This appendix provides additional explanation of the Connecticut Statewide Transportation Study (CSTS) Main datasets, beyond the labels for variables and values in the data dictionaries. This version of the data includes valid, fully complete households; it does not include "incomplete" households (only completed Part 1-Recruit Survey) and "volunteer" households. Only data for complete households were considered in the weighting analysis. A comprehensive list of data variables can be found in the data codebooks.

PRIVACY

This dataset guide accompanies full version of the datasets that include sensitive or confidential information. Specifically, home, school, work and trip (location) addresses and coordinates are considered sensitive information. Other potentially sensitive data may include the household income variables, vehicle fleet composition, names used by respondents to describe locations visited and open-ended comments (if participants shared personal information). As such, these datasets should be treated with utmost care and not shared or posted publicly.

Personally identifiable information that was only used for survey administration includes passwords, sample mailing addresses, email addresses, phone numbers, and person nicknames. These variables have been removed from the dataset.

OVERVIEW

A total of 8,403 invited households completed the entire Main study (Part 1 - Recruit Survey and Part 2 - Travel Diary Survey) online or over the phone in spring 2016. The remainder of this reference document only refers to the information corresponding to these complete households.

The data deliverable includes six distinct datasets:

1. Household-level data
2. Person-level data
3. Unlinked Trip-level data
4. Linked Trip-level data
5. Tour-level data
6. Vehicle-level data

Missing Data

Blank or null cells are intentionally missing data (e.g., a question was not asked, or an answer choice was not shown to the respondent). Only a small number of sensitive questions, such as household income, home ownership, and detailed age variables offered respondents the option to select "Prefer not to answer"; any "Prefer not to answer" responses have been coded as 99 in the dataset.

If data are missing because of a logic error or technical issue, the value was coded to -999 (unintentionally missing). One person in the valid dataset is missing school location data due to an error in the Google address geocoder. Due to the limits of the Google driving distance and duration calculator, 17 trips are missing Google distance and Google duration.

Weights

The household-level dataset also includes two sets of weights, namely, expansion weight and final weight. Expansion weight is based on the parameters of the sampling plan. Final weight is obtained by adjusting expansion weight to match marginal distributions of known household and person level variables of interest. Final weight has also been included in all of the remaining files to ease the process

of generating weighted distributions. Details regarding the expansion weight and final weight can be obtained from the final report.

DESCRIPTION OF DATA FILES AND KEY VARIABLES

Household-Level Dataset

The household-level dataset has 8,403 records from valid, complete households – one row per household.

Unique identifier: hhid

Example = 16100001. All hhid numbers start with '16', marking the year of the main study recruitment and data collection.

Sample segment (segnum)

Segnum variable refers to the household's assigned segment based on the sampling plan. For the Main study, the sample segment is based on the location of the mailing address to which study materials were sent.

Recruit survey start time, end time, duration (recruit_start_et, recruit_end_et, recruit_duration)

The recruit survey start and end timestamps are automatically recorded in the survey database (based on when the respondent viewed the first and last page of the survey). The recruit survey duration is calculated as the difference (in minutes) between the survey start and end times. Exclude extremely long durations when interpreting survey duration as the respondent may have paused the survey by leaving the website open and returned at a later time or date.

Household survey status (hh_hts_status)

Filter variable indicating whether or not each household completed Part 2 (the travel diaries). The dataset only includes valid, and complete households, therefore, all records assume a value of one.

Household number of trips on travel day (hh_tripcount)

The number of trips reported by each household is the count of trip records associated with each household's diaries (collected from all household members).

Home location variables (home_address, home_lat, home_lng, bg_geoid, town, county)

Respondents reported their home address during the recruit portion of the survey. Participants could search for either an address or an intersection, or they could place a marker on a Google map indicating their home location. The latitude and longitude of each home address were auto-calculated in the survey (in WGS84, the coordinate system used by the Google API). Other home location variables (block group ID, town, county) were derived from the spatial location of the reported home coordinates.

