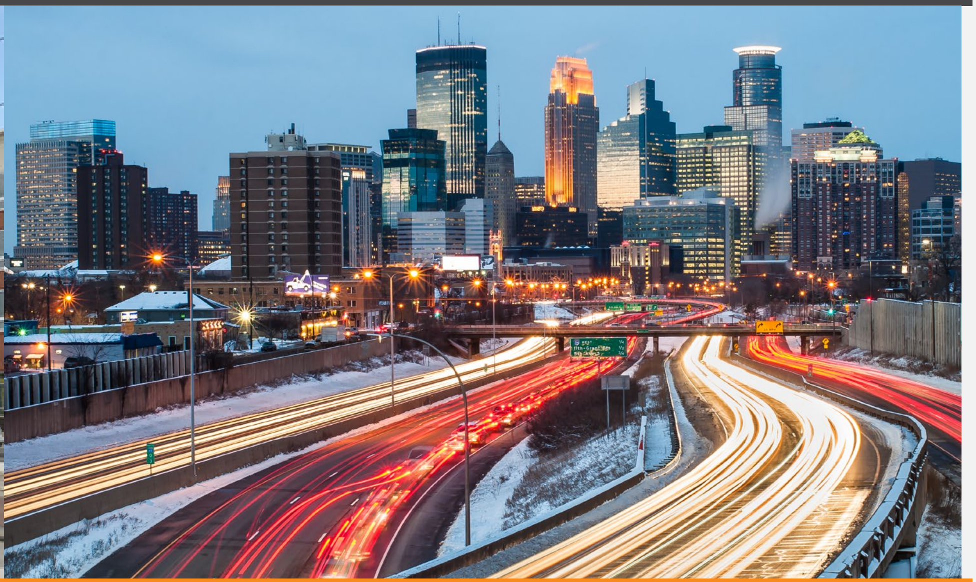
Metropolitan Council

2018-2019 TRAVEL BEHAVIOR INVENTORY

Technical Methodology Report



PREPARED FOR:
METROPOLITAN COUNCIL

SUBMITTED BY:
RSG

55 Railroad Row
White River Junction, VT 05001
802.295.4999
www.rsginc.com

Metropolitan Council
2018-2019 TRAVEL BEHAVIOR INVENTORY

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1.0 EXECUTIVE SUMMARY

This report documents the data cleaning, imputation, and weighting processes used to prepare the Wave 1 Travel Behavior Inventory (TBI) dataset for the Metropolitan Council. This report contains four sections, which are described below.

Trip-level data cleaning and processing

This section summarizes the methods used to clean and process the data in the TBI, particularly focusing on the methods used to clean and process rMove trip data. The following topics are described in this section:

- Data cleaning and processing, including automated data cleaning, splitting “loop” trips, and imputing habitual locations
- Unlinking transit trips
- Trip detection and “late pickups”
- Imputing departure times

Trip purpose imputation

This section describes the process used to impute trip purposes in the TBI rMove survey data in cases where (a) the reported trip destination purpose was not consistent with the trip destination location (home, work, school or other), or (b) where the destination purpose was not reported for the trip. This section gives an overview of the process and provides tables summarizing the results. Appendix A describes the purpose imputation rules in more detail.

Trip mode imputation

This section describes the estimation and application of the mode imputation model developed by RSG for the TBI study geography to be used as a validation of the trip survey reported mode. The sole purpose of this model is to be applied to person-days with all complete trips (no trips with unanswered surveys) and used to identify reported modes that are unlikely given the reported information. However, a secondary benefit is that the model can also be used to impute mode for trips where mode was not reported in the trip survey.

Weighting methodology

This section describes the analysis, recommendations, and methodology used to expand the data collected in the 2018-2019 TBI to the 2018 American Community Survey Public Use Microdata Sample (ACS PUMS) data. The weighting methodology applied adjusts for survey non-response, survey participation mode, and geographic bias due to oversampling and other factors.

