U.S. DEPARTMENT OF ENERGY

SMARTMOBILITY

Systems and Modeling for Accelerated Research in Transportation

Victor Walker, Idaho National Laboratory
Dave Black, Idaho National Laboratory
Saika Belal, Lawrence Berkeley National Laboratory
C. Anna Spurlock, Lawrence Berkeley National Laboratory

December 2020

INL
Idaho National Laboratory

DISCLAIMER

This information was prepared as an account of work sponsored by an agency of the U.S. Government. Neither the U.S. Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness, of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. References herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the U.S. Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the U.S. Government or any agency thereof.

CONTENTS

FIGURES

TABLES

1. Introduction

In order to engage in transportation system planning and policy design, it is essential to understand travel behavior. Travel surveys are a common mechanism used to gather data on travel behavior. However, these surveys often cannot provide a complete picture of what drives a set of choices and how these choices drive mobility decisions. Additionally, many travel behavior details are often missing from self-reported survey-based trip diaries. Specific trips or activities may be reported through these mechanisms, but details about the timing of those trips, the routes used, etc., are generally not included. The WholeTraveler Transportation Behavior Study, conducted in the San Francisco Bay Area in the spring of 2018, was designed to collect data on a broad range of behavioral choices, and a multitude of behavioral factors that might drive or limit certain mobility decisions or preferences. In addition to a wide range of stated preferences and self-reported behavior, the study also included revealed behavior of its participants via a week of Global Positioning System (GPS) data collection.

Understanding mobility actions from these data is very valuable, especially when combined with the stated behavioral choices, detailed characteristic information, and stated motivational preferences from the online survey that preceded the GPS data collection phase of the WholeTraveler study. Comparing stated and revealed behaviors can build a greater understanding of behavioral mechanisms and tracking the nuances of motion provides more detailed insights than surveys alone.

There are benefits and limitations from using a variety of alternative methods for tracking high-resolution revealed travel behavior. This report documents the choices made to tackle this challenge for the WholeTraveler Transportation Behavior Study, including the method used to collect the data and the post-collection processing steps taken to convert the data from raw inputs into meaning. The data described here, along with the data from the online survey from the WholeTraveler study, anonymized to ensure protection of the survey participants, are available publicly on the Department of Energy (DOE)-funded Livewire data repository.

Section 2 of this report documents the methodology used to collection the revealed behavior location data for the WholeTraveler Transportation Behavior Study. Section 3 provides a summary of the location data collected. Section 4 describes the methodology developed to identify and define Trips and Trip Chains from the location points. Finally, section 5 summarizes the publicly available information and data and how to access it.

2. Data Collection Methodology

The WholeTraveler Transportation Behavior Study developed and conducted a transportation-based survey with the support of the U.S. Department of Energy's (DOE's) Energy Efficient Mobility Systems (EEMS) program as part of the SMART Mobility Consortium, which strives to clarify energy implications and opportunities related to advanced mobility solutions. The survey was developed, administered, and analyzed by a collaborative team of researchers at Lawrence Berkeley National Laboratory (LBNL), Idaho National Laboratory (INL), and National Renewable Energy Laboratory (NREL).

2.1 Survey Administration Details

The WholeTraveler Transportation Behavior Study was designed to generate insights about the underlying drivers of and barriers to different types of transportation behaviors. Of particular interest were preferences around emerging transportation technologies and services, such as vehicle electrification and automation technologies, ride-hailing, car sharing services, and e-commerce engagement and preferences. The data collection for this study progressed in two phases. Phase 1 was an online survey followed by Phase 2, which was the location data collection phase which is the focus of this report.

2.1.1 The Phase 1 WholeTraveler Transportation Behavior Study Survey

The full WholeTraveler survey instrument can be found on https://livewire.energy.gov/project/wholetraveler, and in the supplemental material of Spurlock et al. (2019). The survey included questions around each user's travel behaviors, mode choices, preferences over mode attributes, commute locations, car ownership, e-commerce behavior, and interest in new mobility technologies and services. It also included questions associated with demographic and household characteristics, personality traits, risk attributes, discount rates, and a life-history calendar that looked at life events and travel behaviors undertaken while the respondent was between the ages of 20 and 50. In addition, the following published papers and reports describe results from analyses of the online survey from the WholeTraveler Study. Spurlock et al. (2019) presents results from a broad analysis of a variety of factors that predict adoption of and interest in a wide range of emerging transportation technologies and services. Spurlock et al. (2020) presents results form an analysis of e-commerce behavior by study participants, assessing the extent to which home deliveries of four different categories of products replace or are in addition to household shopping trips. Jin et al. (2020) presents results form an innovative analysis using the life history calendar data from the survey; respondents were clustered into archetypal lifecycle trajectory patterns and the impact of different life stage transitions (e.g., having children) on mode use, differentiated by these different trajectory patterns, were analyzed. Finally, Department of Energy (2020) is a broad capstone report summarizing information about the WholeTraveler project overall, as well as a number of analyses conducted using data from this project including some using the life history calendar data.

The administration of the survey began with the identification of a sample of randomly selected addresses in the nine Bay Area California counties (Alameda, Contra Costa, Marin, Napa, San Francisco, San Mateo, Santa Clara, Solano, and Sonoma). This sample was recruited to respond to an online survey via a mailed invitation letter followed by a reminder postcard. The invitation asked the household member above the age of 18 who most recently had a birthday to respond to the survey. To complete the survey, the respondent went online through a web browser on a desktop or laptop computer. The survey was only administered in English. Respondents received a $10 Amazon gift card for completing the survey.

Recruitment letters were sent to 60,000 households. Of these, 997 residents completed the entire survey, and 48 completed only the first portion of the survey instrument (the part used for this analysis) for a total of 1,045 responses (1.74%). All responses were completed during the period between March and June 2018, with a median completion time for those that finished the full survey of 28 minutes. The response rate, while low, is consistent with other implementations using similar unsolicited mailings to recruit, and with similar incentive payment levels. For example, the 2015-2017 California Vehicles Survey has a 1.5% response rate overall (Fowler et al. 2018). Those who completed the survey were disproportionately highly educated and high income even within the San Francisco Bay Area.

Those who completed the initial online survey, referred to as Phase 1 of the WholeTraveler Study, were then offered the chance to complete Phase 2 of the study, which involved recording their movements and travel for one week using Google Location History that is tracked using the Google Maps App on either Android or Apple mobile devices. Participants in this phase received an additional $20 Amazon gift card for completing the survey and uploading the requested data. Participants were asked to provide one week of "typical" travel from the research period and then asked about what modes of transportation they used during that week and why.

The remainder of this report summarizes the design of the Phase 2 WholeTraveler Study revealed behavior location data collection, the data collected in this second Phase of the study, and the processing steps undertaken to extract meaning from those data.

2.2 Location Data Collection During Phase 2 of the WholeTraveler Transportation Behavior Study

2.2.1 Options for gathering location data

To aid in determining the best possible approach for gathering the movement of study participants in the WholeTraveler Study, members of the WholeTraveler team at NREL investigated different methods for collecting GPS and location data. They considered options such as dedicated GPS sensors sent to every user, using phone GPS sensors via a dedicated custom-developed app, and the use of Google Location History readings on existing phones. Their entire study and analysis of the options is documented in Kwasnik, Carmichael and Isley (2019).

Based on this assessment, and after some further testing, it was determined that Google Location History data would be the preferred method used for the WholeTraveler Study. There were a variety of factors that contributed to this choice including: cost of implementation, timeline to implementation, burden and ease of completion for participants, Institutional Review Board (IRB) human subjects research considerations, and the nature of the data that would be collected. WholeTraveler contributors also at NREL developed a custom web application to collect and parse the data. The data were collected and housed on a secure LBNL server. The database created to house the collected location data was administered by researchers at INL.

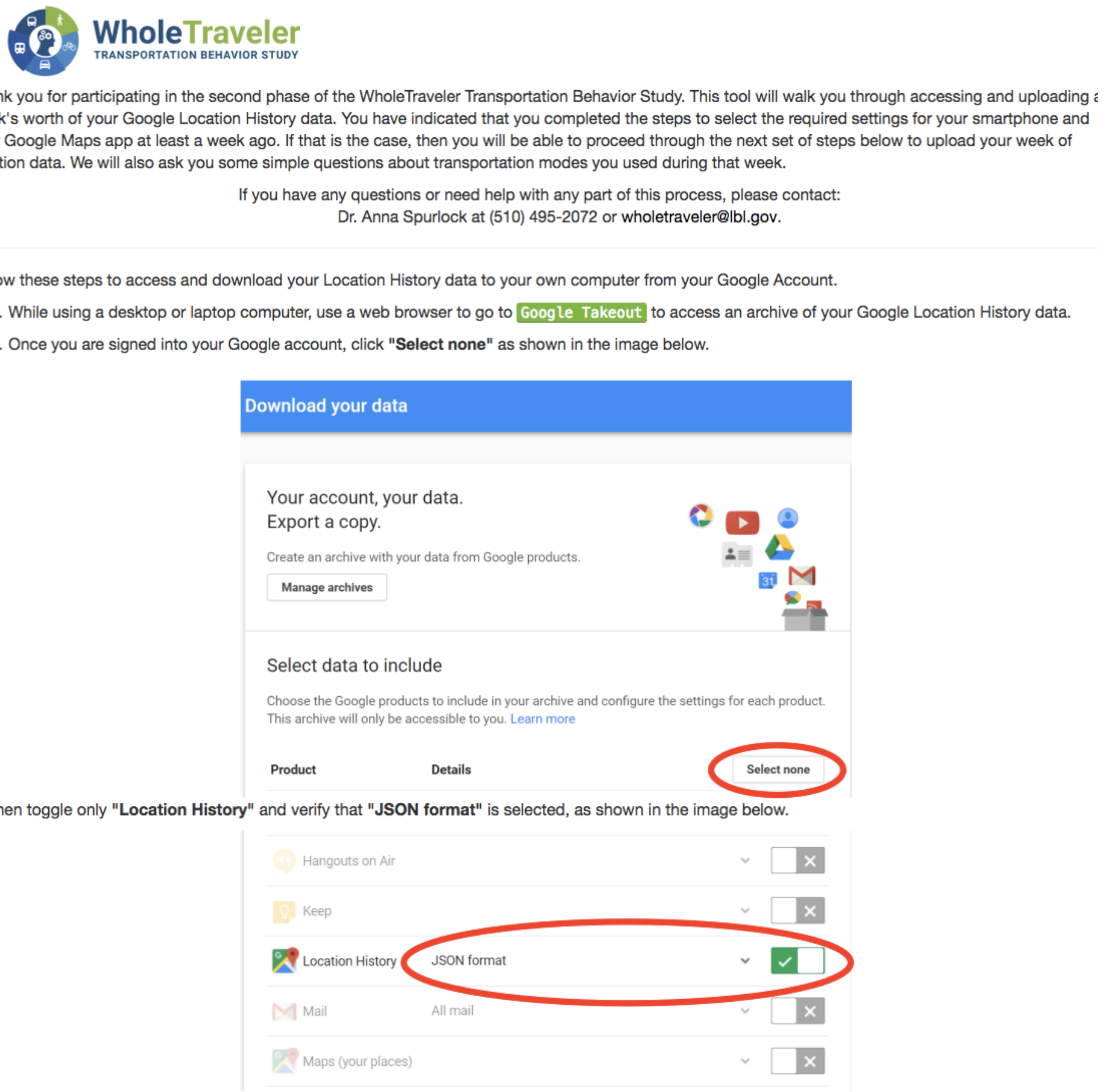
2.2.2 Location data collection process

The data collection process proceeded as follows. Participants, upon joining Phase 2 of the WholeTraveler Study, were provided with step-by-step instructions for how to configure the settings in their Google Account via their Google Maps app and on their smartphone. This would enable Google to collect and store their location history data. The instructions included links to a web site where they could see screen-shots for how to enable the feature on both iPhone and Android devices. The instructions needed to be updated during the collection period as the Google Maps application changed during that period.

