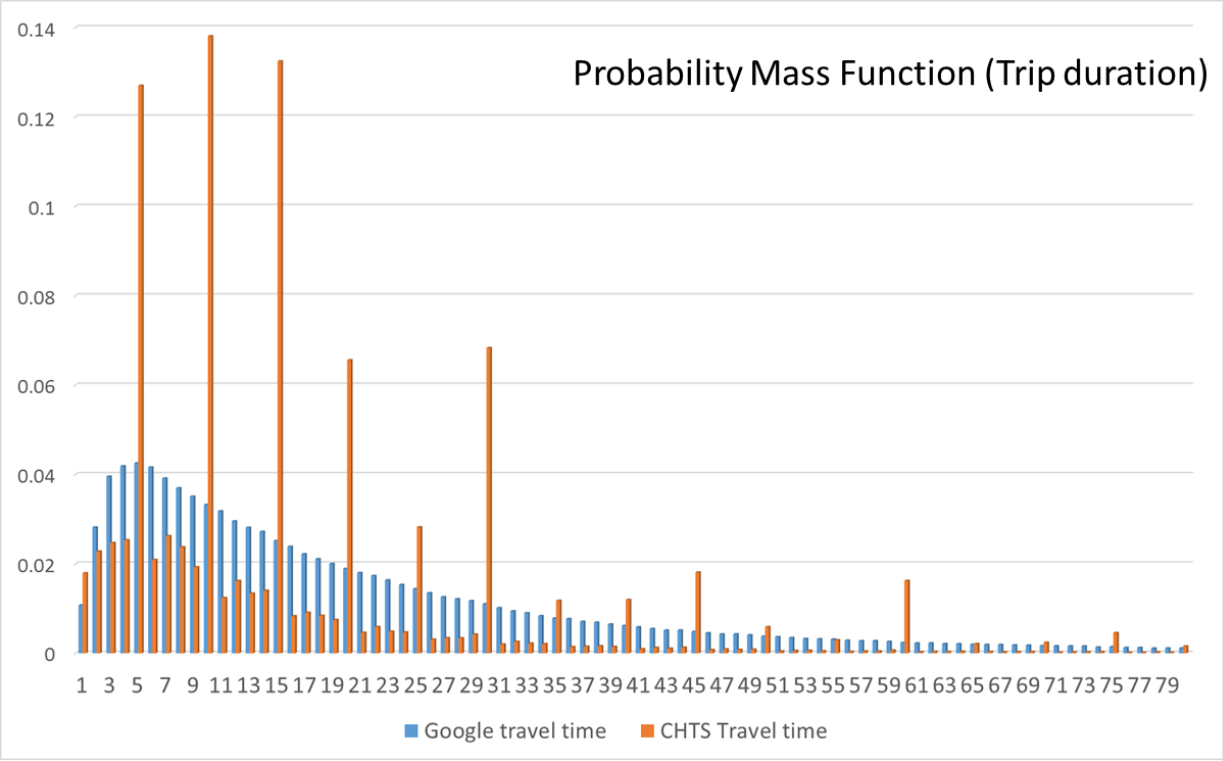


Figure 13. Probability Mass Functions for Trip duration (less than 80 minutes only)

8.2. Twitter Trips and Synthetic Population

As discussed earlier, the population synthesis produced the total number of trips and the sum of travel distances by households, and their residences at the Census block group (23,209 zones). Therefore, it is possible to enumerate the total number of trips and the sum of travel distances for each block group, and compare it with the extracted Twitter trips' attributes. These trips per block group do not exactly correspond to the number of trips originated to each block group, but it represents the trips from each block group's trips of residents. However, home is still one of the most important origins of daily trips and many trips tend to originate from a person's residence.

We use a Tobit regression model to understand the relationship between the Twitter trips and the number of trips by different modes per block group because the Twitter trips have a censored distribution at zero. The number of the Twitter trips was used as a dependent variable, and synthetic population's trips by different modes are the explanatory variables in this model. As a result, we found positive partial effect for the number of trips by walking, biking, driving alone, airplane, and all other modes and negative one was found from driving as passenger (Table 8). Another Tobit model was estimated with trip distance variables: dependent variable – Sum of Twitter trip distances in block group, independent variables – sum of trip distances from synthetic population. The significant partial effects from this model are very similar to the previous model except for Driving with others and all other modes (Table 9). Overall, we found more trips from Twitter data where people make more trips by walking, biking, driving alone, and airplane, but less trips with driving as passenger or driving with others.

Table 8. Estimated Partial Effect for the Number of Twitter Trips per Block Group

Independent Variables: Synthetic Population		Partial Effect		Standard Error	Z	Prob. |z|>Z*	95% Confidence Interval	
Number of trips in Block group	Walking	0.612	***	0.155	3.940	0.000	0.307	0.916
	Biking	1.829	***	0.272	6.730	0.000	1.296	2.362
	Driving alone	0.032	***	0.008	3.830	0.000	0.016	0.048
	Driving with Others	0.002		0.020	0.090	0.929	-0.038	0.042

	Driving as passenger	-0.062	***	0.009	-7.210	0.000	-0.079	-0.045
	Airplane	0.082	**	0.039	2.100	0.036	0.005	0.158
	All other modes	0.068	***	0.023	2.900	0.004	0.022	0.113

Note: ***, **, * ==> Significance at 1%, 5%, 10% level.

Table 9. Estimated Partial Effect for the Sum of Twitter Trips' Distance per Block Group

Independent Variables: Synthetic Population		Partial Effect		Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Number of trips in Block group	Walking	0.015	***	0.006	2.630	0.008	0.004	0.026
	Biking	0.146	***	0.019	7.550	0.000	0.108	0.184
	Driving alone	0.027	***	0.003	10.010	0.000	0.022	0.032
	Driving with Others	-0.065	***	0.010	-6.320	0.000	-0.085	-0.045
	Driving as passenger	0.000		0.004	0.070	0.942	-0.008	0.009
	Airplane	0.224	**	0.093	2.400	0.016	0.041	0.407
	All other modes	0.022		0.014	1.570	0.116	-0.006	0.050

Note: ***, **, * ==> Significance at 1%, 5%, 10% level.

8.3. Weekday vs Weekend in Twitter Trips

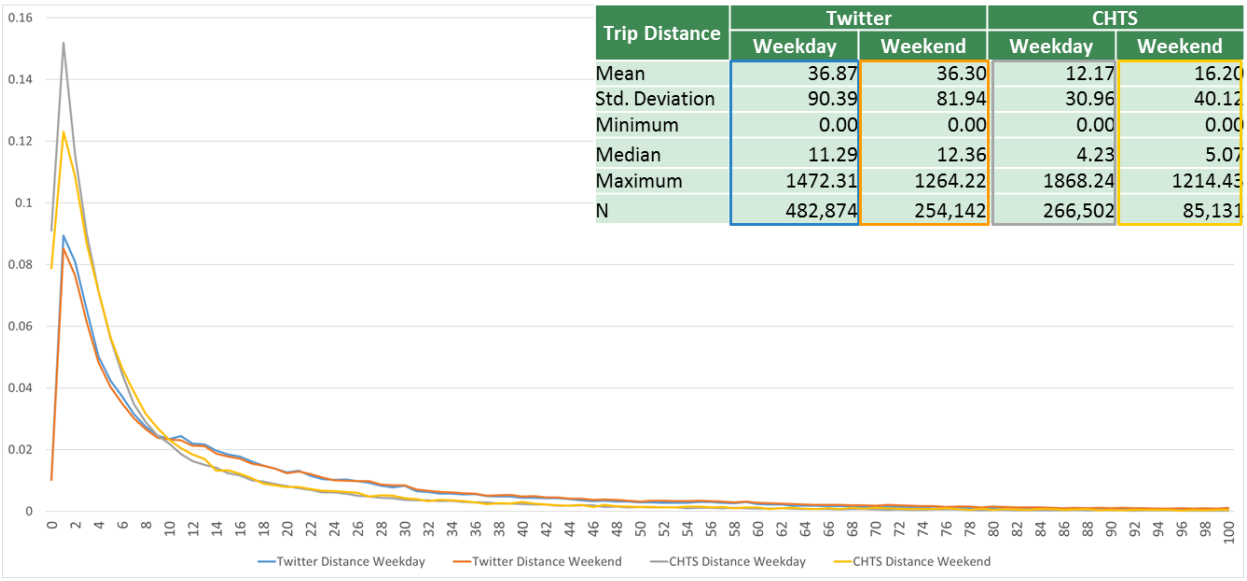
Because social media services are also popularly used during the weekend, we can use this data to explorer day-by-day dynamics of travel behavior (Lee et al, Forthcoming). In order to compare trip distances between weekday and weekend from two different sources, we created probability density functions and overlaid those (Figure 14). Similar to the weekday and weekend variability in descriptive statistics in CHTS data, the shapes of density functions are also quite different especially for the short distance trips. In other words, people make more short distance trips during the weekday and make more long distance trips in the weekends. However, it is very difficult to observe the day-by-day variability in Twitter trip records in both descriptive statistics and probability density functions. In terms of trip duration, it also has similar patterns with the trip distances. In CHTS trip records, people spend less time per trip (Figure 15). Like trip distance, there was not significant differences in trip duration in Twitter trip data.