2.0 TRIP-LEVEL DATA CLEANING AND PROCESSING

2.1 DATA CLEANING AND PROCESSING

This section summarizes the methods used to clean and process the data in travel survey datasets, particularly focusing on the methods used to clean and process rMove trip data. RSG initially “shapes” the raw rMove survey data into preliminary versions of the six tables described above (household, person, vehicle, etc.), ensuring administrative information (e.g., hh_id) and survey questions end up in the right tables (e.g., residence_duration goes in the household table).

Data cleaning and processing the trip data occurred in several sequential stages:

1. **Automated trip data cleaning:** The first stage of data cleaning employs a machine-learning algorithm to automatically classify trips that have a high likelihood of needing to be dropped from the dataset (false/spurious trips) or a high likelihood of requiring no edits. This algorithm is based on trip data that has been reviewed and labeled from previous datasets and is employed conservatively to minimize the rate of false positives (dropping trips that are valid) and false negatives (keeping trips that should be dropped). Any trips that do not have a high likelihood (80% or higher) of either classification are manually reviewed by analysts in the next stage.
2. **Manual spatial review and correction:** Analysts review trips to determine if one of three possible corrections should be applied to a trip:
    - **Drop/remove a trip from the dataset**: Invalid trips, e.g., a participant walking around their yard or a trip that was generated due to an errant Wi-Fi signal.
    - **Split a trip where an additional stop is apparent**: E.g., a participant stops to drop off another household member at school on the way to work. In these cases, the answers from the initial trip are applied to all resulting trips after splitting. Trip purposes are later re-derived based upon known home/work/school where possible.
    - **Join two adjacent trips where a stop is not apparent**. E.g., rMove loses signal on the highway and cuts out, but picks up a moment later further along the highway, or a golfing trip with many different segments. In these cases, the analyst chooses which trip survey answers should be applied to the resulting joined trip. Typically, the answers are the same for both surveys.
3. **Automated processing and derivations:** Various automatic trip modifications and derivations were performed on the initial cleaned dataset. These include:
    - Removing point locations with unreasonable derived speed or large accuracy radius (based on a proprietary algorithm) and removing redundant point locations that do not change the trajectories along a trip’s path.

- Unlinking transit trips where possible (described below).
- Splitting “loop” trips (e.g., a walk around the block) into an outbound and return trip to and from the furthest point. Loop trips are defined in these studies as trips that are more than 150 meters in total distance where the distance between origin and destination are less than 85 meters apart. Having distinct outbound/return trips, even on a valid loop trip, is often useful for building tours & other downstream modeling purposes. Purposes on these trips are coded based on a set of rules described below. Note that loop trips are only split once, i.e., trips with multiple looping patterns will not be split multiple times. Additionally, trips split during the unlinking process (described later) are not further split. These exceptions are to avoid over-processing trips beyond what the participant originally saw as a basis for their survey answers.
- Deriving variables such as whether the day was complete or not, the number of days the person completed, and other aggregate statistics (e.g., num_trips).
- Missing value coding based upon the survey questionnaire logic (e.g., which records should get a code of -9998 because the participant saw, but did not answer, a survey question).
- Appending geographic identifiers for latitude/longitude information, such as block group or county.

4. **Recoding purposes on split loop trips**
    - Purposes on outbound and inbound loop trips split during processing are coded with the logic in Table 2-1. The purpose “exercise/drop-off” is a stand in for exercise, recreation, drop off/pick up (or other escort purposes), and dining out. These purposes are singled out as being plausible “short dwell” trips that might cause a loop – e.g., stopping to pick up food (dine out), dropping someone off, or going for a run around the block (exercise). The logic accounts for cases where the user may have coded the entire trip (e.g., an exercise trip that starts and ends at home) with the purpose associated with the outbound leg (in this example, the outbound purpose would be coded to “exercise” and the inbound purpose to “home”, if the original origin purpose was “home”.) When these purposes are not part of the trip, a new “split loop purpose” is derived for the outbound leg, indicating that this purpose isn’t known.