Household income variables (hhincome_detailed, hhincome_followup, hhincome_broad)

Households had the option of reporting income in twelve categories or selecting "prefer not to answer" (hhincome_detailed). If "prefer not to answer" was selected, a follow-up question offered the option of reporting income in one of five broad categories (hhincome_followup), though households were again allowed to select "prefer not to answer". A third derived income variable aggregated the responses from the two questions into a single variable with broad categories (hhincome_broad). Additionally, income was imputed (imputed_income_for_missing) for all observations with missing detailed income

(hhincome_detailed=99) information using a Multinomial Logit-based stochastic regression approach. Additional details regarding the income imputation can be obtained from the final report.

Call center completed recruit survey (callcenter_end)

If the user's IP address upon exiting the survey was determined to be located in Olathe, KS (the call center's location), the household is recorded as using the call center to complete the survey. Note: To protect privacy, the IP address is not provided with the dataset.

Foreign language household (recruit_nonenglish)

This flag was derived to indicate if the household was a foreign language participant. The household is flagged if their web browser used a non-English language setting – 17 households used a browser with a non-English language setting when completing the recruit survey.

Other derived variables (numadults, numkids, numstudents, numworkers, numdrivers, hhr_age, and transithh among others)

Other household-level variables were derived to summarize data collected at the person level. Additionally, other derived variables were generated to support weighting analysis and/or for reporting purposes.

Vehicle-Level Dataset

The vehicle-level dataset has 14,540 records from valid, complete households – one row per vehicle reported households that own vehicles.

Unique identifier: vehid

Example = 16100001101. The unique identifier for the vehicle-level dataset is the household's hhid (16100001) with a '1' and the vehicle number (101, 102, etc.) appended for each household vehicle. The '1' is appended to distinguish the unique vehicle ids from the unique person ids (see below).

Vehicle number (vehnum)

Vehicle number (1 through numvehicles). Unique within each household.

Person-Level Dataset

The person-level dataset has 17,481 records from valid, complete households – one row per person (all adults and all children). These person-level variables include data from both the recruit survey and the travel diary.

Unique identifier: personid

Example = 1610000101, where each personid is the household's hhid (16100001) with the person number (01, 02, etc.) appended for each household member.

Person number (person_num)

Person number (1 through hhsize). Unique within each household. Person number 1 completed the recruit survey.

Number of trips reported by person on travel day (person_tripcount)

The number of trips reported by each respondent is the count of trip records associated with each person's personid.

Person age variables (age, age_followup, age_broad)

Participants had the option of reporting individual members' ages in one-year increments or selecting "prefer not to answer" (age). If "prefer not to answer" was selected, a follow-up question offered the option of reporting ages in one of twelve broad categories (age_followup). A third derived age variable aggregated the responses from the two questions into a single variable with broad categories (age_broad).

School location variables (school_address, school_lat, school_lng)

The primary school address was reported in the recruit survey for each member who commutes to a school or daycare (excluding members who are homeschooled or only take classes online). As previously noted, the intent was to also collect school addresses for all participants who commuted to a school or daycare, but due to an error that data is missing for one person in the Main study dataset. As with the home addresses, participants could search for an address or intersection or place a marker on a Google map indicating their school location. Latitude and longitude of each address were auto-calculated in the survey (in WGS84, the coordinate system used by the Google API). Additional variables representing the corresponding block group ID, town, county were derived for the school location in the final version of the dataset.

Work location variables (work_address, work_lat, work_lng)

The primary workplace address was reported in the recruit survey for each member who commutes to a fixed workplace at least once per week. As with the home and school addresses, participants could search for an address or intersection or place a marker on a Google map indicating their work location. Latitude and longitude of each address were auto-calculated in the survey (in WGS84, the coordinate system used by the Google API). Additional variables representing the corresponding block group ID, town, county were derived for the work location in the final version of the dataset.

Travel diary start time, end time, duration (diary_start_et, diary_end_et, diary_duration)

As with the recruit survey, the diary survey start and end timestamps are automatically recorded in the survey database (based on when the respondent viewed the first and last page of the survey). The diary survey duration is calculated as the difference (in minutes) between the survey start and end times. Exclude extremely long durations when interpreting survey duration as the respondent may have paused the survey by leaving the website open and returned at a later time or date.