Another way to examine day-by-day variability with Twitter trip data is to compare the number of Twitter trips with synthetic population computed trips. We use weekday and weekend Twitter trip per block group as dependent variables, and synthetic population's number of trips are independent variables; two models are estimated in the exactly same settings except for dependent variables. By comparing the significant independent variables between two models,

we can find the differences between weekday and weekend Twitter trips in terms of their spatial distribution. As a result, we were not able to find significant differences between weekday and weekend in Twitter data because the significant explanatory variables in two models are exactly the same except for all other modes (Tables 10, 11). However, this variable was almost significant at the level of 0.010 in the weekend model (p-value: 0.0107).

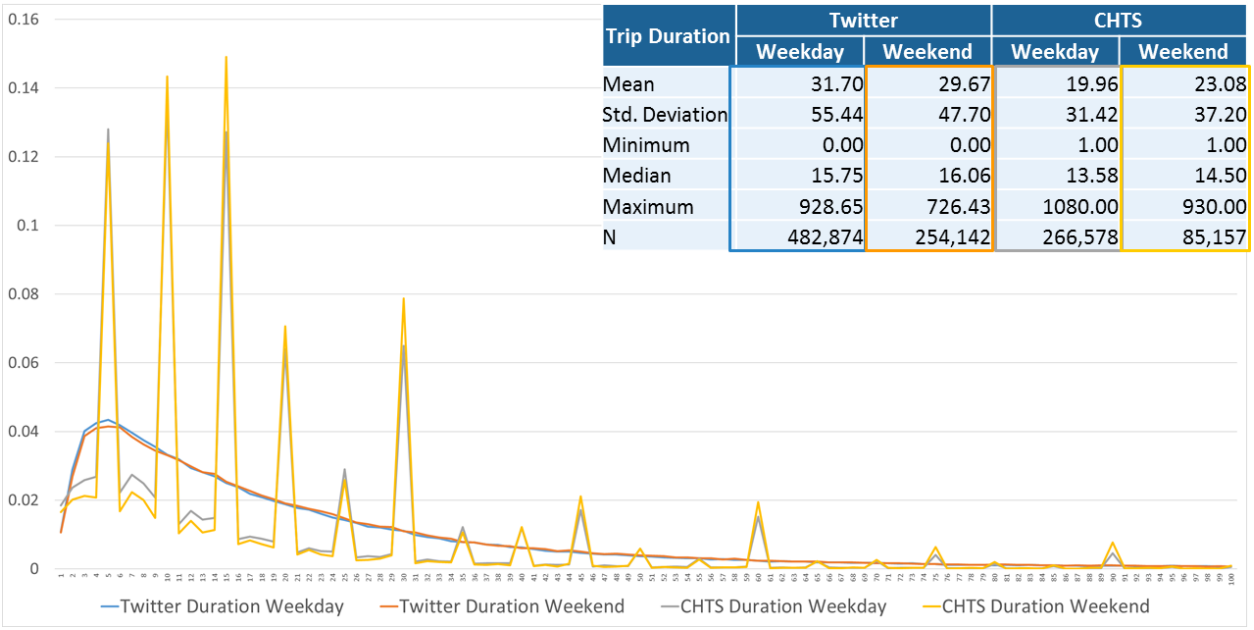
We are also interested in comparing Twitter trips during the weekday and the trips in special occasion such as Thanksgiving days. We select the Twitter trips in Thanksgiving Day period (October 28th to November 1st, 2015), and compare it with weekday Twitter trips data. Like the weekday and weekend comparison, we were not able to find significantly different patterns of Twitter trips between weekday and Thanksgiving Day. (Tables 10, 12). Although the magnitude of partial effects of explanatory variables are different, this is presumably due to the size of the Twitter trip data.

Overall, Twitter data do not seem to provide the functionality needed to observe day-by-day variability in terms of the number of trips per block group statewide.



Trip Distance	Twitter		CHTS	
	Weekday	Weekend	Weekday	Weekend
Mean	36.87	36.30	12.17	16.20
Std. Deviation	90.39	81.94	30.96	40.12
Minimum	0.00	0.00	0.00	0.00
Median	11.29	12.36	4.23	5.07
Maximum	1472.31	1264.22	1868.24	1214.43
N	482,874	254,142	266,502	85,131

Figure 14. Weekday vs Weekend Trip Distances in Twitter and CHTS data



Trip Duration	Twitter		CHTS	
	Weekday	Weekend	Weekday	Weekend
Mean	31.70	29.67	19.96	23.08
Std. Deviation	55.44	47.70	31.42	37.20
Minimum	0.00	0.00	1.00	1.00
Median	15.75	16.06	13.58	14.50
Maximum	928.65	726.43	1080.00	930.00
N	482,874	254,142	266,578	85,157

Figure 15. Weekday vs Weekend Trip Duration in Twitter and CHTS data

Table 10. Weekday Model for Comparison with Synthetic Population

Independent Variables		Partial Effect		Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Number of Weekday trips in Block group	Walking	0.010	***	0.004	2.710	0.007	0.003	0.017
	Biking	0.093	***	0.013	7.430	0.000	0.068	0.117
	Driving alone	0.017	***	0.002	9.910	0.000	0.014	0.021
	Driving with Others	-0.040	***	0.007	-6.090	0.000	-0.054	-0.027
	Driving as passenger	0.000		0.003	-0.060	0.954	-0.006	0.005
	Airplane	0.124	**	0.060	2.060	0.040	0.006	0.241
	All other modes	0.017	*	0.009	1.830	0.068	-0.001	0.034

Table 11. Weekend Model for Comparison with Synthetic Population

Independent Variables		Partial Effect		Standard Error	z	Prob. |z|>Z*	95% Confidence Interval	
Number of Weekend trips in Block group	Walking	0.006	***	0.002	3.040	0.002	0.002	0.009
	Biking	0.048	***	0.006	7.480	0.000	0.036	0.061
	Driving alone	0.010	***	0.001	11.470	0.000	0.008	0.012
	Driving with Others	-0.021	***	0.003	-6.110	0.000	-0.028	-0.014
	Driving as passenger	-0.001		0.001	-0.840	0.399	-0.004	0.002
	Airplane	0.076	**	0.031	2.470	0.014	0.016	0.137
	All other modes	0.008		0.005	1.610	0.107	-0.002	0.017

Table 12. Thanksgiving Day Model for Comparison with Synthetic Population

Independent Variables		Partial Effect		Standard Error	Z	Prob. |z|>Z*	95% Confidence Interval	
Number of Thanksgiving trips in Block group	Walking	0.002	***	0.000	3.380	0.001	0.001	0.002
	Biking	0.010	***	0.002	6.800	0.000	0.007	0.013
	Driving alone	0.003	***	0.000	12.510	0.000	0.002	0.003
	Driving with Others	-0.005	***	0.001	-6.090	0.000	-0.007	-0.003
	Driving as passenger	0.000		0.000	-0.990	0.323	-0.001	0.000
	Airplane	0.019	***	0.007	2.620	0.009	0.005	0.034
	All other modes	0.001		0.001	1.220	0.224	-0.001	0.004

9. Twitter Trips vs CSTDM output

9.1. Models for Matrix Comparison

Because the number of trips between zones depend on the spatial structure, the presence of spatial autocorrelation among zones is problematic in developing the conversion methods between Twitter and CSTDM daily OD trips. To address this within our conversion model, we construct several spatial lag variables, and then we use them as explanatory variables in a regression model and a latent class regression model. We use CSTDM ODs as a dependent variable and Twitter trips, land use, and zonal demographic characteristics as independent variables along with the spatial lag variables. In this way, we create a three-way comparison among three different sources of information about Origins and Destinations for cross-validation. Land use is described with indicators of business density and diversity. Demography is captured by population density. In this way the regression coefficient associated with a Twitter OD trip is the multiplier that needs to represent CSTDM OD trips from a zone and this is the net multiplier taking into account land-use effects. In other words, a unit contribution of a Twitter trip can be derived from a regression model while controlling for land use and other demographic variables of the residents in each zone.