TABLE 2-1: LOGIC FOR RECODING TRIP PURPOSE FOR INBOUND AND OUTBOUND LOOP TRIPS SPLIT DURING PROCESSING

ORIGINAL ORIGIN PURPOSE	ORIGINAL DESTINATION PURPOSE	AFTER SPLITTING, OUTBOUND TRIP ORIGIN PURPOSE	AFTER SPLITTING, OUTBOUND TRIP DESTINATION PURPOSE	AFTER SPLITTING, INBOUND TRIP ORIGIN PURPOSE	AFTER SPLITTING, INBOUND TRIP DESTINATION PURPOSE
Exercise/drop-off	Exercise/drop-off	Same as original origin purpose	Same as original destination purpose	Same as original destination purpose	Same as original origin purpose
Anything else	Exercise/drop-off	Same as original origin purpose	Same as original destination purpose	Same as original destination purpose	Same as original origin purpose
Exercise/drop-off	Anything else	Same as original origin purpose (may get imputed as a different purpose later)	Same as original origin purpose	Same as original origin purpose	Same as original destination purpose
Anything else	Anything else	Same as original origin purpose	Split loop purpose	Split loop purpose	Same as original destination purpose

5. **Recoding of open-ended trip purpose comments**
   - Some trip surveys include written descriptions of "other" trip purposes. RSG recodes a majority of these to detailed trip purposes by using keyword matches to determine the detailed trip purpose. About 5% of the trip purposes were originally reported as "other purpose" comments, the majority of which are now coded to have purposes besides "other". Data users can identify these comments by looking at records with values for d_purpose_other or o_purpose_other. These may get imputed as a different purpose later as part of the trip purpose imputation step.
6. **Habitual location imputation for primary home, work, and school locations**
   - As was decided in the study design phase of this project, this survey did not ask a separate geocoded location for the participant's primary home, work, or school addresses for participants who completed the recruit survey in rMove (participation group 3). RSG is imputing these primary locations based upon survey responses.
   - Primary home locations are those locations where:
     - The person indicated a trip purpose of going home and
     - The dwell time is the longest of such trips and is at least 3 hours or
     - The location is closer to the sample home address than any other location with an indicated trip purpose of going home and a dwell time of at least 3 hours. Candidate locations are all at least 150m apart.

- Work and school locations are derived similarly, except there is no minimum dwell time, and distance from sample home address is not used. Thus, primary work (school) locations are those where:
    - The person indicated a trip purpose of going to work (school)
    - The dwell time is the longest of such locations
- This imputation is for primary home, work, and school locations. People may have multiple home, work, or school locations, however RSG is deriving the "primary" locations.
- Data users should also look at the variables related to these imputed location fields, including second_home and residence_months for home location, job_type for work location, and school_type for school location. These variables provide important context for interpreting those results, including whether the person has a second home and how many months a year they dwell in the region; whether they have a singular work location, work from home, or have some other job location type; and what type of school they attend.

2.2 UNLINKED TRIPS AND TRANSIT TRIPS

What are unlinked trips?

This travel survey dataset is delivered in "unlinked trips" format. "Unlinked trips" are person-trips with each mode of travel identified as a separate trip and thus provided as a separate row in the trips table. By identifying unlinked trips, the dataset contains more detail about multi-modal trips that share a trip purpose (e.g., car->transit->walk to get to work). This is significant for transit trips, where access/egress modes and trip distances are important in understanding travel behavior. Other trips, such as car-to-walk trips (e.g., a long walk from the parking space to the destination) or air travel are also multimodal, but transit trips are typically the most important to capture unlinked in regional travel surveys. Many agencies later choose to "relink" trips or generate "tours" for their travel demand models. Capturing the dataset as a series of unlinked trips gives data users the maximum flexibility and detail when working with the data.