The survey requested that the participates collect and share data for one "typical" week. Once sufficient time had elapsed for them to collect a week's worth of location data, an email prompt was sent to them with instructions for how they could access and download their personal Google Location History data file to their own computer. The download consisted of a single data file in a JavaScript Object Notation (JSON) format. A Uniform Resource Locator (URL) was provided, containing a random unique identifier that linked their email address to their submitted data on the LBNL server prepared for the purpose of collecting these data. Once the JSON data file was downloaded to the participant's computer, they then followed the provided instructions to upload the data file using the Java-based web application in their browser window. This web application allowed them to select and confirm the date range of the data they were agreeing to submit. Once they confirmed this information, they were guided through a short series of questions regarding the transportation modes they used during that date range, and the application then encrypted their data locally on their own computer and uploaded the data to LBNL's dedicated secure server.

The Google Location History JSON data file consisted of a series of location observations that captured the following variables: timestamp, latitude, longitude, accuracy, velocity, altitude, heading, vertical accuracy, estimated transportation mode taken, and the associated confidence with each mode. Not all timestamp observations captured all of these variables; it depended on Google's algorithms and transportation modes seem to only be collected by phones running the android operation system. In addition, the timestamp location observations were collected at variable time increments that ranged between 1 second to many hours. The timestamped location observation collection interval was also determined by Google's algorithms. More detail on the data is available in section 3.

The following series of images included in Figure 1 below show the instructions that users were presented with during the upload process, and the survey that participants were presented with after they had uploaded their location history data.



WholeTraveler
TRANSPORTATION BEHAVIOR STUDY

Thank you for participating in the second phase of the WholeTraveler Transportation Behavior Study. This tool will walk you through accessing and uploading a week's worth of your Google Location History data. You have indicated that you completed the steps to select the required settings for your smartphone and your Google Maps app at least a week ago. If that is the case, then you will be able to proceed through the next set of steps below to upload your week of location data. We will also ask you some simple questions about transportation modes you used during that week.

If you have any questions or need help with any part of this process, please contact:
Dr. Anna Spurlock at (510) 495-2072 or wholetraveler@lbl.gov.

Follow these steps to access and download your Location History data to your own computer from your Google Account.

1. While using a desktop or laptop computer, use a web browser to go to `Google Takeout` to access an archive of your Google Location History data.
2. Once you are signed into your Google account, click **"Select none"** as shown in the image below.

Download your data

Your account, your data.
Export a copy.

Create an archive with your data from Google products.

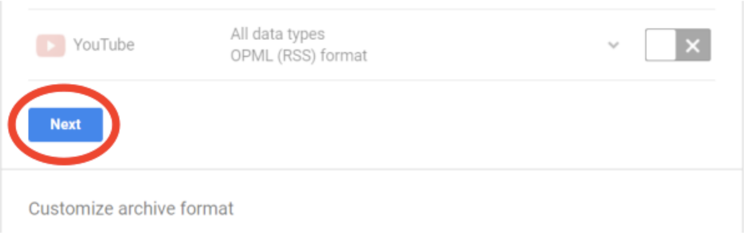
Select data to include

Choose the Google products to include in your archive and configure the settings for each product. This archive will only be accessible to you. Learn more

3. Then toggle only **"Location History"** and verify that **"JSON format"** is selected, as shown in the image below.

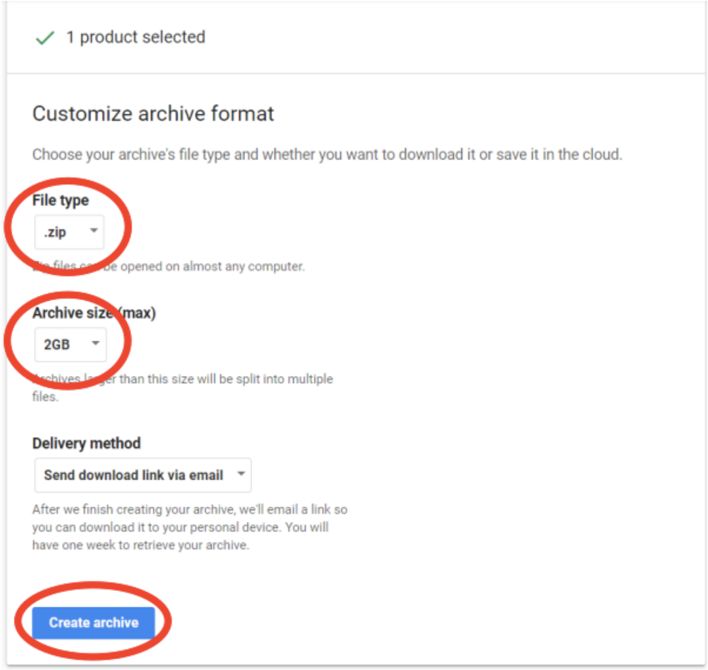
Hangouts on Air ☒
Keep ☒
Location History JSON format ✓
Mail All mail ☒
Maps (your places) ☒

4. Click **"Next"**, as shown in the image below.



YouTube — All data types, OPML (RSS) format

5. Confirm the file type is set to **".zip"**, the archive size (max) is set to **"2GB"**, and select the delivery method you prefer, then click "Create Archive", as shown in the image below.



✓ 1 product selected

Customize archive format

Choose your archive's file type and whether you want to download it or save it in the cloud.

File type: .zip

Zip files can be opened on almost any computer.

Archive size (max): 2GB

Archives larger than this size will be split into multiple files.

Delivery method: Send download link via email

After we finish creating your archive, we'll email a link so you can download it to your personal device. You will have one week to retrieve your archive.

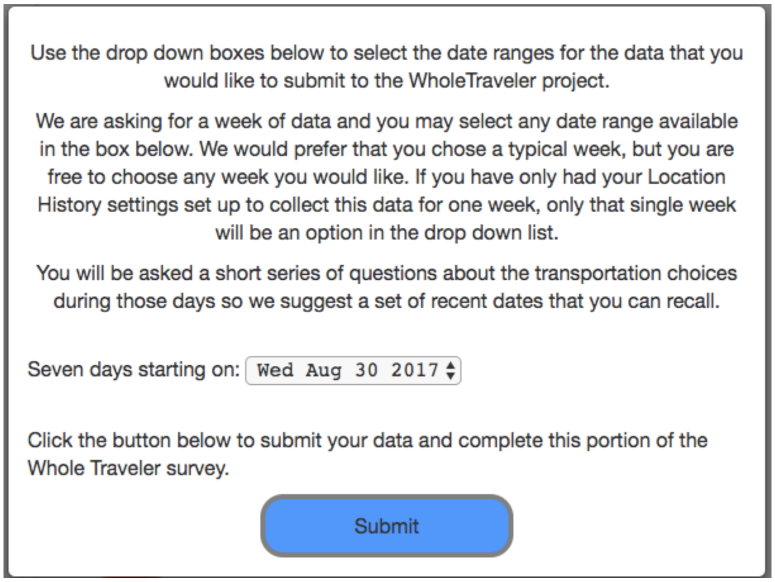
6. Google will prepare this archive of your data and send you an email with a link to download it.

Alternatively, you can stay on the same browser page and wait for the archive to be created. Then you can click **"Download"** once that option becomes available. You will have to enter your Google/Gmail password to download the file.

7. Once the file is downloaded, locate the .zip file where it was downloaded on your computer.
   It will be named something like "takeout-20170317T171533Z-001.zip"
8. Drag the .zip file into this browser window and drop it into the box below. After you have done so, you will have the opportunity to review and select the dat range of the data you want to submit before continuing.

Drop Google takeout.zip here

Complete the above steps to advance to the next section of this survey.



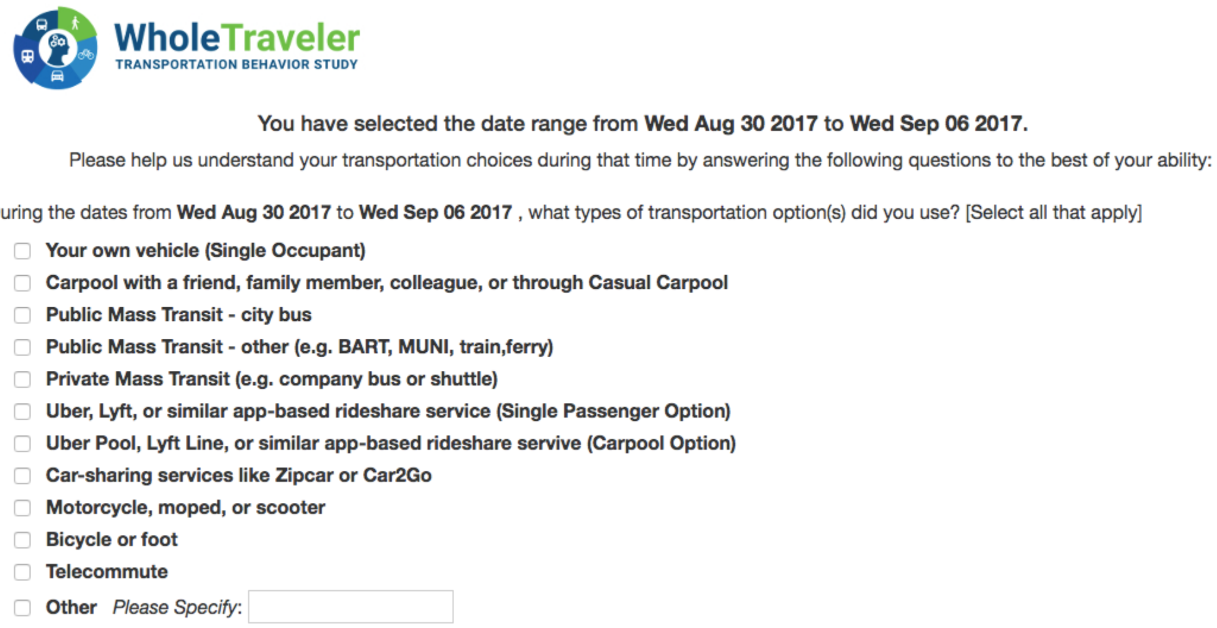
Use the drop down boxes below to select the date ranges for the data that you would like to submit to the WholeTraveler project.

We are asking for a week of data and you may select any date range available in the box below. We would prefer that you chose a typical week, but you are free to choose any week you would like. If you have only had your Location History settings set up to collect this data for one week, only that single week will be an option in the drop down list.

You will be asked a short series of questions about the transportation choices during those days so we suggest a set of recent dates that you can recall.

Seven days starting on: Wed Aug 30 2017

Click the button below to submit your data and complete this portion of the Whole Traveler survey.



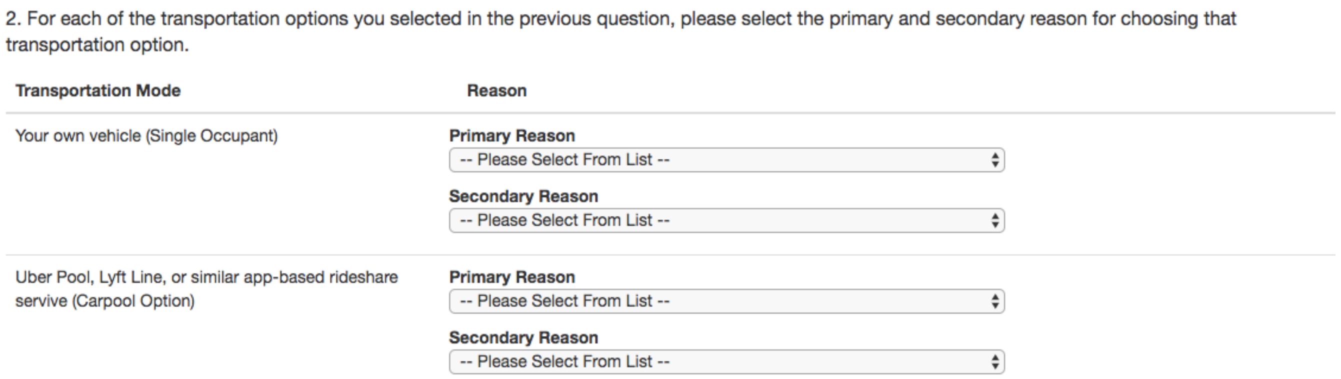
WholeTraveler
TRANSPORTATION BEHAVIOR STUDY

You have selected the date range from **Wed Aug 30 2017** to **Wed Sep 06 2017**.