9.1.1. Defining Spatial Lag Variables

Both OD-trips from Twitter and CSTDM model are spatially dependent, because the number of OD trips is correlated with the number of trips between neighboring Origins and Destinations; we address this using spatially lagged variables. General ways of defining spatial lag variables can be found in Anselin (1988). However, our OD-trips are doubly dependent upon space (Origins' and Destinations' Neighborhood). Figure 16 illustrates the method for defining spatial lag variables. The first image (a) represents an OD-trip that we want to compute its values for spatial lag variables. The images (b) and (c) show two ways of calculating spatial lag variables, which indicates the number of trips from an origin to the adjacent zones of the destination (O_Dadj), and the adjacent zones of the origin to a destination (Oadj_D). The last image (d) in Figure 16 illustrates the method to compute the third spatially lagged variable: the number of trips from the neighborhood area of the origin to the neighborhood area of the destination (Oadj_Dadj). We defined as neighbors all of pairs of zones with centroids that are located within three miles

(Euclidean distance between centroids of the zones is sufficient for this initial testing). For example, the zone located in the southern area of destination was not included as neighbor area for destination zone (Marked with a green asterisk mark on the maps (b) and (d)) because the distance between centroids of the two zones was longer than three miles. We use 10 miles radius because it was the median of CHTS trip distance. We computed these spatial lag variables for both Twitter trips and CSTDM trips, yielding six variables: T_O_Dad, T_Oad_D, T_Oad_Dad, C_O_Dad, C_Oad_D, C_Oad_Dad. We use these variables as explanatory variables in a regression model along with other land use and population characteristics. In order to define the neighboring zones of each zone, we need to define the centers of each zone, and compute the distance between zones. Instead of using artificial centers of zones (such as centroid), we use business employment centers for each zone, and route distances between these centers of the zones. Figure 17 shows the centers of PUMA zones in California that are enumerated based on business establishment employment data (NETS, 2013). Unlike the simple centroids of zones, all centers are located within each zone with this method.

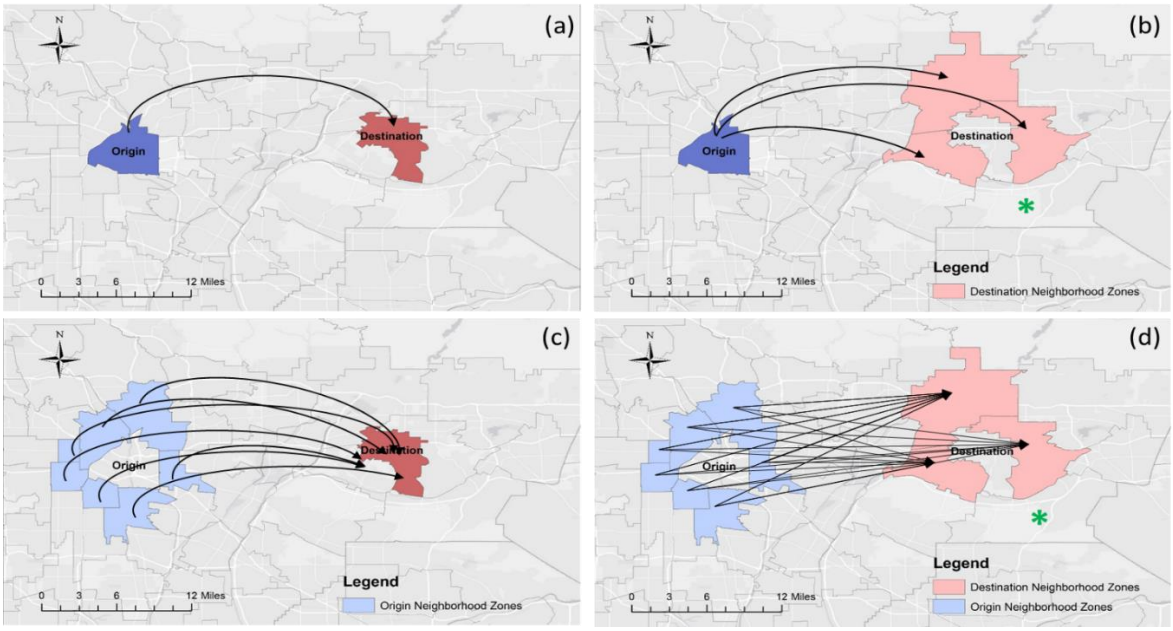


Figure 16. Defining Spatially Lagged Variables for the OD trip

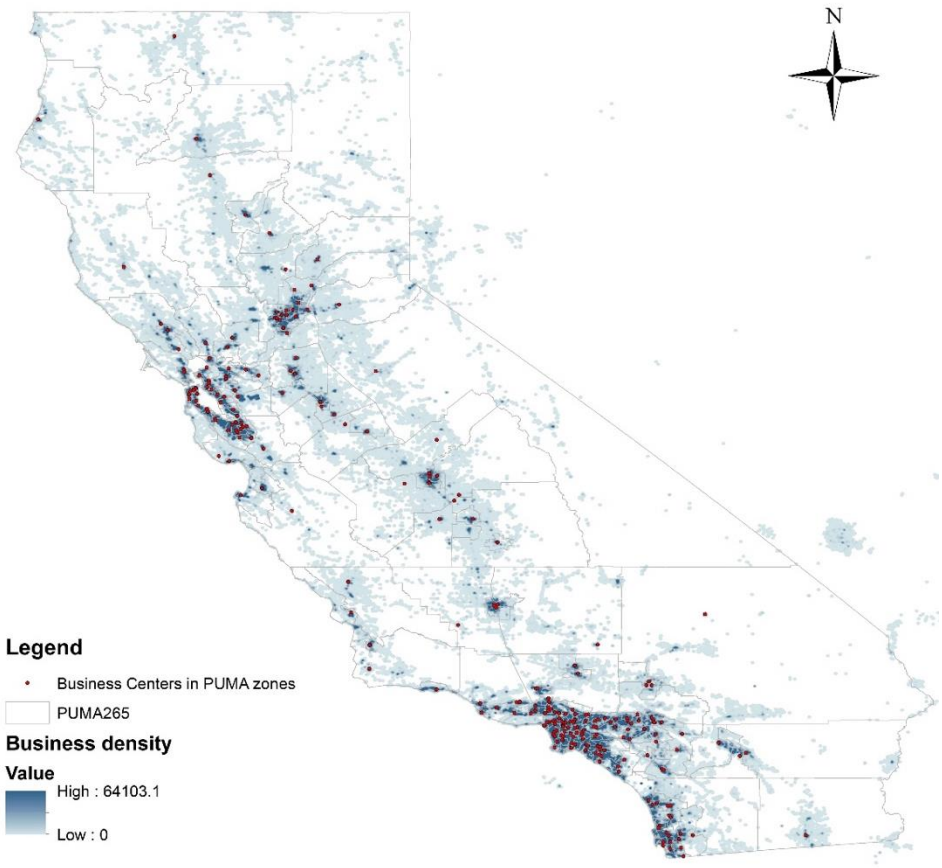


Figure 17. Business Employment Centers of PUMA zones in California

9.1.2. Spatial Lag Tobit Model

In the CSTDM OD database we find 17 percent OD pairs with number of trips smaller than one. A dependent variable like this is limited from below and can be considered as censored at zero (Monzon et al., 1989). There are exact and approximate ways to estimate regression models accounting for limits and censoring (Goulias and Kitamura, 1993). With this type of dependent variable, the Tobit model is generally recommended because the model provides functionality to handle censored distributions (Maddala, 1986, Greene 2003). It is worth noting that we use spatially lagged explanatory variables of the neighboring zones. According to Xu and Lee (2014), the Maximum Likelihood estimator (MLE) produces consistent estimates for Spatial Autoregressive Tobit Model. The marginal effects of explanatory variables x or Z are the partial derivative of the expected value of y with respect to variables included in the specification and is

a function of the coefficient and the probability of a unit to be a nonzero zone (Maddala, 1986). In this way we can obtain the unit contribution of a Twitter OD trip on CSTDM model output.

9.1.3. Latent Class Regression Model

Though the spatial lag regression model provides a suitable framework to find the conversion of Twitter OD trips to CSTDM OD trips, it may be limited in reflecting the heterogeneous nature of space. The Latent Class Analysis provides a method to capture many different types of heterogeneities because this model allow us to classify observational units into a set of latent classes and estimate class-specific regression models simultaneously. This is particularly useful when we attempt to capture spatial heterogeneity (Deutsch-Burgner and Goulias, 2014). With our dataset, spatially similar OD trips can be grouped into latent classes and regression coefficients are estimated for each class simultaneously with the determination of the number of classes (Vermunt and Magidson, 2013). In this way, we can test if each hypothetical class has different Twitter trip conversion multiplier.

As we did in the models described earlier, we use CSTDM OD trip as the dependent variable in the latent class model. This model features two distinct types of exogenous variables: 1) covariates – variables affect the latent variable defining classes; and 2) predictors – variables that affect the dependent variable (CSTDM OD trips). Model estimation follows the method described by Vermunt and Magidson (2002). The likelihood function of a multi-class latent regression model has many local maxima and we test multiple models with different sets of initial values of parameters (Goulias, 1999). Since the degrees of freedom rapidly decrease as we increase the number of parameters, this may lead to a variety of operational problems with model identification (inability to estimate a parameter) or failure to converge (subsequent estimation step parameters are not close enough). Therefore, we use a hierarchical iterative process to estimate this model as follows:

a) Start with one-class assumption without covariates;
b) Proceed by increasing number of classes for the models until any parameter fails to be identified and the size of a class becomes too small to be meaningful;

c) Estimate a series of Latent Class Regression with different combinations of exogenous variables and select the most suitable number of classes based on changes in goodness of fit criteria, such as Bayesian Information Criterion (BIC), Akaike Information Criterion (AIC) and the Consistent Akaike Information Criterion (CAIC), following McCutcheon, 2002, and Nylund et al., 2007;

d) Compare the models with different specifications and select the best model based on multiple statistical goodness-of-fit measures like the second step as well as classification errors and R-square values. The higher R-square indicates better model in predicting the endogenous variable, but the lower classification error means better model in classifying spatially homogenous groups.