Capturing unlinked trips has several implications for the dataset. First, unlinked trip rates will always be higher than linked trip rates, although sometimes they will not differ by much. Related, given how total person-miles of travel should not change based upon the decision to report linked or unlinked trips, the average travel distances for unlinked trips should be shorter than for linked trips. Second, trip purposes of "change mode" are possible in unlinked trip datasets, but that trip purpose is notably reduced from linked trip datasets (since any "change mode" trips would be linked together). Third, because most transit access/egress is by walking, the walk mode shares tend to be higher for unlinked trips than for linked trips.

Finally, accurately and uniformly capturing unlinked trips is hard to do. Most people do not think of each leg of their journey as a separate trip. Ensuring that people accurately describe their travel modes and purposes for each leg is difficult. Research has shown that using smartphone-based travel surveys helps increase the level of detail captured (i.e., more short stops and transfers), but there are still limitations. People will sometimes incorrectly answer questions about their trip purpose by not answering "change mode purpose" or will merge two consecutive trips that, ideally, would have been captured separately. Additionally, there is evidence that walk-access-to-transit is captured more consistently via smartphone than walk-egress due to the expected dwell times while waiting for the train or bus during the access leg. The degree to which these imperfections matter often relates to the size of the transit mode share in a region. RSG continues to improve our travel survey methods to tackle these issues. The next section provides an overview of how transit trips are processed to handle these circumstances.

Additional steps to ensure transit trips are delivered as unlinked trips

This section describes the unlinking process for transit trips. RSG uses the Google Directions API for unlinking transit trips where transit routes are known by Google. Unlinking trips that used a transit mode helps to enforce consistency in the dataset. A linked trip would comprise, for example, a trip in which a participant walks to a bus stop (point A), takes a bus to point B, then walks to point C (their destination). The unlinking process splits this sequence into three separate trips to present a more accurate representation of regional travel behavior, including mode shares, trip distance, etc. In cases where the trip was captured as unlinked, RSG takes steps to avoid "double-counting" access or egress trips (step 5 below). In all cases, the original origin and destination locations and times are unchanged, there is just added detail in between. Finally, because this survey asked access and egress modes, those modes are incorporated into the transit trip unlinking process.

RSG's current transit trip unlinking methodology involves:

1. Selecting the set of trips that can be unlinked:
   - rMove trips with transit modes were eligible for unlinking, including bus and rail transit.[1]
   - Currently, the Google Directions API cannot return drive-to-transit or drive-from-transit directions, so **trips that involved a driving mode were excluded and are thus not unlinked**. Thus, **only trips that included a transit mode and/or walking and/or bike modes were unlinked in this process**. Often, driving trips to transit appear as separate trips due to transit wait times and do not need unlinking; however, vehicle egresses are harder to capture as distinct trips unless the user splits these trips themselves.
2. Comparing rMove trip data to the Google Directions API, including departure times and origin/destination locations.
3. Receiving the following information from the Google Directions API using the schedule of the next weekday that is the same as the weekday of the trip for up to four suggested trip routes:
   - Boarding/alighting location and time for each leg of travel.

[1] The following modes were eligible for unlinking: intercity bus, public bus, express bus, local bus, Bus Rapid Transit, employer/university shuttle, private bus, other bus, vanpool, medical transportation service, Metro Mobility, intercity rail, light rail, Northstar, and other rail.

- Assumed travel mode for each leg of travel (type of transit or walking/biking).
- The geometric “path” of the suggested route.

4. Selecting the most viable route from the suggested alternatives based on the degree of temporal and spatial overlap between the suggested route and that of the original transit trip being unlinked. The spatial overlap comparison uses a 100-meter buffer around each trip path (the path along rMove locations and the path provided by Google). This selection is done during the manual cleaning process, and the “best match” route will show up to data cleaners first (that with the most spatial and temporal overlap with the trip to be unlinked). Often this route matches the trip trace and will be selected at this stage.