Please help us understand your transportation choices during that time by answering the following questions to the best of your ability:

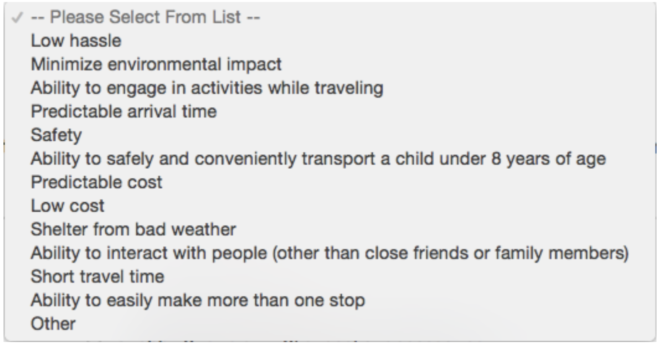
. During the dates from **Wed Aug 30 2017** to **Wed Sep 06 2017** , what types of transportation option(s) did you use? [Select all that apply]

- □ **Your own vehicle (Single Occupant)**
- □ **Carpool with a friend, family member, colleague, or through Casual Carpool**
- □ **Public Mass Transit - city bus**
- □ **Public Mass Transit - other (e.g. BART, MUNI, train,ferry)**
- □ **Private Mass Transit (e.g. company bus or shuttle)**
- □ **Uber, Lyft, or similar app-based rideshare service (Single Passenger Option)**
- □ **Uber Pool, Lyft Line, or similar app-based rideshare servive (Carpool Option)**
- □ **Car-sharing services like Zipcar or Car2Go**
- □ **Motorcycle, moped, or scooter**
- □ **Bicycle or foot**
- □ **Telecommute**
- □ **Other** *Please Specify*: ______

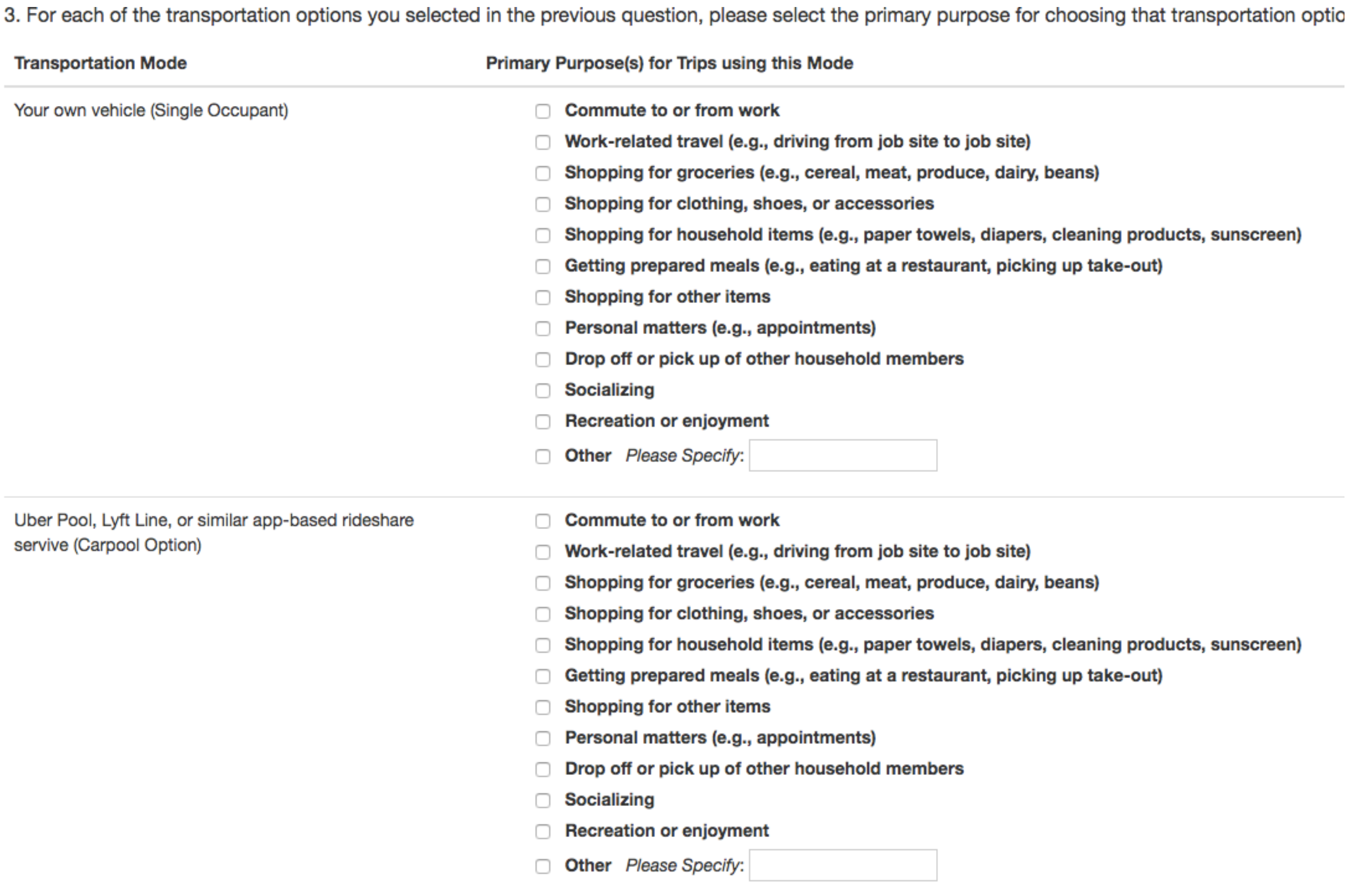


2. For each of the transportation options you selected in the previous question, please select the primary and secondary reason for choosing that transportation option.

Transportation Mode	Reason
Your own vehicle (Single Occupant)	**Primary Reason** -- Please Select From List -- **Secondary Reason** -- Please Select From List --
Uber Pool, Lyft Line, or similar app-based rideshare servive (Carpool Option)	**Primary Reason** -- Please Select From List -- **Secondary Reason** -- Please Select From List --

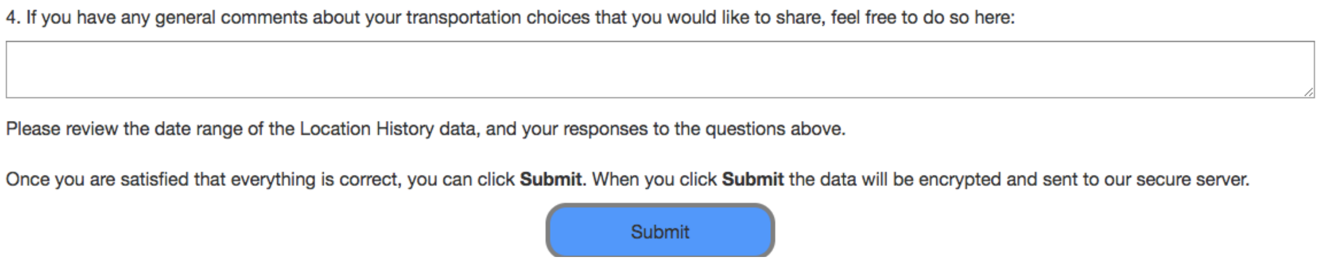


✓ -- Please Select From List --
Low hassle
Minimize environmental impact
Ability to engage in activities while traveling
Predictable arrival time
Safety
Ability to safely and conveniently transport a child under 8 years of age
Predictable cost
Low cost
Shelter from bad weather
Ability to interact with people (other than close friends or family members)
Short travel time
Ability to easily make more than one stop
Other



3. For each of the transportation options you selected in the previous question, please select the primary purpose for choosing that transportation optio

Transportation Mode	Primary Purpose(s) for Trips using this Mode
Your own vehicle (Single Occupant)	□ **Commute to or from work** □ **Work-related travel (e.g., driving from job site to job site)** □ **Shopping for groceries (e.g., cereal, meat, produce, dairy, beans)** □ **Shopping for clothing, shoes, or accessories** □ **Shopping for household items (e.g., paper towels, diapers, cleaning products, sunscreen)** □ **Getting prepared meals (e.g., eating at a restaurant, picking up take-out)** □ **Shopping for other items** □ **Personal matters (e.g., appointments)** □ **Drop off or pick up of other household members** □ **Socializing** □ **Recreation or enjoyment** □ **Other** *Please Specify*:
Uber Pool, Lyft Line, or similar app-based rideshare servive (Carpool Option)	□ **Commute to or from work** □ **Work-related travel (e.g., driving from job site to job site)** □ **Shopping for groceries (e.g., cereal, meat, produce, dairy, beans)** □ **Shopping for clothing, shoes, or accessories** □ **Shopping for household items (e.g., paper towels, diapers, cleaning products, sunscreen)** □ **Getting prepared meals (e.g., eating at a restaurant, picking up take-out)** □ **Shopping for other items** □ **Personal matters (e.g., appointments)** □ **Drop off or pick up of other household members** □ **Socializing** □ **Recreation or enjoyment** □ **Other** *Please Specify*:



4. If you have any general comments about your transportation choices that you would like to share, feel free to do so here:

Please review the date range of the Location History data, and your responses to the questions above.

Once you are satisfied that everything is correct, you can click **Submit**. When you click **Submit** the data will be encrypted and sent to our secure server.

Figure 1 - The WholeTraveler Phase 2 location data upload and survey process

As shown in these images, the application would guide the user through each step of creating the zip file, uploading the file, and choosing the date range. The application then parsed the file and collected all the JSON data associated with that date and 6 days forward. The file was then discarded so no other information was stored.

The survey then asked the user to identify what modes they used during the selected week. They were then asked to identify the primary and secondary reasons they had chosen each mode selected. It then asked them to select the primary purposes for which they used each mode of transportation. These options were intended to be matched with options from our Phase 1 survey questions. Details about the data that were recorded is in the following section.

3. Data Description

3.1 Location readings

The Google location data for phase two were collected for 301 participants over a range of about 4 months. Six of those participants had uploaded files that contained only data from before the study started (with dates as early as 2013), so readings were removed before the first possible date of the study, March 15, 2018. An additional 13 participants completed the data submission process but due to unresolvable technical difficulties with that process their data were not recorded in the server. This left 282 participants with usable location data and the statistics recorded below were from this final cleaned set. The earliest data recorded was on March 15, 2018 and the last data was recorded on July 25, 2018. The total number of points recorded was 360,012. The average number of locations per person recorded was 1,385. The minimum number of locations for a single person recorded was 42 readings, and the maximum was 23,355. Figure 2 shows the distribution of number of readings across study participants. This distribution varied greatly based on the mobile phone operating system used by the participants. Android phones collected more location readings while iPhones collected less, in part due to the fact that the location collection algorithm for iPhones seemed to favor collection to times when the phone was moving from one location to another. Figure 3 shows the differences in these distributions between these two platforms.