9.2. Spatial Lag Tobit Model Results

The Tobit regression model with spatial lag variables was estimated with NLOGIT 5.0, and the partial effects are shown in Table 13. The partial effect of the Twitter OD trips is B = 33.021, significant at the level of 0.001. This indicates that every trip from twitter corresponds to a larger number of CSTDM OD trips, while, accounting for other influencing factors. All of the spatial lag variables are significantly different than zero at 0.01 level of significance. This means that both spatial lag variables of the exogenous and endogenous variables play significant roles and are able to control for spatial autocorrelation in this model. In terms of land use and demographic characteristics, positive coefficients are found for the size of origins and destinations area (significant at the level of 0.05) as expected. On the other hand, negative coefficients are found for population densities at both origins and destinations, presumably because CSTDM denser zones also have a larger number of Twitter trips. In other words, CSTDM trips are adjusted by land use and demographic variables if there is a large number of Twitter trips. Finally, the negative coefficient associated with the route distances between PUMA zones indicates that there are more trips between nearby zones than zones further apart.

Table 13. The List of Estimated Partial Effect

Independent Variables		Partial Effect					
		Coef.	S.E.	z	P[Z|>z]	Conf. Interval	
Twitter OD	T_OD	33.021	0.285	115.870	0.000	32.462	33.579
Spatial Lag Variables from CSTDM OD	C_Oad_Dad	-0.294	0.009	-31.010	0.000	-0.313	-0.276
	C_Oad_D	0.536	0.009	61.930	0.000	0.519	0.552
	C_O_Dad	0.572	0.009	63.180	0.000	0.554	0.590
Spatial Lag Variables from Twitter	T_Oad_Dad	21.991	1.526	14.410	0.000	19.000	24.981
	T_Oad_D	-40.263	0.884	-45.540	0.000	-41.997	-38.531
	T_O_Dad	-41.251	0.889	-46.380	0.000	-42.995	-39.509
Demographic Characteristics from US Census	O_AREA km^2	0.030	0.003	8.620	0.000	0.023	0.036
	D_AREA km^2	0.029	0.003	8.400	0.000	0.022	0.036
	O_Population	0.004	0.001	3.300	0.001	0.001	0.006
	D_Population	0.003	0.001	2.780	0.006	0.001	0.005
	O_Housing	0.005	0.002	1.940	0.052	0.000	0.010
	D_Housing	0.007	0.002	2.660	0.008	0.002	0.011
	O_POP_Density (person/km^2)	-0.255	0.035	-7.380	0.000	-0.323	-0.188
	D_POP_Density (person/km^2)	-0.311	0.035	-8.970	0.000	-0.379	-0.243
	O_Housing Density (House/km^2)	0.425	0.084	5.090	0.000	0.261	0.588
	D_Housing Density (House/km^2)	0.512	0.083	6.130	0.000	0.348	0.675
Land Use Characteristics from NETS dataset	O_Number of Employees	1.855	0.444	4.170	0.000	0.983	2.726
	D_Number of Employees	0.800	0.445	1.800	0.073	-0.073	1.673
	O_Business Diversity	306.117	107.016	2.860	0.004	96.369	515.864
	D_Business Diversity	91.372	106.899	0.850	0.393	-118.146	300.891
Distance	Route Distance Between OD (km)	-3.629	0.070	-52.200	0.000	-3.765	-3.492

9.3. Spatial Lag Latent Class Regression Model Results

We used a stepwise approach to develop the Latent Class Regression Model. The first step is identifying a suitable number of classes describing this OD trip dataset. Similar to the spatial lag Tobit model we use CSTDM OD trips as the dependent variable, and estimated a series of Latent Class models (also called mixture regression models) starting with one-class and increasing the classes until we find an optimal model. No explanatory variable was added in this step and eight models were identified (Table 14). Although model fit improves with each additional class, goodness of fit indices (BIC, AIC, AIC3) ceased to improve dramatically beyond the four-class

model, reaching an asymptote. In other words, this indicates that it is possible to explain the heterogeneous nature of the CSTDM trips efficiently with four latent classes representing three different groups of zones. Therefore, the subsequent latent class regression models are estimated using the four-class assumption.

Table 14. The List of Estimated Partial Effect

	LL	BIC(LL)	AIC(LL)	AIC3(LL)	CAIC(LL)	Npar	Class.Err.
1-Class	-628029	1256258	1256094	1256112	1256276	18	0
2-Class	90927.72	-181376	-181769	-181726	-181333	43	0.0049
3-Class	106110.6	-211462	-212085	-212017	-211394	68	0.0046
4-Class	111243.8	-221450	-222302	-222209	-221357	93	0.0096
5-Class	114619.8	-227923	-229004	-228886	-227805	118	0.0161
6-Class	114784.9	-227974	-229284	-229141	-227831	143	0.0145
7-Class	117099.5	-232324	-233863	-233695	-232156	168	0.0247
8-Class	117687.8	-233222	-234990	-234797	-233029	193	0.0256

The covariates and predictors play different roles in estimating Latent Class Regression Model as discussed earlier; covariates influence the latent classes and predictors influence the dependent variable. Since we use latent class analysis to capture the spatial heterogeneity, covariates in this model reflect spatial characteristics of Origins and Destinations. All of our exogenous variables contain zonal information, therefore all of them could be used as either covariates, predictors, or both. Although there is no consensus in the literature about which exogenous variables should be used for this type of analysis, our previous experiment in Southern California Association of Governments area found that the latent regression model using spatial lag variables as covariates and all others as predictors produced the best results in OD matrix conversion.

As mentioned earlier, the model was estimated with four classes, and their estimated membership proportions are reported in Table 15. The largest proportion of the sample (OD pairs) was found in the first class (88 %, 61,995 OD pairs), followed by the second, third and fourth class (6% 4,388 OD pairs, 4% 2,851 OD pairs, and 2% 991 OD pairs, respectively). However, in terms of CSTDM OD trips, by far the largest proportion of OD trips (67.3%, 61,038,429 OD trips) were found in the fourth class followed by the third, second, and first class (26.6%, 24,067,710 OD trips 4.6%, 4,196,934 OD trips and 1.5%, 1,339,309 OD trips, respectively). Although the fourth class consists of the smallest number of OD pairs, it has the largest number of CDTDM OD trips.

Because we use spatial lag variables as covariates, these latent classes represent the homogeneous groups of OD flows with respect to their neighbors' OD flow patterns. The right hand side of Table 15 provides the descriptive statistics of both CSTDM and Twitter OD trips and covariates for each class. The first class captures OD pairs with relatively few trips; these pairs have relatively small numbers of both CSTDM and Twitter OD trips and are adjacent to similarly low-traffic OD pairs. The second and third class captured zone pairs with a moderate and mid-high level of CSTDM trips, and the fourth class consists of the OD pairs with the largest number of trips by both measures as well as large interactions between their neighbors.

Table 15. Proportion of Latent Classes and Descriptive Statistics of Each Class

Class modal		CTrips	C_Oa_Da	C_Oa_D	C_O_Da	T_OD	T_Oa_Da	T_Oa_D	T_O_Da
1 (N=61,995, 1,339,309 CSTDM trips)	Mean	21.6	31.5	25.3	22.1	0.9	0.6	0.7	0.6
	Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Median	3.0	0.8	1.5	0.6	0.0	0.0	0.0	0.0
	Max	813.9	8481.1	3007.5	4054.7	265.0	67.0	128.0	134.0
	Std.Dev.	53.5	155.9	89.5	85.8	4.4	2.3	2.9	2.8
2 (N=4,388, 4,196,934 CSTDM trips)	Mean	956.5	1765.2	1281.5	1233.6	11.2	14.9	12.6	12.6
	Min	6.8	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Median	767.5	1038.9	867.8	814.8	4.0	6.0	5.0	4.0
	Max	9489.1	27872.0	16419.2	11263.3	588.0	274.0	524.0	525.0
	Std.Dev.	659.4	2159.3	1351.7	1369.8	29.1	27.5	30.6	32.5
3 (N=2,851, 24,067,710 CSTDM trips)	Mean	8441.8	7815.0	8230.8	8022.3	27.7	34.8	30.6	29.6
	Min	149.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Median	5071.7	6116.4	6228.4	5955.6	14.0	17.0	17.0	16.0
	Max	416149.1	65133.9	70473.9	70473.9	2161.0	424.0	565.0	561.0
	Std.Dev.	13521.0	7341.8	7465.9	7976.3	68.2	48.7	46.7	46.2
4 (N=991, 61,038,429 CSTDM trips)	Mean	61592.8	16504.3	20061.4	19035.9	301.7	79.6	119.6	116.7
	Min	905.7	0.0	0.0	0.0	5.0	0.0	0.0	0.0
	Median	35309.9	13430.1	16995.5	16255.8	117.0	47.0	52.0	51.0
	Max	492169.6	85610.1	87855.6	88592.9	8194.0	589.0	2134.0	1980.0
	Std.Dev.	71505.5	14768.0	16505.1	16178.6	613.9	97.7	193.8	194.0
Total (N=70,225, 90,642,383 CSTDM trips)	Mean	1290.7	688.3	719.7	690.9	6.9	4.0	4.3	4.1
	Min	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
	Median	4.3	1.0	2.0	1.0	0.0	0.0	0.0	0.0
	Max	492169.6	85610.1	87855.6	88592.9	8194.0	589.0	2134.0	1980.0
	Std.Dev.	11591.1	3408.9	3773.5	3706.8	82.8	20.5	30.3	30.2