5. Storing information for each of the unlinked trip segments of the selected transit route in separate rows. Very short unlinked trip segments (less than two minutes in duration) are discarded to avoid re-unlinking a trip that a user has already split into unlinked trips. The data associated with unlinked legs include:

- Timestamp and locations of the point nearest to the boarding/alighting points, stored as the “origin” and “destination” time and location for that leg;
- Origin/destination purpose, which are recoded to “change travel mode”, except for origin purpose on the first unlinked segment and destination purpose on the last unlinked segment;
- Mode and mode_type on access and egress trips resulting from unlinking are coded as follows:
    - If the user indicated that they accessed by transit, but the access/egress trip is flagged as a non-transit trip by the Google API, other selected modes are used to identify access and egress.
    - If no other modes are selected, the access/egress mode are coded as “walk”.
    - Whenever possible, the user’s input is chosen; otherwise the best effort is made to interpret mode information provided by the user in order to retain non-walking modes used for access and egress on unlinked trips.
- All other originally provided survey answers.

The variable “leg_num” is the leg number of the unlinked trip within the linked trip. RSG also derived a variable “linked_trip_num” to demonstrate which trips were part of an overall “linked” trip, whether they were unlinked by this process or prior to processing (that rMove or the user “unlinked”). The following logic was used to link trips:

- Trips that were unlinked by the process above were linked together;
- Transit trips with a “change mode” purpose and a dwell time less than one hour were linked to the next trip;

- Non-transit trips with a “change mode” purpose and a dwell time less than 30 minutes were linked to the next trip;
- Trips with a home or work purpose followed by a trip the same purpose with a dwell time less than 15 minutes were linked to the next trip (note that this was based on the originally reported purpose rather than the imputed purpose.)

Use of the “linked_trip_num” variable is up to the discretion of the data user. To determine trips that were unlinked in processing, refer to the variable “unlinked_trip”.

2.3 TRIP DETECTION AND “LATE PICKUPS”

This section describes the way rMove begins detecting a trip and why there are sometimes gaps between the origin location and subsequent location.

To avoid overly depleting a user’s battery, rMove will “ping” devices to get location information at a lower frequency when devices appear to be stationary versus when they are moving. rMove checks location every 3-5 minutes when a device is inside of a “geofence”, which is a radius established around the last known location. If the device leaves that geofence, rMove will detect this movement and start collecting location more frequently (every few seconds) until the user remains within a geofence for longer than 3 minutes. The last known location before the trip was detected is used as the trip start point.

Additional factors beyond rMove’s control can also impact the time it takes for rMove to detect movement outside a geofence. Battery saving features on certain devices, for example, can lead to rMove detecting movement when the user is far from their prior location, as well as environmental factors like underground travel. If the device travels more than one mile before rMove starts collecting locations, the user is asked to provide a start time (as rMove does not know when the user left the origin). Users can add a stop along the “pickup” distance for any trip if rMove missed a stop between when the user left the origin and when rMove began recording the trip.

In RSG’s datasets, roughly 10% of rMove-collected (non-user-added) trips in the dataset have a pickup distance of more than one mile. This varies by mode, where transit and LD passenger trips are more likely to have a pickup distance more than a mile. This is likely due to transit trips traveling underground and plane trips where a person’s phone is off or in airplane mode until they land.

2.4 IMPUTED DEPARTURE TIMES

As a result of the “late pickup” condition described above, recorded start times for trips that are not user added are not always reliable. rMove users can travel a significant distance before rMove recognizes that they’re making a trip, and this can yield invalid or extreme values for trip duration and speed. The fields depart_time_imputed, speed_mph_imputed, and duration_imputed provide updated values that can be analyzed where the original speed or duration appear invalid.

Departure time was imputed using median speed between all locations along the trip, excluding the origin point, and the distance between the origin and the next point on the trip. For trips with fewer than three recorded locations, imputed departure time is set three minutes earlier than the original departure time to compensate for rMove's 3-5-minute ping interval. Note that some trips that are the result of split loop trips may only have three or fewer points but will use the imputed depart time from before the loop trip was split and thus may not be included in this rule.