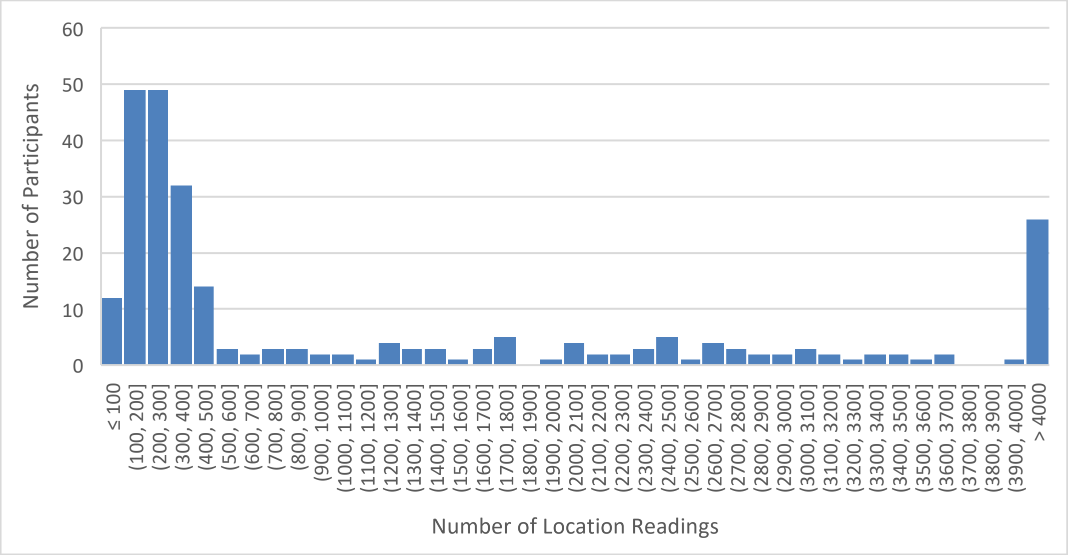


Figure 2 – Distribution of number of location readings collected across participants

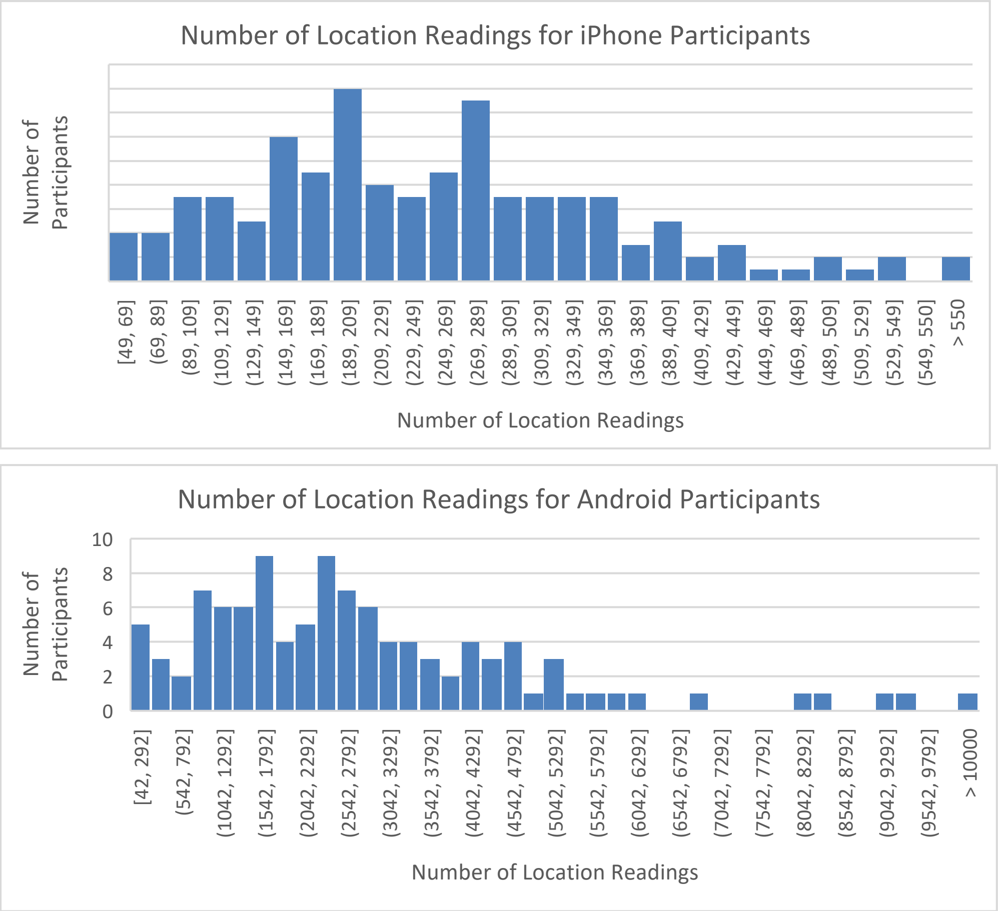


Figure 3 - Distribution of numbers of locations readings across mobile phone platforms

Some study participants, while residents of the Bay Area, traveled outside that region during the date range of data submitted. Figure 4 shows a map of all points collected during the study, showing the cases in which travel occurred outside the region. Figur 5 focuses in on the San Francisco Bay Area and shows a heatmap of the spatial distribution of points collected in that region.

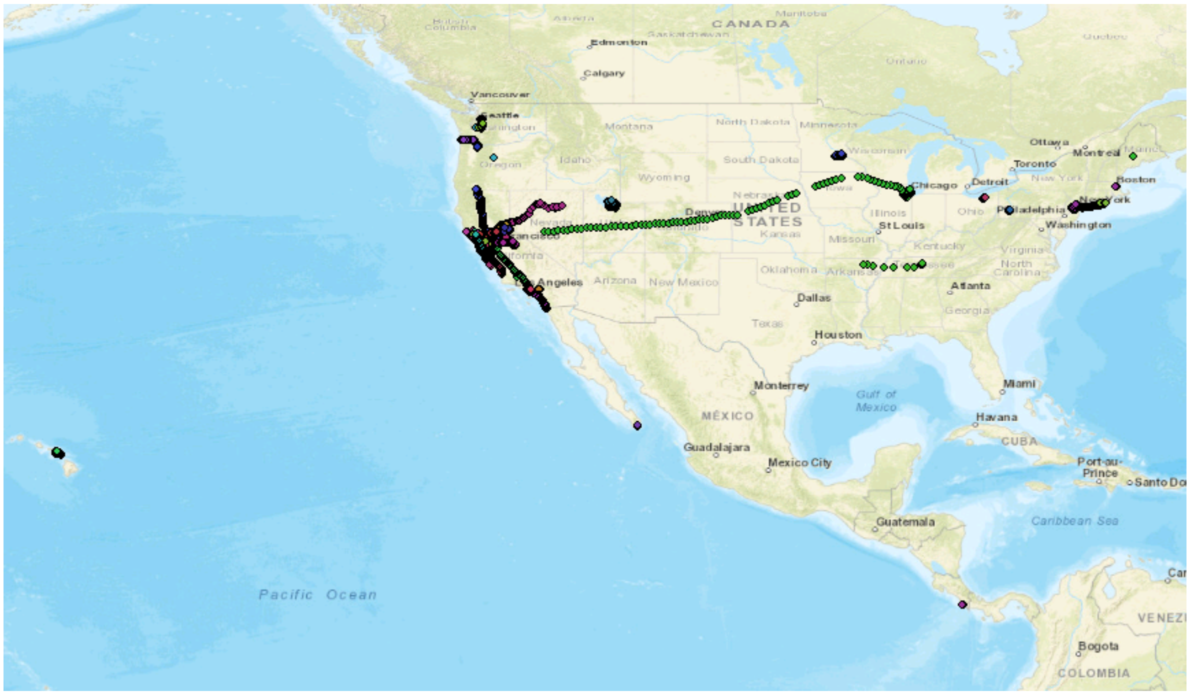
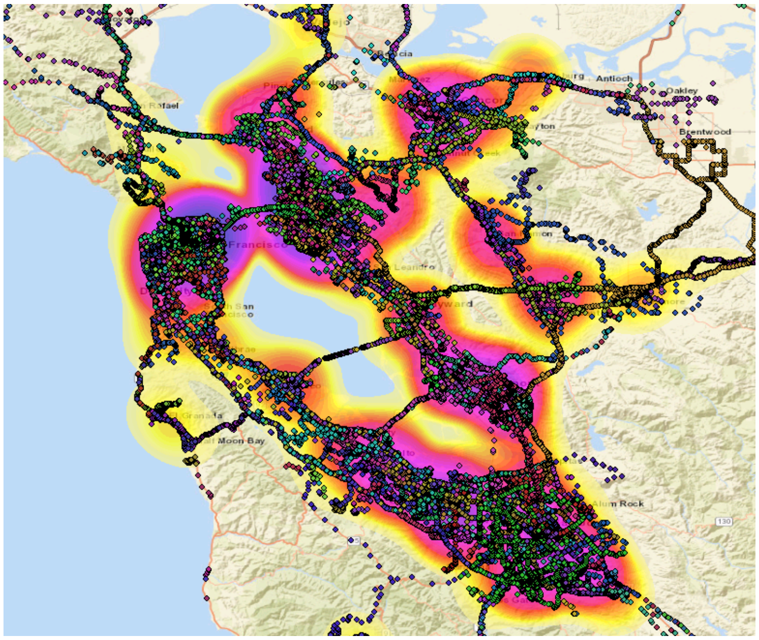


Figure 4 - Total map of location readings



Figur 5- Heat-map of number of readings in the San Francisco Bay area

3.2 Survey Results

One of the valuable elements of the survey was a report on what modes of transportation the participants used during the week that they recorded data. As shown above, participants were asked to specify what type of activities or trip purposes they used the transportation mode to perform, and what the key reasons were for why they selected a particular mode of transportation.

3.2.1 Transportation modes used and trip purpose

Figure 6 shows the number and percentage of users that selected that they used each particular mode of transportation during the week for which data were submitted. The most common modes used were personal vehicle (used by 83% of respondents), biking or walking (used by 54% of respondents), carpool (used by 36% of respondents), and mass transit (used by 34% of respondents). The least common modes used were private mass transit, motorcycles, and carshares.

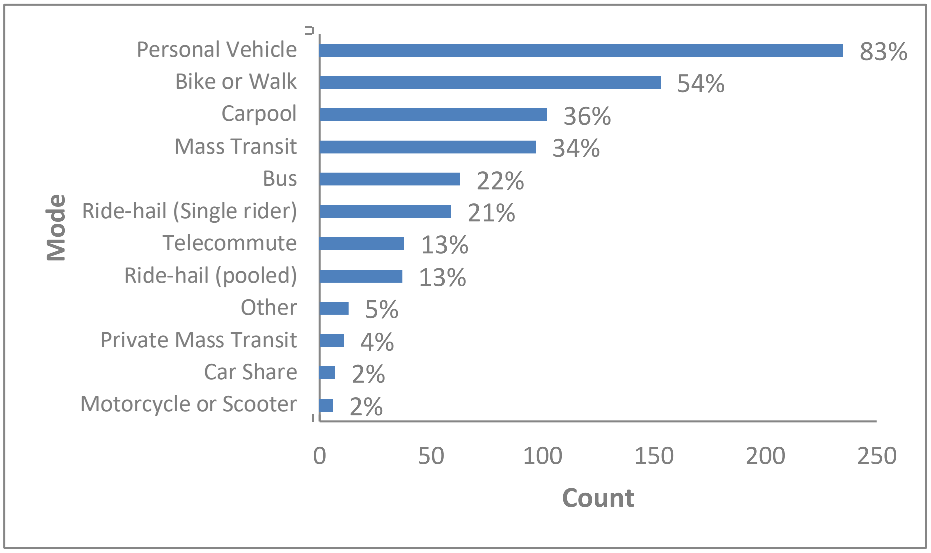


Figure 6 - Type of transportation modes used (percentage of users in bar label)

For each of the modes that the participant selected, they were asked to indicate the primary activities for which they were using that mode of transportation. Figure 7 shows the share of each mode used for each particular purpose reported. For many of the trip purposes, the travel was largely accomplished (over 50% of the time) in a personal vehicle, either alone or carpooling. This was especially true for dropping off or picking someone up (accomplished over 90% of the time in a personal vehicle either alone or carpooling) and shopping (almost 70% of the time in a personal vehicle either alone or carpooling). In contrast, over 50% of the respondents reporting trip purposes of socializing, recreation, and commuting accomplished those purposes via bicycling, walking, using public transit, or using ride-hailing. Figure 8 shows this same summary for shopping trips, separating out different types of shopping trips. Personal vehicles tend to be used for the majority of shopping trips regardless of the type of shopping, though this is particularly true for household item shopping trips, for which personal vehicles (either alone or carpooling) were used 75% of the time. In both Figure 7 and Figure 8 the "other" mode category includes private mass transit, car share, motorcycle or scooter, as well as the "Other" response that respondents were able to select. Telecommuting is not included in Figure 7 and Figure 8, and ride-hail and ride-hail (pooled) are combined.