Table 16 shows latent class-specific coefficients of predictors as well as Wald statistics providing the results of the statistical test if the coefficients are different between the latent classes. The coefficients with grey color shading indicate significant value at the 5% level. Significant

predictors of CSTDM OD-trips are the classes themselves (i.e., the overall class specific averages are different). Based on the Wald test statistics, all of the predictors are different among latent groups except for the housing related variables and population size in origins. Most importantly, the coefficients of Twitter trips turned out to be very different. The smallest unit contribution of Twitter OD trips was found in the first class, the biggest one was found in the third class (2.1279, 183.3002). Although the CSTDM OD pairs in the fourth class has the largest number of trips per OD pair, the coefficient on Twitter OD trips was relatively small because large number of Twitter OD trips were also found in the fourth class. This indicates that a Twitter based OD trip should be used in a different way depending on the underlying spatial structures when we validate model-based OD trips. This result also shows the necessity of using a methodology that is able to reflect the heterogeneous nature of geography and the people living in different geographies. Although the first latent class regression model had the smallest R-square value (0.2960) among classes, it included a variety of significant predictors (14 in total), and their signs are the same with the output of the spatial lag Tobit model in the previous section. However, the magnitude of coefficients is smaller than the spatial lag Tobit model results (e.g., the unit contribution of a Twitter OD trip for this latent group was 2.1279 and the difference with the Tobit is 30.893). The density of housing and population in both origins and destinations have different directional effects in Tobit model and the latent class model especially in the first class. Higher housing density and lower population in origins and destinations indicate higher number of CSTDM trips between two zones in the Tobit model, but their effects in the first latent class were the opposite.

The smallest number of significant predictors was found in the second model with the moderate R-square value (0.4713). Among 16 predictors, only two significant predictors were found in the second model, but the Twitter trips play the most important roles in this class. Also, the negative coefficient was found at the number of employees in destinations. This means that a higher number of employees in this class imply more number of trips.

The highest R-square value (0.9317) was found in the third model with ten significant predictors; Twitter OD trips, area and population sizes in origins and destinations have the positive coefficients, but the number of houses, business employees in both origins and destination, and

route distances between zones are the negatively associated with the trips in CSTDM output. However, all of the density and diversity variables are not significantly related to the CSTDM OD matrix. The highest coefficient of Twitter OD trips across all the classes was found in this class (183.3002). Presumably, this is associated with higher number of CSTDM trips and lower number of Twitter trips (Table 15).

The fourth class regression model was estimated with R-square 0.5450, with ten significant predictor variables. This model produced the closest unit contribution of a Twitter OD trip to the Tobit model (32.6650). All of the significant coefficients in the fourth class were much bigger impact on CSTDM trips than all of the other classes, for example, route distance between Origins and destinations were (class1: -0.1380, class2: -0.0235, class3: -133.5920, class4: -1,782.6917). This is presumably due to the shortest mean distance between origins and destinations in class 4 (class1: 424.1, class2: 44.5, class3: 27.2, class4: 13.9).

Finally, the spatial lag variables played important roles as covariates in this model, the coefficients can be found in Table 16. Based on Wald statistics, the amount of trips from neighborhood area to the destinations from CSTDM model was the most important variable to classify the latent classes followed by two other spatial lag variables from CSTDM data based on Wald statistics.

In terms of spatial distributions of the OD pairs of latent classes, those are distributed differently across California (Figure 18). In this map, the straight lines between OD pairs are used to illustrate the distributions of the OD pairs. The first class seems to represent all of the long distance OD pairs as we found it at the route distances' coefficients in latent class regression model. The straight lines in this class cover the entire state of California. However, from the second class to the fourth class, the spatial distributions of the OD pairs tend to be much shorter than the first class. Although the second and third classes cover some inter-regional OD pairs between zones, the fourth latent class seems to cover inner zone trips as well as the shortest OD pairs.

Figure 19 shows a set of maps describing California with bar charts indicating the amount of trips with the origins in red and destinations in blue. Each map shows each latent class's CSTDM OD

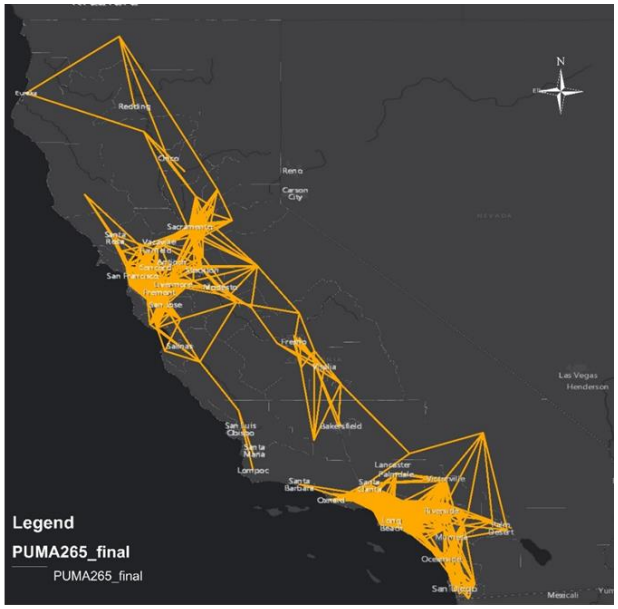
trips. The first class OD pairs are widely distributed across California. The second and third classes are more densely concentrated in the City of Los Angeles and San Francisco Area and the fourth class is quite evenly distributed like the first class. These maps also show spatial clusters of the zones that have similar OD trips patterns with their neighbors' trip patterns. Also this classification reflects the effect of size and relative location of the zones because those were captured via the spatial lag variables and used as covariates.

Table 16. Class-specific Coefficients of Predictors

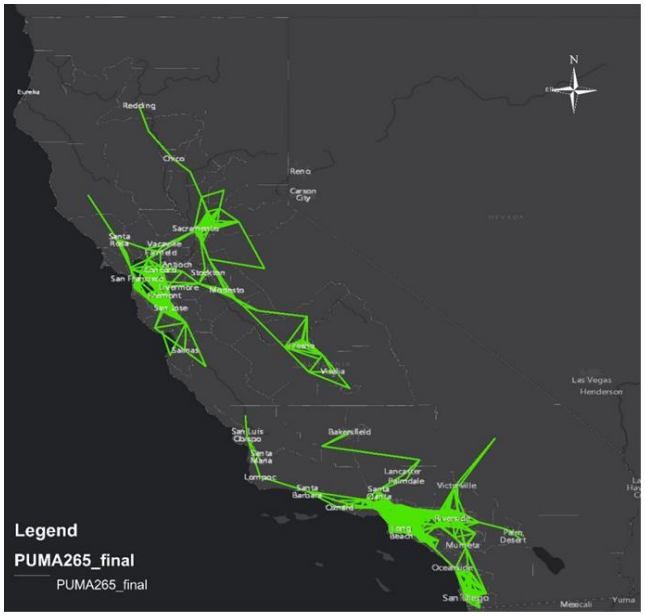
Independent Variables		Class1		Class2		Class3		Class 4		Mean	Std.Dev.
		Coef. 1	z-value	Coef. 2	z-value	Coef. 3	z-value	Coef. 4	Z-value		
Twitter OD	T_OD	**2.1279**	34.9053	**16.0510**	38.4447	**183.3002**	129.2272	**32.6550**	11.2504	11.086	36.492
Demographic Characteristics from US Census	O_AREA km^2	**0.0004**	9.1234	0.0007	0.2895	**0.2338**	6.5674	**3.1397**	2.8830	0.058	0.386
	D_AREA km^2	**0.0004**	9.2788	-0.0007	-0.2948	**0.2170**	6.1182	**4.6225**	4.2129	0.080	0.567
	O_Population	0.0000	0.8308	0.0000	-0.0745	**0.0209**	3.4520	**0.4783**	3.8747	0.008	0.059
	D_Population	0.0000	1.2282	-0.0002	-0.2673	**0.0178**	3.1310	**0.4484**	3.6379	0.008	0.055
	O_Housing	**0.0001**	4.1581	0.0023	1.5021	**-0.0635**	-4.1519	-0.4975	-1.7992	-0.010	0.062
	D_Housing	**0.0001**	4.4004	0.0023	1.5228	**-0.0539**	-3.6925	-0.4967	-1.8179	-0.010	0.062
	O_POP_Density (person/km^2)	**0.0034**	6.9663	0.0243	1.3675	-0.2222	-1.2095	**-16.3914**	-4.1901	-0.255	2.008
	D_POP_Density (person/km^2)	**0.0020**	4.0745	0.0238	1.3434	-0.1473	-0.8058	**-15.8827**	-4.0761	-0.245	1.946
	O_Housing Density (House/km2)	**-0.0062**	-5.3681	-0.0600	-1.4023	-0.6906	-1.4664	**22.3358**	2.6187	0.302	2.745
	D_Housing Density (House/km^2)	**-0.0037**	-3.1935	-0.0529	-1.2387	-0.8863	-1.8646	**21.3824**	2.5227	0.282	2.631
Land Use Characteristics from NETS dataset	O_Number of Employees (1,000)	**0.1032**	16.0630	-0.3568	-1.8687	**-7.2615**	-3.5004	-41.8349	-1.5795	-0.875	5.306
	D_Number of Employees (1,000)	**0.0722**	11.2905	**-0.5693**	-3.0191	**-7.9954**	-3.6742	-35.3275	-1.3562	-0.847	4.586
	O_Business Diversity	**20.9423**	14.0173	48.8464	0.9275	-117.6351	-0.2690	4870.4160	0.4791	90.693	595.338
	D_Business Diversity	**17.3669**	11.6541	25.7911	0.4881	-433.7490	-1.0066	2543.8045	0.2483	37.335	324.780
Distance	Route Distance Between OD	**-0.1380**	-113.8530	-0.0235	-0.6398	**-133.5920**	-12.7773	**-1782.6917**	-17.4804	-32.934	219.330