Foreign language travel diary (diary_nonenglish)

As with the recruit survey, this flag indicates if the diary was reported by someone who may speak a foreign language. A diary is flagged if the respondent's web browser used a non-English language setting – 36 diaries were completed on a browser with a non-English language setting.

Call center completed person's diary survey (diary_callcenter)

As with the recruit survey flag, if the user's IP address upon exiting the diary was determined to be located at the call center's location, that person is recorded as using the call center to complete their diary.

Proxy variable (proxy)

The proxy variable indicates if respondents took the diary for themselves, if other people filled out the answers while they were present, or if people filled out the answers and they were not present.

Copied trip variables (copytrips_confirm_none, trips_first, trips_last)

These variables indicate that the person had the opportunity to copy trips from previous household members. Copytrips_confirm_none is a flag indicating that the person was shown a list of copied trips from previous members; when coded '1', this means that the person chose not to copy any of the trips where other members reported them as part of the travel party; otherwise they copied at least one trip from a previous member. Trips_first and trips_last recorded the response to the question "Was [this copied trip] the first/last trip of the travel day?" If the respondent answered "No" to either question, they then reported the location where they started and/or ended the travel day (the same as members who did not have any copied trips).

Had additional trips to report (added_trip flags)

Indicates if the respondent went back and added more trips in the roster after seeing the prompt asking whether they had made any additional trips not already reported. Each variable name indicates a different type of trip a respondent could have added.

Other derived variables

Other derived variables were generated to support weighting analysis, for reporting purposes, and in response to data cleaning and processing carried out.

Unlinked Trip-Level Dataset

The trip-level dataset has 66,175 trips reported by valid, complete households – one row per (one-way) trip. These trip-level variables are reported in the diary for each person for their assigned travel date. The travel date starts at 3 AM on the assigned travel date and ends at 3 AM the following day.

Unique identifier: tripID

Example = 161000010101, where each tripid is the household's hhid (16100001) with the person number (01, 02, etc.) and the trip number appended for each person-trip (01, 02, etc.)

Trip number (trip_num)

Trip number (1 through person_tripcount). Unique within each person.

Trip copied from other household member (prepop)

Flag to indicate that a respondent copied this trip from another household member who had already reported them in the travel party on the original trip. This option is available to all household members (except for the first member to complete a travel diary), and reduces the respondent burden of repeating trip details.

Origin and destination location variables (o_address, o_lat, o_lng, d_address, d_lat, d_lng)

The destination addresses were reported in the diary survey for each trip to a new location. Respondents were only asked to record each location once; the addresses were automatically copied for each trip that returned to the same location (e.g. if a person left and returned home multiple times during the day). Additionally, respondents did not need to re-record locations for trips to home, work, or

school locations already recorded in the recruit survey – these addresses were prepopulated when respondents reported trip to these locations. Destination addresses were also automatically recorded for the origins of the next sequential trip.

As with the home, school and work addresses in the recruit survey, participants could search for an address or intersection or place a marker on a Google map indicating their trip locations. Latitude and longitude of each address were auto-calculated in the survey (in WGS84, the coordinate system used by the Google API). Additional variables representing the corresponding block group ID, town, county were derived for the trip destination in the final version of the dataset.

Origin, destination, and overall trip purposes (o_purpose, d_purpose, t_purpose)

Respondents report the destination trip purpose. The origin purpose is derived from the destination purpose of the previous trip, except for first trip in the day. For the first trip in the day, origin purpose is instead coded as home, work, or other, based on the origin-location description for the starting location of the day; it is most often home. Overall trip purpose was derived for modeling purposes to identify home-based, work-based, and other-based trips.

It must be noted that o_purpose and d_purpose went through significant data cleaning. Recoded variables for these and any derived variables were created. These are appropriately labeled in the final versions of the dataset

Trip departure and arrival time variables (depature_time, arrival_time, departure_time_hhmm, arrival_time_hhmm, departure_hour, arrival_hour)

Respondents reported the time when they departed (began traveling) on each trip and when they arrived at their destination (stopped traveling). They were asked to report their travel times in five-minute increments. Departure_time and arrival_time variables are the numeric codes recorded in the raw dataset (value labels are provided in standard time). For convenience, these raw codes were then recoded to text fields displaying the reported 5-minute increments in military time (departure_time_hhmm and arrival_time_hhmm). Departure_hour and arrival_hour are derived variables that place a trip's departure or arrival time into 1-hour bins for each hour of the travel day (again provided for convenience).