If the imputed departure time would overlap with the previous trip's arrival time, the previous trip's arrival time was instead used as the departure time. Regardless of the number of locations along a trip, if the imputed departure time was later than the initially reported departure time, the imputed departure time is set to the original departure time. User-added trips as well as long distance passenger mode trips are also set to the original departure time, as user-added trips are not subject to "late pickup" conditions, and long distance passenger modes are often plane trips where all collected traces contain speed information from other modes and thus are less reliable (as rMove cannot collect locations when a phone is in "airplane mode").

Imputed speed and duration are calculated based on the imputed departure time.

3.0 TRIP PURPOSE IMPUTATION

This section provides an overview of the process used to impute trip purposes in the TBI rMove survey data in cases where (a) the reported trip destination purpose was not consistent with the trip destination location (home, work, school or other), or (b) where the destination purpose was not reported for the trip. This section gives an overview of the process and provides tables summarizing the results.

3.1 OVERVIEW OF THE IMPUTATION PROCESS

The trip purpose imputation approach was applied to all rMove trips in person-days with at least 1 complete trip and no more than 10 incomplete trips. (“Incomplete” trips for which the respondent did not answer the trip-specific survey questions about purpose, mode, etc. for the given trip.)

The approach was to apply a logical sequence of “tests” or rules to trips for which the reported purpose is not consistent with the imputed location type based on the smartphone trip trace data. Appendix A describes the approach in more detail.

In general terms, the rules were designed to:

- Check the respondent’s reported destination purpose when it conflicts with the destination location type. (The details of the rules depend on the trip purpose, with different criteria used for change-mode trips, escort trips, linked transit trips, trips with home destinations but other reported purposes, etc.)
- Identify cases where respondents swapped the order of two or more trips when reporting their details.
- Identify cases where respondents may have omitted a trip and shifted remaining reported trip details by one trip when reporting the rest of their trips.
- Filling in missing data by sampling destination purposes from other trips made to the same locations, either by the same respondent or by other respondents.

Problematic trips were identified by comparing the destination purpose category to the destination location type. For example:

- For any destination purpose reported as “work” that was not at the primary work location, the destination purpose and purpose category were changed to “work-related”. (This was the original intention of offering different purpose categories for work (at the usual workplace) versus work-related.
- Similarly, a new purpose category of “school-related” was added to apply to any destination purpose reported as “school” that was not at the primary school location.

These recodings avoided some of the apparent mismatches between reported purposes and locations for work and school activities.

3.2 OVERVIEW OF THE IMPUTATION RESULTS

Table 3-1 below shows the frequencies and percentages of the various possible types of purpose/location mismatches before and after purpose imputation.

TABLE 3-1: LOCATION PURPOSE MISMATCH TYPE FOR TRIP DESTINATIONS

MISMATCH TYPE	FREQUENCY BEFORE IMPUTATION	PERCENT BEFORE IMPUTATION	FREQUENCY AFTER IMPUTATION	PERCENT AFTER IMPUTATION
Invalid day	3,832	1.1%	3,832	1.1%
No mismatch	263,839	79.0%	329,462	98.5%
Location = Home / Purpose = Not Home	10,792	3.2%	542	0.2%
Purpose = Home / Location = Not Home	11,257	3.4%	0	0.0%
Location = Work / Purpose = Not Work	2,651	0.8%	186	0.1%
Purpose = Work / Location = Not Work	3,727	1.1%	0	0.0%
Location = School / Purpose = Not School	878	0.3%	19	0.0%
Purpose = School / Location = Not School	614	0.2%	0	0.0%
Purpose = Missing	36,451	10.9%	0	0.0%
Total	**334,041**	**100.0%**	**334,041**	**100.0%**

As shown in Table 3-1, 9% of trips in the TBI dataset were identified as having a destination location/purpose type mismatch. The majority of trips (79%) have no destination location/purpose type mismatch and 11% have incomplete data for destination purpose meaning the participant did not complete the survey for that trip. For the 9% of trips with mismatches, the two most common types are trips where 1) the destination location is home but the reported purpose was not home (about 3%), and 2) the destination location was not home but the reported purpose was home (about 3%). About 1% of trips have a location at the usual workplace but the reported purpose was not work and 1% have a destination location not at the usual workplace but the reported purpose given was work. Overall, there are very few school-related mismatches.