Covariates	Class1	z-value	Class2	z-value	Class3	z-value	Class4	z-value	Wald	p-value
C_Oa_Da	0.000	-12.596	0.000	5.709	0.000	8.952	0.000	15.321	236.844	0.000
C_Oa_D	-0.004	-36.172	0.001	26.749	0.002	37.944	0.002	38.663	1539.952	0.000
C_O_Da	-0.003	-26.300	0.001	18.993	0.001	28.227	0.001	28.584	896.758	0.000
T_Oa_Da	0.054	12.174	-0.009	-5.233	-0.017	-9.592	-0.027	-13.882	194.247	0.000
T_Oa_D	-0.006	-1.524	0.003	2.341	-0.003	-1.981	0.006	3.653	62.943	0.000
T_O_Da	-0.012	-3.251	0.005	3.763	-0.001	-0.758	0.008	5.579	77.923	0.000

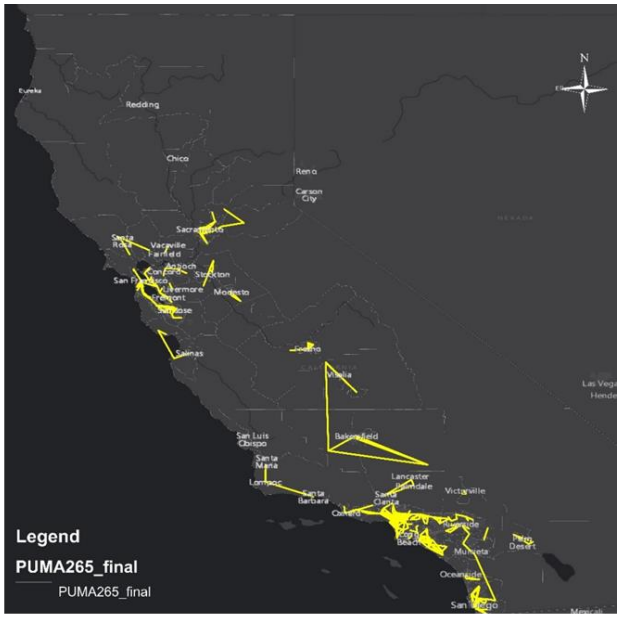
Class 1



Class 2



Class 3



Class 4

Figure 18 Spatial distribution of the OD pairs of each latent class

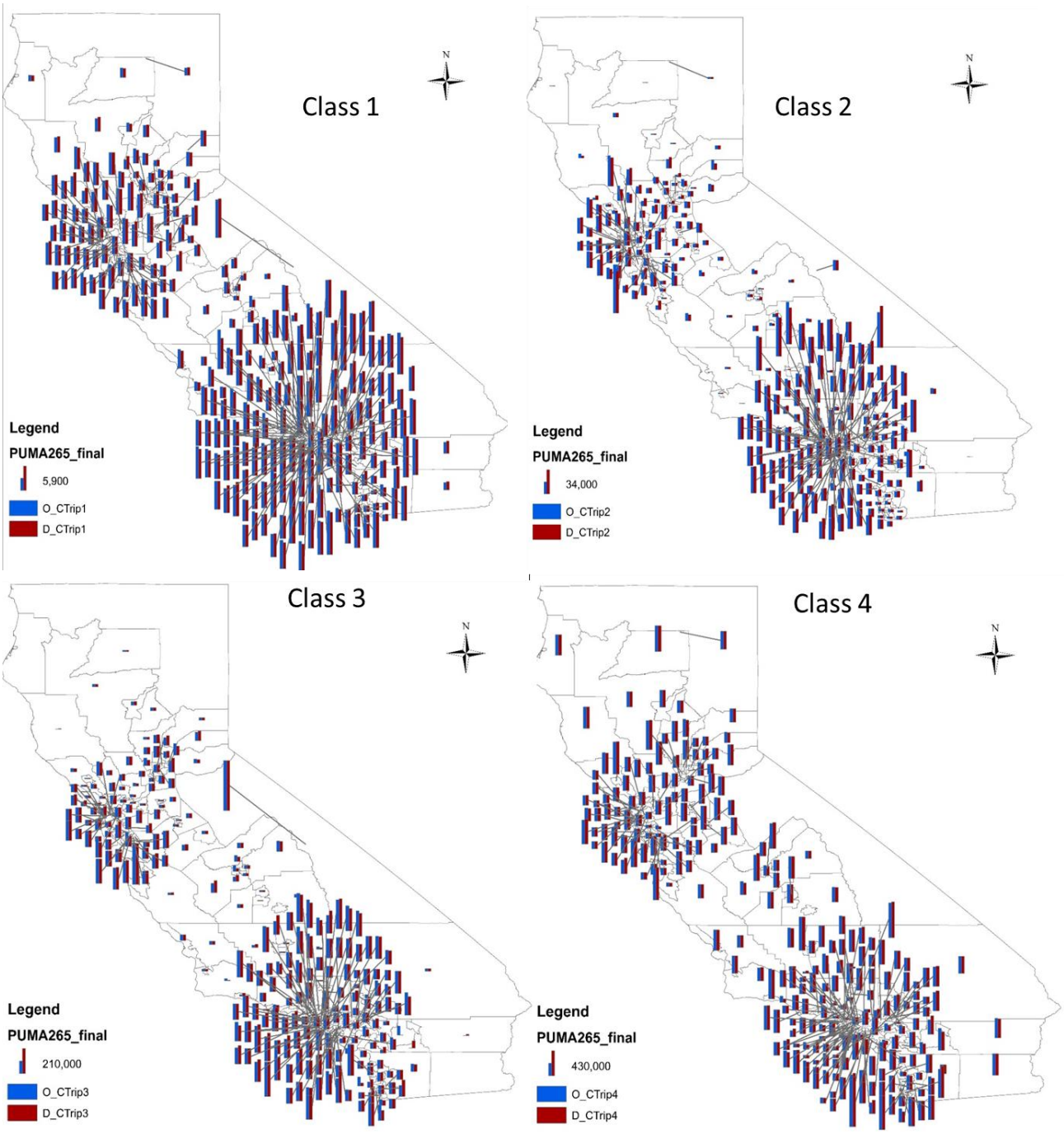


Figure 19. Spatial distributions of the CSTDM OD trips in each class

Figure 20 describes the proportion of the OD pairs that are classified by the latent class regression model within each metropolitan planning organization in California. The total number of OD pairs are also provided underneath the proportional bar chart. The larger MPOs seem to consist of diverse latent classes, for example SCAG, MTC, SANDAG, and SACOG. On the other hand, smaller MPOs are mainly accounted for by third and fourth classes. This is presumably because the larger MPOs consist of diverse OD pairs from short distance to long distance OD pairs, and urban and rural area. This result reinforces the fact that spatially heterogeneous OD pairs require different conversion coefficients from Twitter trip to CSTDM trips. Moreover, the OD pairs in different MPOs may need their unique conversion coefficients because their combination of latent classes are different from each other. In this regard, we estimated Tobit models for four large MPO area separately, and found different conversion coefficients (Table 18). As a result, SCAG model has the lowest conversion coefficient (24.3), but highest one was found at SACOG model (191.4). This result verifies the necessity of using conversion models which account for spatial heterogeneity. In other words, we need different conversion for different regions.