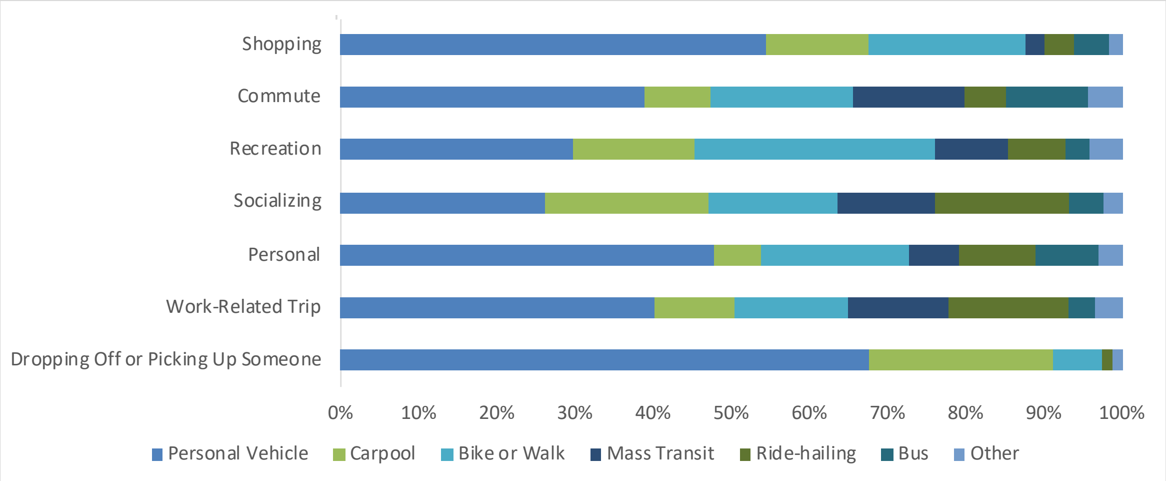


Figure 7 – Modes used for different trip purposes

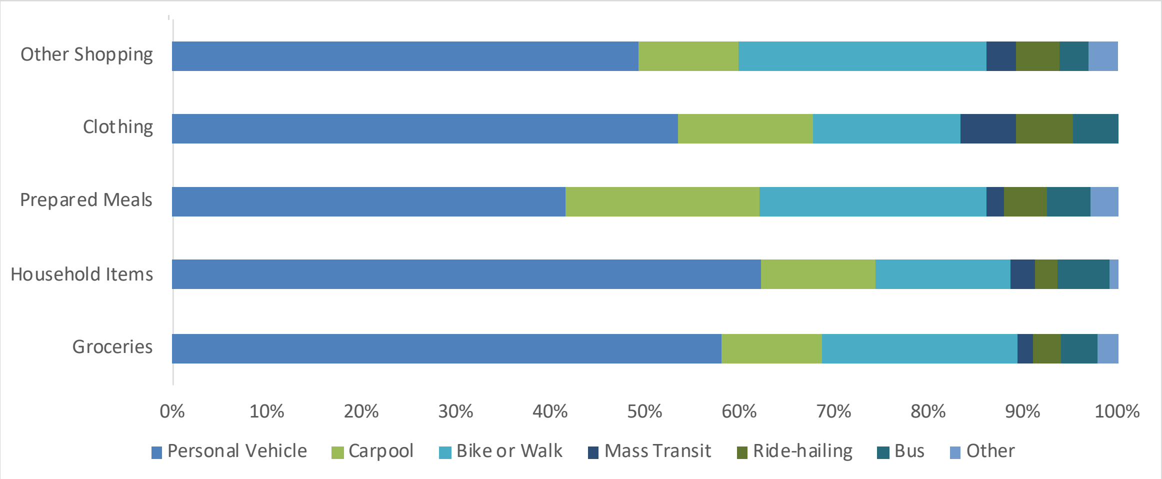


Figure 8 – Modes used for different shopping trip purposes

3.2.2 Reasons for using each mode

For each of the modes that the participant selected, they were also asked to choose the primary and secondary reasons that they chose that particular mode of transportation. The options provided were the following: low hassle, minimize environmental impact, ability to engage in activities while traveling, predictable arrival time, safety, ability to safety and conveniently transport a child under 8 years of age, predictable cost, low cost, shelter from bad weather, ability to interact with people, short travel time, ability to easily make more than one stop, and other.

Figure 9 shows the number of times respondents selected each of these reasons as their primary reason for choosing a given mode. Low hassle was the most frequently cited reason to take a given mode. Figure 10 shows this broken out by percentages. From Figure 10 it is clear that while low hassle is a highly desirable attribute, depending on the respondent there are a large number of types of modes that can be described that way. While close to 40% of the time low hassle was selected it was for a personal vehicle mode (either alone or carpooling), it was selected for other modes the rest of the time. In contrast, while the reasons "Ability to safely and conveniently transport a child under 8 years of age" and "Ability to make more than one stop" were not as frequently cited as a primary reason to take a given mode, those that did cite this reason overwhelmingly did so for private vehicle modes (either alone or carpooling). This suggests that for those for whom these two reasons are driving factors for their mode choice, there tends to be little alternative than to take a private vehicle. Not shown in these figures, but worth noting, is the fact that for those that indicated that they rode the bus, about a third of them selected "low cost" as the primary reason they chose that mode, suggesting that for many of these respondents, public buses fill a need for those placing a high importance on low cost.

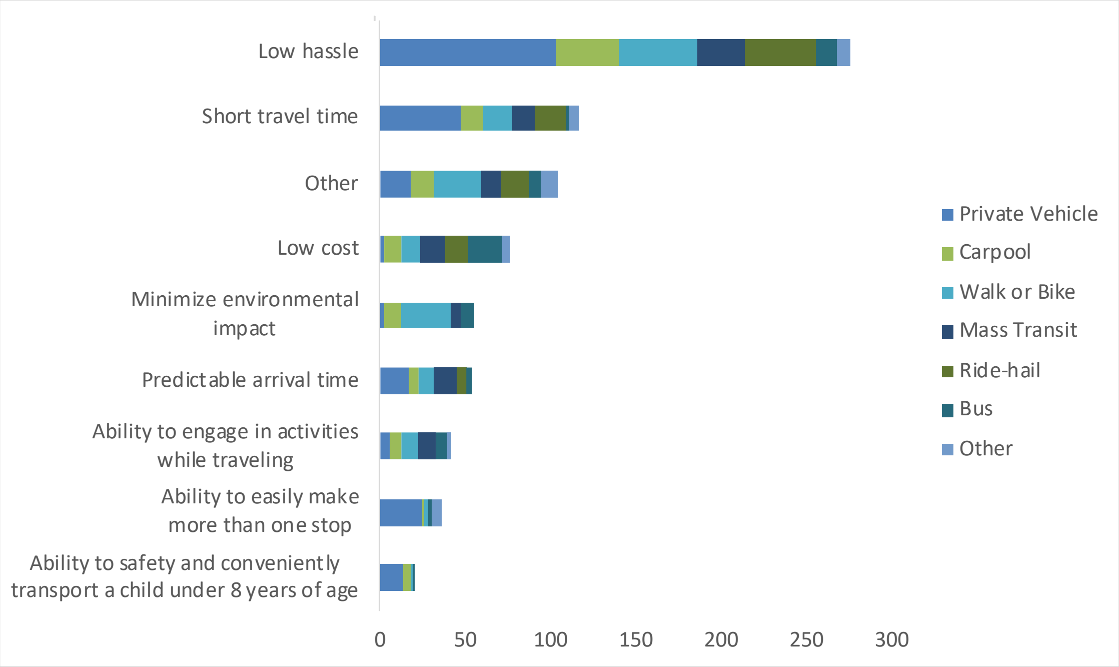


Figure 9- Frequency of each primary reason selected by mode

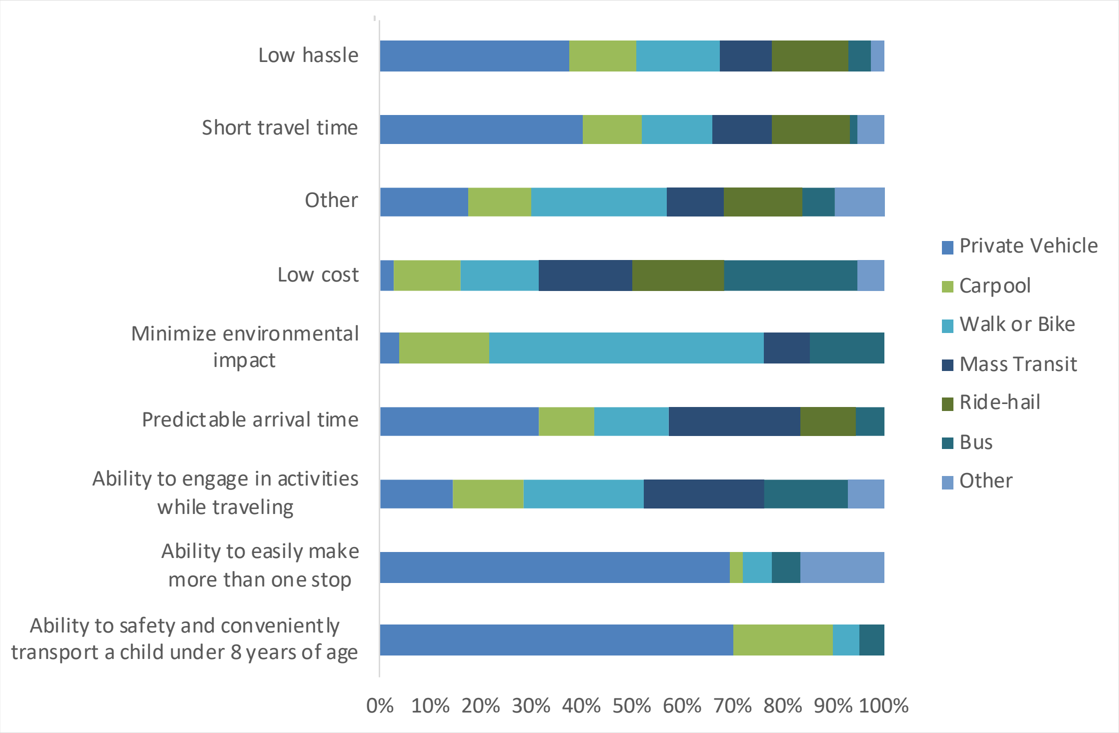


Figure 10- Mode share for each primary reason selected

4. Deriving meaning from location data

Location data points alone capture an unconnected series of moments in time and space. To understand travel behavior, the data needed to be analyzed and augmented. This included resolving issues with the data, adding information about each point encountered, converting the location readings into trips, and looking at sequences of behaviors such as trip chaining during commute trips. Methods used to address each of these are described in this section.

4.1 Data cleaning and accuracy considerations

All data contain some errors, noise, and imperfections. This is particularly true with location data from mobile phones. In this section the particular challenges associated with location data from Google Location History are described.

4.1.1 Google shared accounts

As the research team worked to evaluate the movements of users, they encountered many unrealistic readings which would indicate jumps of large distances in very short amount of time; sometimes jumping many miles in a matter of seconds. It was not initially clear why this was happening or the appropriate ways to filter for them. Eventually it was noted that some of the researchers on our team had shared a Google account with a family member so they could access the same google resources on both phones. It was quickly obvious that this was also the case with some survey participants, with some phones reporting location to the same Google account. Since they could be logged into two or more phones with the same account, the recordings uploaded contained readings that appeared separated by sometimes many miles.

Because this possibility was not considered before-hand, the survey did not include a criterion that a Google account be logged into on only one phone. Neither did it include instructions on how to avoid this kind of issue during the setup phase. Data evaluation and research identified that there was no programmatic way of distinguishing between the phones that were reporting if there was more than one phone logged into the account. This meant that it was very difficult to separate the actions of one phone-holder versus the other. To help alleviate the issues, the researchers took the step of flagging accounts that were suspected of being shared and removed them from several evaluations and reports.

Participants suspected of sharing an account were initially identified by checking the speed of motion that would be required to move from one point to the next in the submitted points (dividing the distance of succeeding location points by the time between the recordings). If motion repeatedly indicated speeds in excess of 200 mph then researchers would manually look at the locations reflected in the reporting. If the observed locations seemed to indicate more than one user providing location information, then a flag was set on the primary participant record. This analysis resulted in 22 users being flagged as suspected of this type of Google account sharing. The choice to flag the account was generally conservative so there may be others that have this condition that were not flagged.