Reported trip duration (reported_duration)

Reported travel time is derived as the difference between respondent-reported start and end time of the trip. Minimum duration allowed by the survey is 5 minutes.

Google-estimated driving distance and time (gdistance, gduration)

The survey instrument estimated travel distance (in miles) and duration (in minutes) for each trip in addition to the user-reported travel time. The estimates of distance and duration were calculated using the Google Maps API Distance Matrix Service and indicate the distance and duration of a trip for "standard driving directions using the road network" (under free-flow conditions). Google estimated drive time and distance can be used to validate reported trip durations, but does not necessarily reflect the routes or modes used by the respondent (and consequently some differences in reported versus Google estimated travel times should be expected). As previously noted, 17 trips are missing this information when a trip was made where driving paths and times could not be estimated (e.g. overseas trips).

For 1,855 trips in this version of the dataset, the estimated distance and duration were 0 miles and/or 0 minutes. This means that the respondent had a trip where the origin and destination were in the same location or were very close together (such that the Google service rounded the driving time estimate to 0 minutes). Some of these trips may be considered valid trips, such as "loop trips" (e.g., a jog or bike ride) or a very short trip (e.g. a walk across the block); others of these trips could be erroneous (e.g., the respondent did not understand instructions or made some error in their reporting). However, there are no definitive criteria to indicate respondents' intentions when reporting these trips; therefore, analysts should evaluate these trips carefully to determine how and if they should be included in analyses.

Implied speed in miles per hour (implied_speed_mph)

Google-estimated driving distance over reported travel time. This variable can be used in trip validation to detect trip records with potential issues (i.e. extremely high or low speeds for a given mode may indicate a potential trip record with error). Similar to Google-estimated times and distances, outliers in travel speeds may be valid estimates of travel conditions (e.g. congested traffic). However, analysts may wish to evaluate these outliers when considering analysis of travel times and destinations.

Household member on trip (hhmember1 through hhmember9)

Household members who were reported as traveling on each trip. This was asked for households with more than one member. The member number (hhmember1, hhmember2, etc.) corresponds to the person number in each household (01, 02, etc.) A '1' indicates that the member was reported as traveling on the trip, while a '0' indicates that they were not a traveler on that trip. These variables were derived for "self" (the member the trip was reported for) for all trips from all households (including single-person households).

Linked Trip-Level Dataset

In the CSTS, individual legs of the travel episode were sometimes broken and reported as separate trips for multimodal journeys. For example, an individual going back home from office may have reported walking to the parking lot to get in his/her car and driving home as two separate trips. When in fact they should have been reported as a single trip pursued using the auto mode. The process of identifying these trip legs and consolidating them into a single trip is referred to as trip linking. A total of 65,103 linked trips were derived from the 66,175 unlinked trips reported by valid, complete households.

Unique identifier: linked_tripID

Example = 16100001010101, where each tripid is the household's hhid (16100001) with the person number (01, 02, etc.) and the first unlinked trip that is part of the linked trip chain appended for each person-trip (01, 02, etc.) and linked trip index (1, 2, etc.). If the trip was not part of a trip chain then the linked trip index was left as 0.

Other variables

All other variables in the linked trip file were derived from the unlinked trip file.

Tour-Level Dataset

The linked trips were then joined to form home-based tours and work-based sub-tours. A total of 22,434 tours were identified from the 65,103 linked trips. It must be noted that only full tours are included in this file. Partial tours (i.e. where trip chains did not start/end at the same location) are excluded.

Unique identifier: linked_tripID

Example = 1610000101201, where each tripid is the household's hhid (16100001) with the person number (01, 02, etc.) a dummy value of "2" and the index of the tour for the person (01, 02, etc.).

Other variables

All other variables in the tour file were derived from the unlinked trip file.