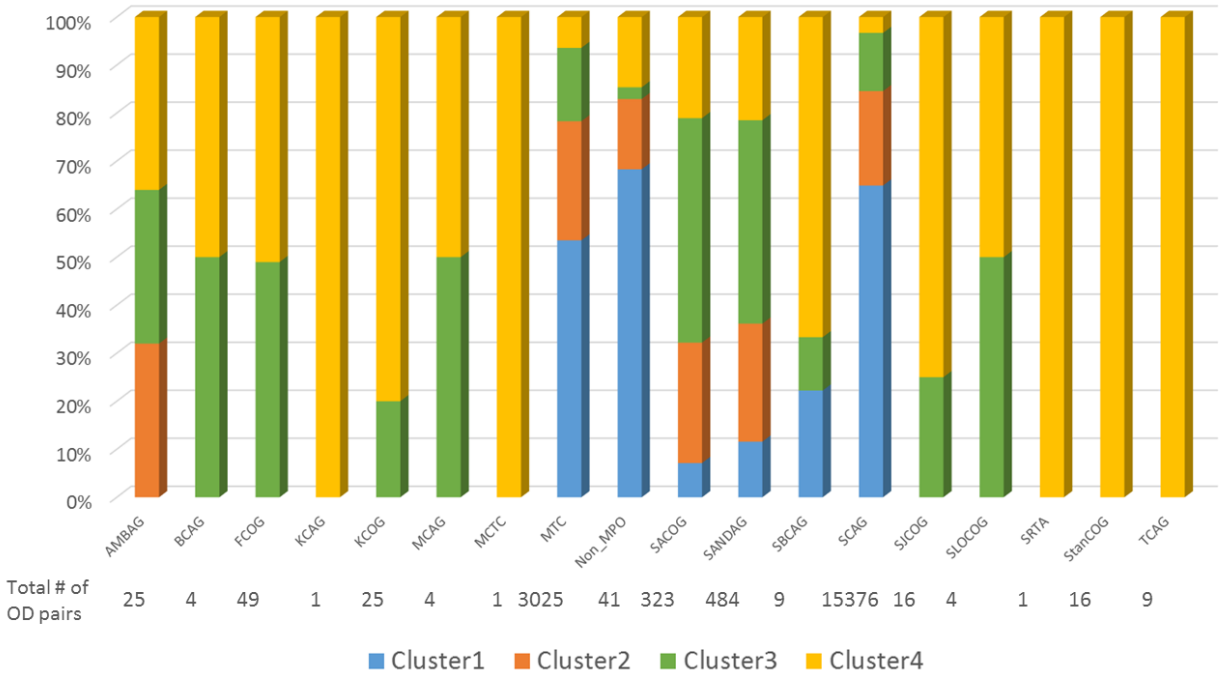


Figure 20. The proportion of OD pairs in each latent class within each MPOs

Table 18. Four MPOs and their conversion coefficients

	MTC	SACOG	SANDAG	SCAG	Total
Mean of Twitter OD trips	27.7	33.2	76.2	18.8	6.9
Mean of CSTDM OD trips	5805.9	16444.3	15924.5	2821.3	1289.4
Conversion Coefficients	55.1	191.4	40.6	24.3	33.1
N	3025	323	484	15376	70225

10. Summary and Findings

10.1. Summary

In this project, a Twitter data harvester was developed with Python code, and data were stored as JSON format with MongoDB server 3.0x; approximately 8 million geotagged tweets were used as an input for this project. Three different Twitter trip extraction algorithms were developed, and Rule #2 produced the most suitable list of Twitter trips from geotagged Twitter data. The list of Twitter trips was transformed into a trip generation table and an OD matrix with spatial aggregation; the former aggregated in block group level (23,092 zones), and the latter in PUMA level (70,225 OD pairs). We compared the Twitter based trip production table with Synthetic Population estimated by "Task 2644: Spatial Transferability Using Synthetic Population Generation Methods." We found positive correlation between Twitter trip and Walking, Bicycling, and Drive alone trips from Synthetic population in both total number of trips and sum of the trip lengths in block groups. Twitter data was not able to capture the differences between weekday and weekend, and weekday and Thanksgiving day. In terms of trip lengths, Twitter trip data has similar distribution to the California household travel survey data, but their trip durations were slightly different, Twitter data produce a smoother distribution than the other because it was computed using Google API. We compared the Twitter based OD matrix with a recent OD matrix (CSTDM output) given by the California Department of Transportation. We used a Spatial Lag Tobit model to develop an unbiased conversion method between Twitter trips and Travel Demand Model output. We also used Latent Class Regression models to take into account of the heterogeneous nature of space. The spatial lag Tobit model produced a single unit-contribution of Twitter trip, but four different unit-contributions depending on spatial structures were obtained with Latent Class Regression model. Tobit models for four MPO area produced different unit-contributions of Twitter trip, and SCAG area has the lowest conversion coefficient and SACOG area has the highest one.

10.2. Recommended Methods

There are three types of methods required in this project including 1) social media data harvester, 2) Twitter trip extraction, 3) OD matrix conversion. For the first method, we recommend to use any programming language that is connected to MongoDB. Although we use Python for this project, it is possible to develop exactly the same program with Java or other programming languages. However, it is very important to use database software, such as MongoDB, so as to store, access, query, and extract large amounts of social media data efficiently. In terms of Twitter trip extraction, we recommend the Rule #2 for smaller input data, but Rule #3 would be the best extraction method, theoretically and practically. Lastly, the OD matrix conversion with spatial lag latent class regression model provides the functionality to account for errors from spatial autocorrelation as well as to capture spatial heterogeneity of OD pairs. However, the spatial lag Tobit model would be helpful to estimate individual MPO's conversion coefficient.

10.3. Next Steps in Research

A variety of research directions have emerged from lessons learned in this project. First, with the comparison of Twitter trips and synthetic population, we found walking trips are strongly related to Twitter trips, so our immediate next step is to perform in-depth analysis of Twitter trip and its relationship with walking trips. It is possible that this relationship is due to different land uses and resident characteristics not captured in the analysis of this project and can explain the relationship.

In addition, although 6-month observation was a long period of data collection, it would be better to collect data for more than a year like the California Household Travel Survey. In this way, we can observe the year-long dynamics of travel behavior. Moreover, we envision the creation of an observatory project in which social media data are collected for more than a year (to mimic CHTS). This could provide valuable information for not only Caltrans but also the MPOs.

Twitter is used heavily by a segment of the population for which we have limited travel behavior data. This segment includes students residing in group quarters, and social media may be the only currently available source to understand their behavior. Developing a small scale survey that is also informed by social media will provide invaluable information for modeling and simulation of travel behavior for this group. In addition, as a first step we could create a hot spot

analysis and identify if many of the trips we estimated in this project have their origins at colleges and universities and then design surveys that target the locations with the highest number of tweets.

This data and the findings from this project will play very important roles in our new Caltrans project (Task Order- 65A0529 TO 047: Long Distance Travel in the California Household Travel Survey (CHTS) and Social Media Augmentation).

Although we focused on using geographic information of tweets to extract Twitter trips regardless of users' information or text, using all other information provided by Twitter streaming API would be very helpful. For example, we may be able to identify trip information from text mining of tweets, and potential home locations of heavy Twitter service users with night time tweets' locations. Correlating business establishment information with tweet text may also give us information about activity participation and human interaction during the activities and travel. Related to this is the possibility of identifying home locations and other frequently visited places by tracking individual tweet IDs.

Moreover, we can impute missing trips when two tweets' time difference is much longer than estimated travel time based on each user's home locations and major tweeting locations. In this way, we can identify social media data with high potential of complementary information to the traditional survey data.

References

Auld, J., Williams, C., Mohammadian, A., & Nelson, P. (2009). An automated GPS-based prompted recall survey with learning algorithms. Transportation Letters, 1(1), 59–79. http://doi.org/10.3328/TL.2009.01.01.59-79

Auld, J., Mohammadian, A., and Wies, K. Population Synthesis with Regional-Level Control Variable Aggregation. Paper presented at the 87th Annual Transportation Research Meeting, Washington D.C., January 2008.

Beckman, R.J., K.A. Baggerly, and M.D. McKay (1996) Creating Synthetic Baseline Populations. Transportation Research Part A: Policy and Practice, 30(6), pp. 415-429.

Cebelak, M. K. (2013). Location-based social networking data : doubly-constrained gravity model origin-destination estimation of the urban travel demand for Austin, TX. Retrieved from http://repositories.lib.utexas.edu/handle/2152/22296

CALTRANS (2013) California Household Travel Survey

http://www.dot.ca.gov/hq/tpp/offices/omsp/statewide_travel_analysis/chts.html (Accessed March 2016)

CALTRANS (2016) California Transportation Plan 2040 http://www.dot.ca.gov/hq/tpp/californiatransportationplan2040/ (Accessed March 2016)

Cambrige Systematics (2014), California Statewide Travel Demand Model, Version 2.0 Model

California Energy Commission (2013) California Light Duty Vehicle Survey http://www.energy.ca.gov/2013_energypolicy/documents/2013-06-26_workshop/presentations/03_Light_Duty_Vehicle_Survey-Aniss_RAS_21Jun2013.pdf

Chen, X., & Yang, X. (2014). Does food environment influence food choices? A geographical analysis through “tweets.” Applied Geography, 51, 82–89.