4.1.2 Accuracy issues

There also seemed to be several issues with location data recordings in general. Some recordings had very close time gaps while others had gaps of several hours. Some phones only reported a few points a day while others reported many an hour. These issues impact the ability to perform meaningful analysis. Thus, several of the analysis methods attempted to mitigate problems due to this variability in data accuracy and frequency.

In particular, the location data submitted includes an estimated accuracy of the data points. This reflects the radius in meters around the recorded point that the true location may be. The higher the “accuracy” number the less accurate the reading is as it could be anywhere inside a larger diameter circle.

Over 39,000 of the readings reported an accuracy over 400 meters (approximately a quarter mile). This represents approximately 8% of the total number of readings. In most of the analyses performed, this was treated as an indication of poor data quality and these points were excluded.

4.2 Data augmentation with location queries

To help enhance the value of the location data, the data were supplemented with queries about the reported locations. Each of the recorded GPS locations was submitted to a Google "place" query through the Google Application Programing Interface (API). The query returned the physical address closest to the GPS point, as well as a list of places and place types which were close to the location.

The physical address data was added to each location data point. The place types and names were returned as a matrix with a list of different addresses and location types that were close to the requested point. This could include items such as residential address, business types (such as grocery store, bank, auto-mechanic, etc.), or transportation-based locations such as a bus stop. For purposes of analysis, the first returned value from the nearby places matrix was recorded as the location type for that location record.

4.3 Trips

To better understand mobility behaviors, it was necessary to convert the location information into movement information. To do this, the analysis defined an approach to convert the location data into "trips." Trips are intended to represent each time the traveler moved from one location to another. This helps to define the types of mobility that users participated in and helps characterize their behavior patterns.

Defining trips from the Google location data proved difficult. A location-based approach was first used to define trips. However, anomalies in the data reporting methods required many changes and refinements to the approach. Later, a trip definition based on movement and motion (distance between each subsequent point) was also assessed which was used to compare trip definitions. The two methods were used in some initial analyses and the comparisons between points helped refine the trip definitions. Both methods of trip definitions are defined below, but the resulting statistics are based on the location-based trip definitions.

4.3.1 Location-Based trips

A location-based trip is defined by an origin and a destination. Origins and destinations are places where the participant "dwelled" for more than 10 minutes. We define dwelling at a location as staying within a 200-meter-radius circle, or "buffer", during the specified time. Any movement from one dwelling location to another constituted a trip. Two hundred meters was chosen for this buffer value as it tended to account for minor drifts of the GPS readings and allowed for movements around local environments, like within a set of buildings at work, without necessarily triggering a new trip.

To create a trip, the algorithm identified the first point that the user reported as a new origin. The origin is assumed to be the center of the first location at which the participant was dwelling. Future points are compared against the origin point to see if they are still within 200 meters of the point (inside the buffer). If a point moves outside this buffer of 200 meters, then a new trip is started. In the comparisons, the algorithm takes into consideration the accuracy of the origin point and the accuracy of the comparison point. So, a trip only starts if the new point has a distance from the origin of more than 200 meters plus the accuracy in meters of the origin plus the accuracy in meters of the comparison point. When a trip starts, the algorithm records the difference in time from the first recorded location time at the origin to the start of the trip as the "dwell time," or the length of time that the participant stayed at the origin before leaving. The starting point of the trip is recorded as the first point that was recorded outside of the 200-meter buffer. Normally, the time this point was recorded is listed as the start time of the trip. However, if this point was after a gap of longer than 10 minutes from the previous reading, then the start time is set as the time recorded in the previous point. This helps mitigate issues where a phone may not record the initial movement or have been out of service for an extended time.

The end of a trip is determined by two primary criteria. First, if the location readings dwell within 200 meters for longer than 10 minutes (as described below) then the trip is considered ended and the dwelling location is labeled as the destination of the trip, and as the origin of the next trip. Alternatively, if during a trip there was a break in location reporting of longer than 20 minutes, then it is assumed that the trip ended at the point before the break. This was to remove some errors where the phone did not report for an extended period, sometimes hours, and may have related to the battery failing or the phone being turned off.

To determine whether the trip had reached an end destination, each point during the trip is considered as a potential end of the trip. For each point, the algorithm looked ahead 10 minutes to see if the participant was still within a 200-meter buffer of the current point. If so, then the current point is marked as the end of the trip. If it was not within this buffer after 10 minutes, then the algorithm moved to the next point and repeated the test.

To improve accuracy, the system discarded trips that were only 2 points – as these were often “jumps” in time and space with no information about the trip. It also discards trips less than 4 minutes, as they were likely related to GPS drift. Any trip with an average speed of over 200 miles per hour, which was potentially a result of shared accounts as mentioned above, was discarded. Data issues may still have created a number of anomalous results from trips, as many of them still tend to be short in duration and distance. Still, they can be used to determine a broad sense of the motion of the participants. The number of points that constitute a trip can be used to help determine how accurately captured a trip’s motions were.

4.3.2 Motion Based Trip

The motion-based trip approach uses an algorithm that treats motion as a function of physical displacement and time. Therefore, the process began by defining two fundamental measures: the time spent at each location in the data and the distance between adjacent location readings in the data. Broadly, locations were then assigned as one of 4 types: the start of a trip, movement within a trip, end of a trip, or being still.

The algorithm identified a location as the start of a trip if the recording moved more than 200 meters from that location to the next location in less than 10 minutes. Once a location had been identified as the starting point of a trip, the subsequent locations were marked as “moving” as long as the time spent at each location did not exceed 10 minutes.

One way that a trip ended, and a location was marked as the stopping point, was if the recording at that location was 10 minutes or longer. In addition, to account for dwelling in a small area, the algorithm considered locations where more than 10 minutes pass without moving more than 200 meters as the end of a trip, and the first location of this period is identified as a stopping point. Subsequent to a stopping point, locations are marked as “still” until there is greater than a 200-meter movement in less than 10 minutes.

To deal with potential data errors, the algorithm gave special treatment to a couple of instances of recordings. First, the algorithm identified instances when more than 20 minutes passed between two location observations and considers these as a “time jump.” If a time jump overlaps with an ongoing trip, then it is unclear if the participant had continued a trip or stopped the trip during the lapse (some lapses are over an hour). Therefore, the algorithm began by labelling time jumps to ensure that trips do not begin, end or overlap with them. This resulted in trips ending before a time jump.

At several points the data also experienced “location jumps,” where there were improbably large displacements for the corresponding length of time. The algorithm flagged locations if the implied velocity of the agent’s movements at those locations exceeded 200km/hour (around 124 miles/hour), as this is likely an upper bound of possible speeds for cars on the road. Since many of these movements could be based on GPS drift, the algorithm does not allow them to serve as candidates for the start of a trip, despite the sizable displacement. However, if these jumps occur during the course of a trip, the trip is not negated or ended as they often move back close to the original location soon after.

Despite the challenges, the algorithm identified trips fairly reliably. We assessed the spread of the trips relative to the participant's home and destination locations and found enough overlap to suggest that the identified trips reflect real travel patterns.

4.3.3 Trips statistics

The above-described location-based trip-detection algorithm resulted in 4,888 trips for 258 participants – or an average of approximately 19 trips per participant. Figure 11 shows the distribution of the length of these trips in minutes, Figure 12 shows the distribution of the trips by distance in meters, and Figure 13 shows the distribution of the trip start times throughout the day.