Chen, Y., Frei, A., & Mahmassani, H. S. (2015). Exploring Activity and Destination Choice Behavior in Social Networking Data. Retrieved from http://trid.trb.org/view.aspx?id=1339428

Coffey, C., & Pozdnoukhov, A. (2013). Temporal Decomposition and Semantic Enrichment of Mobility Flows. In Proceedings of the 6th ACM SIGSPATIAL International Workshop on Location-Based Social Networks (pp. 34–43). New York, NY, USA: ACM. http://doi.org/10.1145/2536689.2536806

Cottrill, C., Pereira, F., Zhao, F., Dias, I., Lim, H., Ben-Akiva, M., & Zegras, P. (2013). Future mobility survey: Experience in developing a smartphone-based travel survey in singapore. Transportation Research Record: Journal of the Transportation Research Board, (2354), 59–67.

Davis A. W., J. H. Lee, E. McBride, S. Ravulaparthy and K. G. Goulias (2016) Business establishments survival and transportation system level of services, Final Report

Duggan, M., Ellison, N. ., Lampe, C., Lenhart, A., Madden, M., Rainie, L., & Smith, A. (2015). Social Media Update 2014: While Facebook remains the most popular site, other platforms see higher rates of growth. Pew Research Center.

Fan, Y., Chen, Q., Liao, C.-F., & Douma, F. (2012). UbiActive: A Smartphone-Based Tool for Trip Detection and Travel-Related Physical Activity Assessment. In Submitted for Presentation at the Transportation Research Board 92nd Annual Meeting. Retrieved from http://docs.trb.org/prp/13-4250.pdf

Farooq, B., Bierlaire, M., Hurtubia, R., & Flötteröd, G. (2013). Simulation based population synthesis. Transportation Research Part B: Methodological, 58, 243-263.

Gao, S., Yang, J.-A., Yan, B., Hu, Y., Janowicz, K., & McKenzie, G. (2014). Detecting Origin-Destination Mobility Flows From Geotagged Tweets in Greater Los Angeles Area. Retrieved from http://www.geog.ucsb.edu/~sgao/papers/2014_GIScience_EA_DetectingODTripsUsingGeoTweets.pdf

Goulias, K. G., Bhat, C. R., Pendyala, R. M., Chen, Y., Paleti, R., Konduri, K. C., … Hu, H.-H. (2012). Simulator of Activities, Greenhouse Emissions, Networks, and Travel (SimAGENT) in Southern California. In Transportation Research Board 91st Annual Meeting. Retrieved from http://trid.trb.org/view.aspx?id=1128924

Goulias K.G. and E. L. Morrison (2010) Pre-Survey Design Consultant for the Year 2010 Post-Census Regional Travel Survey. Final Summary Report Project Number 10-046-C1(April 2010 to July 2010). Submitted to Southern California Association of Governments and Caltrans. June, Solvang, CA.

Goulias, K. G., Pendyala, R. M., & Bhat, C. R. (2013). Keynote—Total Design Data Needs for the New Generation Large-Scale Activity Microsimulation Models. Transport Survey Methods: Best Practice for Decision Making, 21.

Goulias, K. G., Ravulaparthy, S. K., Konduri, K. C., & Pendyala, R. M. (2014). Using Synthetic Population Generation to Replace Sample and Expansion Weights in Household Surveys for Small Area Estimation of Population Parameters. In Transportation Research Board 93rd Annual Meeting (No. 14-0501).

Guo, J. Y., and C.R. Bhat (2007) Population Synthesis for Microsimulating Travel Behavior. Transportation Research Record: Journal of the Transportation Research Board, (2014), pp. 92-101

Hasan, S., & Ukkusuri, S. V. (2014). Urban activity pattern classification using topic models from online geo-location data. Transportation Research Part C: Emerging Technologies, 44, 363–381. http://doi.org/10.1016/j.trc.2014.04.003

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. Paper 16-6639 Presented at the 2016 Annual Transportation Research Board Meeting, Washington D.C.

Lampoltshammer, T. J., Kounadi, O., Sitko, I., & Hawelka, B. (2014). Sensing the public's reaction to crime news using the "Links Correspondence Method." Applied Geography, 52, 57–66.

Lee, J. H., Davis, A., Yoon, S. Y., & Goulias, K. G. (2015). Activity Space Estimation with Longitudinal Observations of Social Media Data Paper accepted for presentation at the 95th Annual Meeting of the Transportation Research Board, Washington, D.C., January 10-14, 2016. Also published as GEOTRANS Report 2015-7-01, Santa Barbara, CA.

Leiman, J. M., Bengelsdorf, T., & Faussett, K. (2006). Household Travel Surveys: Using Design Effects to Compare Alternative Sample Designs. Presented at the Transportation Research Board 85th Annual Meeting. Retrieved from http://trid.trb.org/view.aspx?id=776628

Lin, J., & Cromley, R. G. (2015). Evaluating geo-located Twitter data as a control layer for areal interpolation of population. Applied Geography, 58, 41–47.

McBride E., A. W. Davis, J. H. Lee, and K. G. Goulias (2016) Spatial Transferability Using Synthetic Population Generation Methods, Final Report

MTC (2016) Sample Weighting and Expansion Part II: Average Weekday Weights. http://analytics.mtc.ca.gov/foswiki/pub/Main/HouseholdSurvey2012Weights/CHTS1213_BayArea_Weighting_Part_II_Weekday.pdf. Accessed March 2016.

NREL (2016) Transportation Secure Data Center. http://www.nrel.gov/transportation/secure_transportation_data.html. Accessed March 2016

Nitsche, P., Widhalm, P., Breuss, S., & Maurer, P. (2012). A Strategy on How to Utilize Smartphones for Automatically Reconstructing Trips in Travel Surveys. Procedia-Social and Behavioral Sciences, 48. Retrieved from http://trid.trb.org/view.aspx?id=1255210

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

Pritchard, D. R., & Miller, E. J. (2012). Advances in population synthesis: fitting many attributes per agent and fitting to household and person margins simultaneously. Transportation, 39(3), 685-704.

Pendyala, R., Bhat, C., Goulias, K., Paleti, R., Konduri, K., Sidharthan, R., ... & Christian, K. (2012a). Application of Socioeconomic Model System for Activity-Based Modeling: Experience from Southern California. Transportation Research Record: Journal of the Transportation Research Board, (2303), 71-80.

Pendyala, R.M. , C. R. Bhat, K. G. Goulias, R. Paleti, K. Konduri, R. Sidharthan , and K. P. Christian. (2012b) SimAGENT Population Synthesis. Phase 2 Final Report 3 Submitted to SCAG, March 31, 2012, Santa Barbara, CA.

Ravulaparthy S. and K.G. Goulias (2011) Forecasting with Dynamic Microsimulation: Design, Implementation, and Demonstration. Final Report on Review, Model Guidelines, and a Pilot Test for a Santa Barbara County Application. University of California Transportation Center (UCTC) Research Project. Geotrans Research Report 0511-01, May, Santa Barbara, CA

RSG, J. Dill, J. Broach, K. Deutsch-Burgner, Y. Xu, R. Guenssler, D.M. Levinson, and W. Tang. (2015) Multiday GPS Travel Behavior Data for Travel Analysis. FHWA-HEP-015-026.

Shirima, K., Mukasa, O., Schellenberg, J. A., Manzi, F., John, D., Mushi, A., … Schellenberg, D. (2007). The use of personal digital assistants for data entry at the point of collection in a large household survey in southern Tanzania. Emerging Themes in Epidemiology, 4(1), 5.

Southern California Association of Governments(SCAG). (2012). 2012-2035 Regional Transportation Plan/Sustainable Communities StrategyRTP/SCS Adopted April 2012,.

Turner, S. (1996). Advanced techniques for travel time data collection. Transportation Research Record: Journal of the Transportation Research Board, (1551), 51–58.

Widener, M. J., & Li, W. (2014). Using geolocated Twitter data to monitor the prevalence of healthy and unhealthy food references across the US. Applied Geography, 54, 189–197.

Yang, W., & Mu, L. (2015). GIS analysis of depression among Twitter users. Applied Geography, 60, 217–223.

Yang, W., Mu, L., & Shen, Y. (2015). Effect of climate and seasonality on depressed mood among twitter users. Applied Geography, 63, 184–191.

Zhu, Y., & Ferreira, J. (2014). Synthetic population generation at disaggregated spatial scales for land use and transportation microsimulation. Transportation Research Record: Journal of the Transportation Research Board, (2429), 168-177.