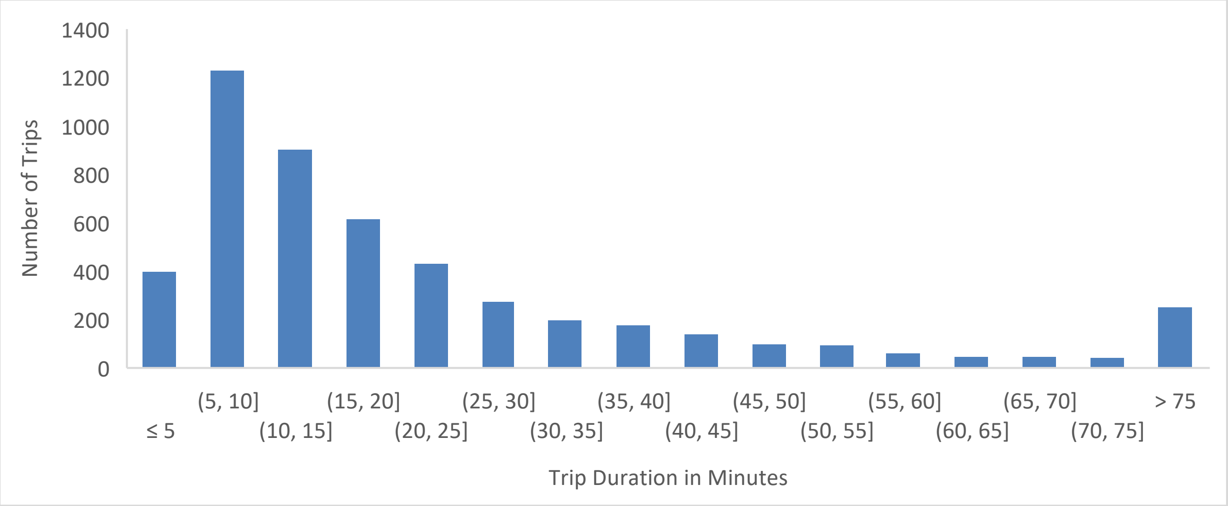


Figure 11 - Distribution of length of trips identified by the location method

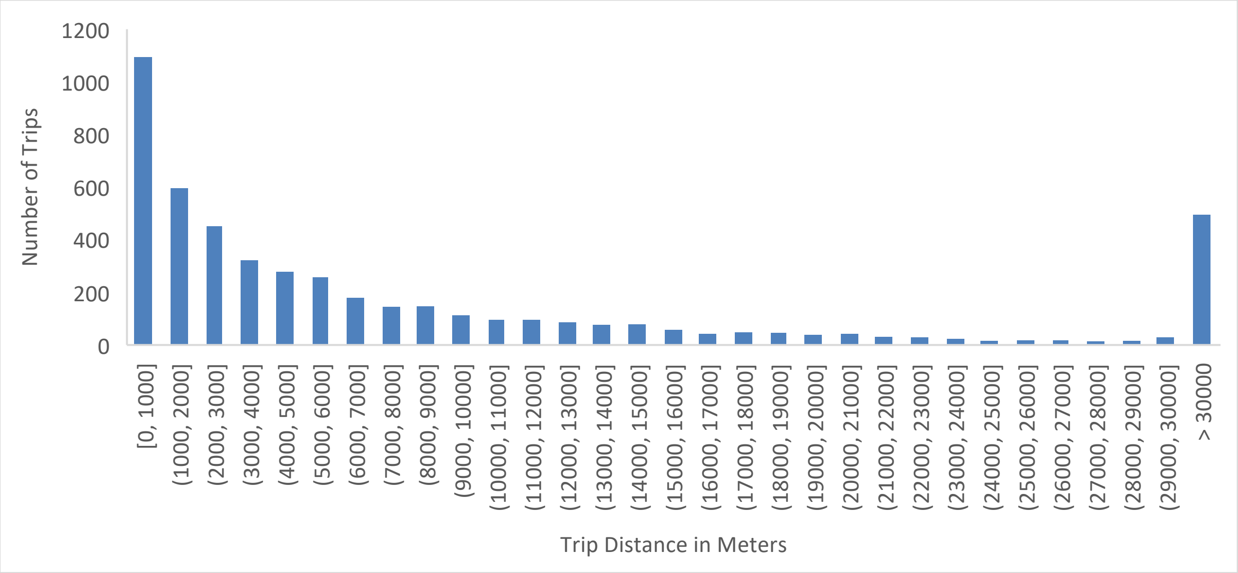


Figure 12 - Distribution of distance of trips in meters

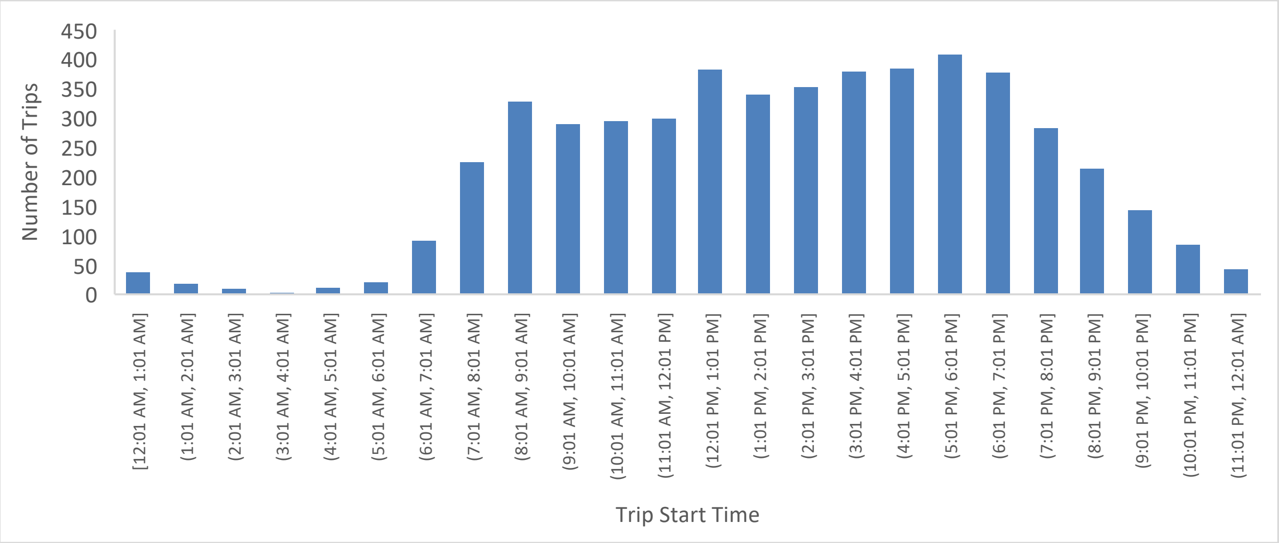


Figure 13 - Distribution of time of the day that trips began

4.3.4 Routes

To assist in understanding how mobility was spread through the area, a routes list was generated for each trip. To accomplish this, the street names for each location point was collected in order. Duplicate street names were merged so that the resulting list included streets visited on the trip in the order of visit. To ensure anonymity of the participants and maintain confidentiality, the first and last street names on the route were then removed.

4.4 Commute Trips

In the Phase 1 WholeTraveler online survey, participants were asked to provide a single “primary destination” to which they traveled most often. This didn’t have to be an employment location, as some respondents were students, unemployed, or retired. After they selected the destination, the participants also reported what type of location it was. In the Phase 2 data processing, commute trips were defined as the sub-set of trips that were identified as being between the participant’s home and their reported primary destination (PD) from the Phase 1 survey. A trip was defined as a commute trip if it started within 400 meters of the respondent’s home location and ended within 400 meters of their PD, or vice versa. This 400-meter cut-off value was chosen to account for the potential impacts of the algorithms that used a 200-meter buffer zone to define a start or end.

In characterizing commute trips, all individuals who had a primary destination that was already within 400 meters of their home were excluded. Direct commutes were cases where this journey was completed in one trip. Chained commutes were identified as a set of trips that would complete a commute circuit. The results of these classifications are described below.

4.4.1 Direct commute

Direct commutes are the set of trips that started at either the home and ended at the PD or started at the PD and ended at home. There were 566 total direct commute trips identified as shown in Table 1. The average duration of these trips was 42.8 minutes when traveling from home to the PD and 33.8 minutes when traveling from the PD to home. Figure 14, Figure 15, and Figure 16 show the distribution of the trips with respect to start time, duration, and distance, respectively.

	Direct Commutes	
	Number of Trips	Average Duration (min)
Home to PD	327	42.8
PD to Home	239	33.8
Total	566	

Table 1 - Direct commutes

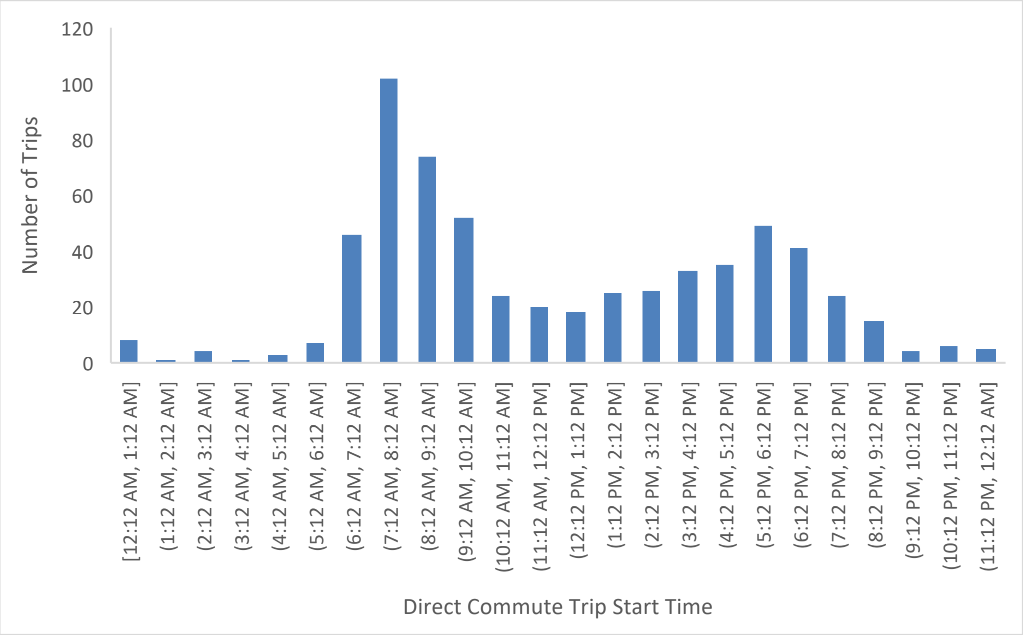


Figure 14 - Distribution of time of the day direct commute trips started

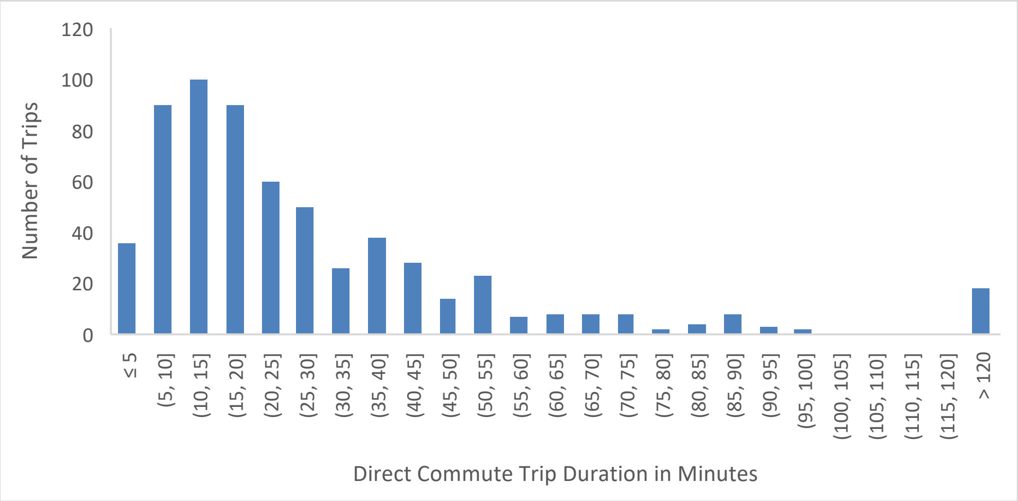


Figure 15 - Distribution of direct commute trip duration in minutes

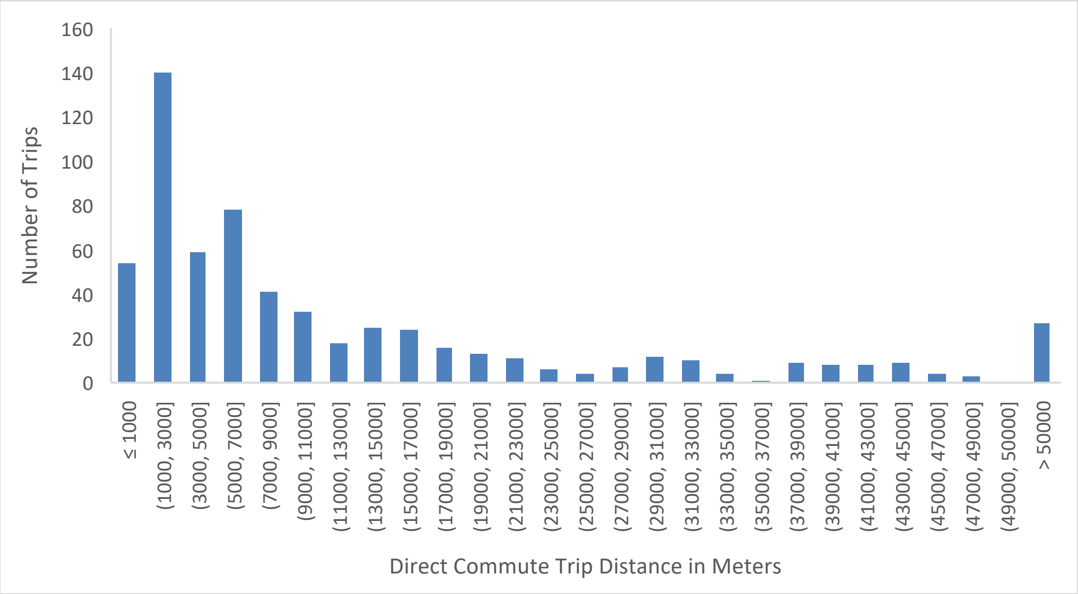


Figure 16 - Distribution of direct commute trip length in meters

4.4.2 Commute Chains

In contrast to a direct commute, where the commute was completed in a single trip, commute chains are defined as a sequence of trips that led from either the home to PD or PD to home, but with intervening stops. A stop is defined by the trip definition, which included dwelling for longer than 10 minutes at a location or where there was at least a 20-minute gap of readings during a trip. The chain was identified by looking at a set of trips where the start of the first trip was within 400 meters of the home or primary destination. The algorithm then collects the set of trips which leads to the other location (either home or PD). If one of the trips returned to the origin before reaching the other destination, then it was not considered a commute chain and was discarded.

Table 2 shows information about the identified chained trips. The chains average 2.62 trips per chain and the average duration is much longer than direct trips. As some trips are much longer and skewed the average, the median value is also included. Figure 17 and Figure 18 show the duration of the commute chains for both travel to the primary destination and to home.

	Chained Trip Commutes				
	# of Commute Chains	# of Trips	Ave Trips/Chain	Average Chain Duration (min)	Median Chain Duration (min)
Home to PD	73	191	2.62	212.5	116.0
PD to Home	123	322	2.62	207.3	120.6
Total	196	513	2.62		

Table 2 - Commute chains

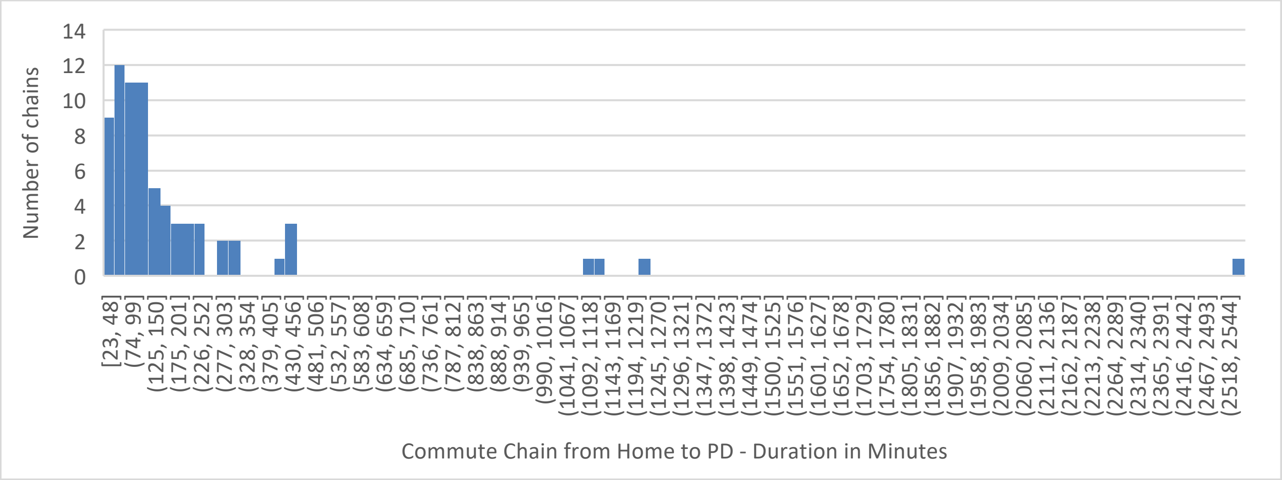


Figure 17- Distribution of commute chain total duration in minutes for chains from home to PD

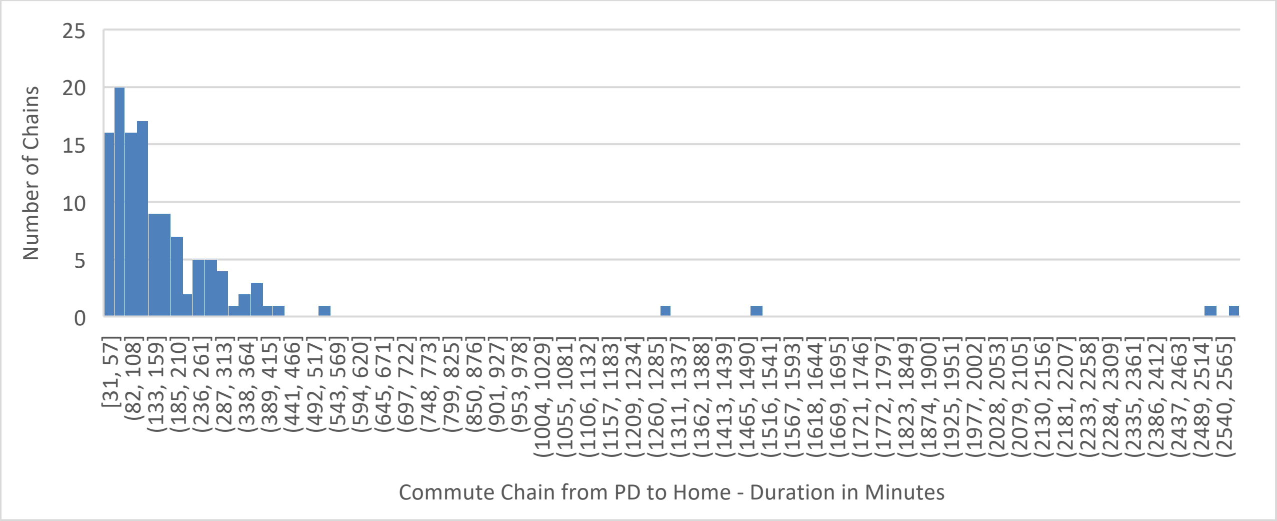


Figure 18 - Distribution of commute chain total duration in minutes for chains from PD to home

As a trip chain is defined by stops in the trips, Figure 19 and Figure 20 show the distribution in stop duration (gap) during the set of commute trip chains from home to the PD and the ones from the PD to home. The median value for the gap in trips from home to the primary destination is 35.2 minutes and from primary destination to home is 42.2 minutes.

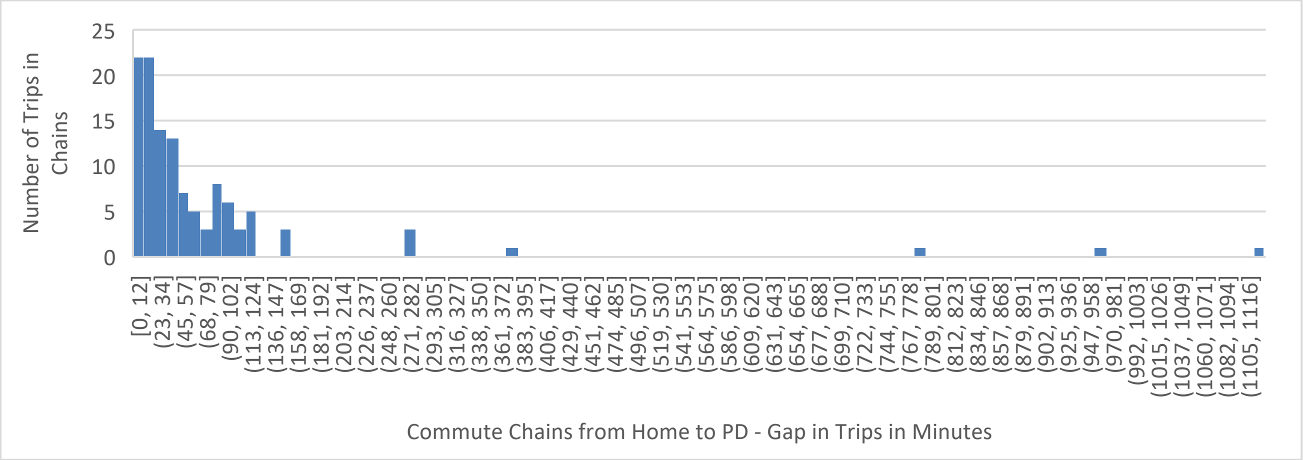


Figure 19 - Distribution of interim stop duration in commute chains from home to PD

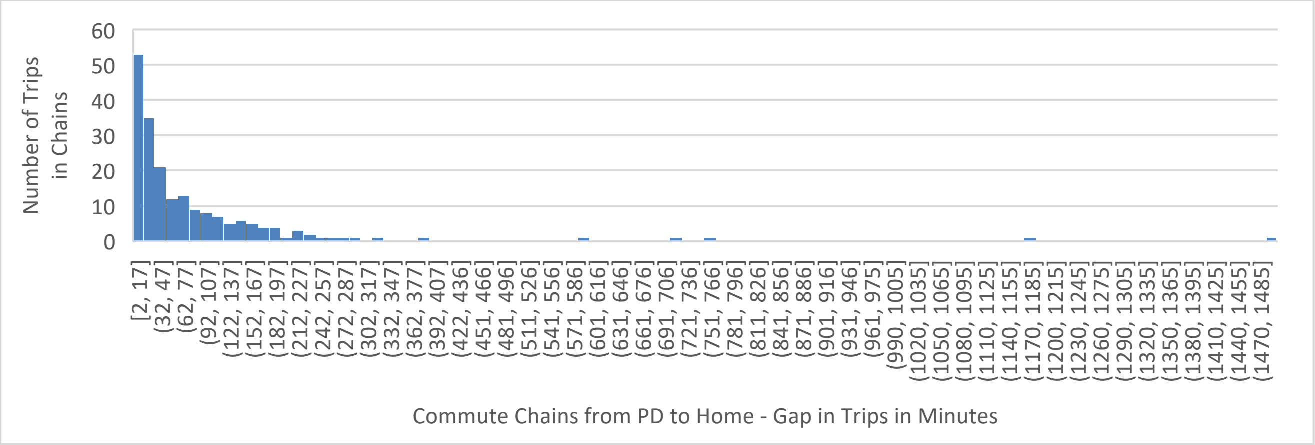


Figure 20 - Distribution of interim stop duration in commute chains from PD to home

5. Publicly Available Material

The data from both the Phase 1 online survey and the Phase 2 data summarized in this report are available as part of the Livewire system data repository. To request access to these data, go to https://livewire.energy.gov/project/wholetraveler.

5.1 Removing sensitive information

To maintain confidentiality and protect the identity of the survey participants, as per the requirement of IRB and human subjects research, the publicly available data have been cleaned of any potentially identifiable information through the following steps:

- The home and primary destination addresses and latitude-longitude locations were removed from the data.
- All names of businesses and places identified in the Google API look-ups were removed from the data.

- The specific latitude-longitude coordinates of the GPS points in the location data were removed, and instead the census block group was reported for each location point. In addition to the census block group, the following information was added into the data for each GPS location point: distance from last location point, distance from home, distance from primary destination.

5.2 Available Data

The data available on the Livewire platform contains an individual participant code that can be used to merge data between the various files, including between the Phase 1 online survey responses and the Phase 2 data files. The Livewire platform contains the following data and meta/data files:

- Background on the WholeTraveler Transportation Behavior Study survey and overview of data collected in Phase 1 (WT_phase1_background)
- Data dictionary for the Phase 1 survey data (WT_phase1_data_dictionary)
- Survey instrument for the Phase 1 survey (WT_phase1_survey_instrument)
- Background report on the Phase 2 locational data (this report) (WT_phase2_background)
- Data dictionary for the Phase 2 locational data (WT_phase2_data_dictionary)
- Survey instrument for the Phase 2 locational data submission (WT_phase2_survey_instrument)
- WholeTraveler Phase 1 Data Package:
    - Responses to the Phase 1 online survey (other than the life-history calendar portion) (WT_phase1)
    - Responses to the life-history calendar portion of the Phase 1 survey (WT_phase1_lifecyclecalendar)
    - Ancillary data collected on the location characteristics and travel times to various key locations and transit hubs for the survey respondents (WT_phase1_ancillary_locational)
    - Ancillary data collected from Google API on the details of a public transit commute from the respondent's home to primary destination and back at multiple times of day (WT_phase1_ancillary_GoogleAPI_public_transit_commute)
    - Ancillary data collected from fueleconomy.gov on the fuel efficiency of the respondent's reported primarily used household vehicle (WT_ancillary_fuelecon)
- WholeTraveler Phase 2 Data Package:
    - Overview data (e.g., number of days with location data collected, average trips per day, flag indicating possible shared Google Account) on the Phase 2 participants (WT_phase2_Participants)
    - Responses to the survey questions asked in the Phase 2 data submission (WT_phase2_SurveyData)
    - Data on the time-stamped location point readings (WT_phase2_Locations)
    - Data on the activities/location types collected from Google API for the location points (WT_phase2_Activities)
    - Data on the trips identified in the Phase 2 data (WT_phase2_Trips)
    - Data on the route details for the trips (WT_phase2_routes)
    - Data on the commute trip chains (WT_phase2_Commute_Chains)

References

Department of Energy, *SMART Mobility: Mobility Decision Science Capstone Report*. Department of Energy Office of Energy Efficiency and Renewable Energy (2020). https://www.energy.gov/sites/prod/files/2020/08/f77/SMART-MDS_Capstone_08.05.20.pdf

Fowler, Mark, Tristan Cherry, Thomas Adler, Mark Bradley, and Alex Richard. *2015–2017 California Vehicle Survey Consultant Report*. CEC-200-2018-006. California Energy Commission (2018). http://web.archive.org/web/20190601182103/https://www.energy.ca.gov/2018publications/CEC-200-2018-006/index.html.

Jin, Ling, Alina Lazar, James Sears, Annika Todd, Alex Sim, Kesheng Wu, Hung-Chia Yang, and C. Anna Spurlock. "Clustering Life Course to Understand the Heterogeneous Effects of Life Events, Gender, and Generation on Habitual Travel Modes." *IEEE Access* 8 (2020) 190964 - 190980. https://doi.org/10.1109/ACCESS.2020.3032328.

Kwasnik, Ted, Scott Carmichael, and Steven Isley. *An Overview of Technologies for Individual Trip History Collection: Mobility Decision Science Pillar SMART Mobility Consortium*. NREL/TP-6A20-70331. National Renewable Energy Laboratory (2018). https://www.nrel.gov/docs/fy19osti/70331.pdf.

Spurlock, C. Anna, James Sears, Gabrielle Wong-Parodi, Victor Walker, Ling Jin, Margaret Taylor, Andrew Duvall, Anand Gopal, and Annika Todd. "Describing the users: Understanding adoption of and interest in shared, electrified, and automated transportation in the San Francisco Bay Area." *Transportation Research Part D: Transport and Environment* 71 (2019): 283-301.

Spurlock, C. Anna, Annika Todd-Blick, Gabrielle Wong-Parodi, and Victor Walker. "Children, Income, and the Impact of Home Delivery on Household Shopping Trips." *Transportation Research Record* (2020): 0361198120935113.