After performing trip purpose imputation, 98.8% of trips have a destination purpose. As another check, not shown in the tables, before imputation, about 1.1% of rMove trips in the dataset have home as both the reported origin and destination purposes. The imputation resolves the majority of those cases decreasing the number of home to home trips from 3,708 to 446.

Table 3-2 is a cross-tabulation of the reported destination purpose category against the imputed destination purpose category. Most cases are on the diagonal, while the most common shifts are trips switching to or from a home purpose and trips shifting from the unknown/missing and incomplete categories to valid purposes.

Finally, Table 3-3 shows the specific imputation types that were applied versus the data mismatch type. The specific imputation types are described in more detail in Appendix A.

TABLE 3-2: REPORTED PURPOSE VERSUS IMPUTED PURPOSE FOR TRIP DESTINATIONS

D_PURPOSE_CATEGORY \ D_PURPOSE_CATEGORY_IMPUTED	Missing: non-imputable	Home	Work	Work-related	School	Escort	Shop	Meal	Social/ recreation	Errand/ other	Change mode	School-related	Spent the night at a non-home location	**Total**
Missing: non-response	3,594	990	255	743	11	359	695	582	1,514	911	185	49	0	9,888
Home	59	81,909	7	561	1	1,145	1,321	776	1,900	423	287	57	2,324	90,770
Work	17	81	18,967	829	0	175	135	139	240	0	48	19	60	20,710
Work-related	22	580	772	21,924	27	3	6	1	1	2	0	0	0	23,338
School	5	161	11	3	3,436	79	2	1	16	0	0	29	0	3,743
Escort	15	2,079	155	1	532	21,333	5	0	2	0	0	1	0	24,123
Shop	9	1,559	142	7	2	2	38,811	4	7	14	1	0	0	40,558
Meal	10	743	362	2	10	7	6	26,087	3	3	2	1	0	27,236
Social/ recreation	23	2,471	238	0	60	9	17	6	48,428	1	3	1	0	51,257
Errand/ other	75	1,207	297	3,245	145	3,599	7,978	4,208	8,390	3,066	836	437	0	33,483
Change mode	3	84	93	1	12	0	0	0	4	3	6,986	0	0	7,186
School-related	0	96	14	1	13	0	0	0	0	1	1	1,623	0	1,749
Total	3,832	91,960	21,313	27,317	4,249	26,711	48,976	31,804	60,505	4,424	8,349	2,217	2,384	334,041

TABLE 3-3: IMPUTATION TYPE VERSUS MISMATCH TYPE FOR VALID DAYS[2]

DESTINATION PURPOSE IMPUTATION CODE	LOCATION / PURPOSE MISMATCH TYPE								
Location:	No mismatch	Home	Work	School	Not Home	Not Work	Not School		**Total**
Purpose:		Not Home	Not Work	Not School	Home	Work	School	Missing	
1-OK-Use as is	263,491	0	0	0	0	0	0	0	**263,491**
2-Test 1-change mode ok	0	79	72	2	0	0	0	0	**153**
3-Test 2A-change to location type	0	2,781	1,299	176	0	0	0	0	**4,256**
4-Test 2B-for last trip of day-change to home	0	3,157	0	0	0	0	0	0	**3,157**
5-Test 2C-2A modified for missing-change to location type	0	1,303	433	43	0	0	0	0	**1,779**
6-Test 2D-2B modified for missing-change to home	0	993	0	0	0	0	0	0	**993**
7-Test 3A-prev trip at same location- longer dwell - keep as is	0	21	28	43	0	0	0	0	**92**
8-Test 3B-next trip at same location- longer dwell - keep as is	0	47	12	2	0	0	0	0	**61**
9-Test 4-prev and next trips are opposite mismatch - swap around	0	21	0	0	42	0	0	0	**63**
10-Test 5-prev trip is opposite mismatch - swap around	0	480	38	0	480	38	0	0	**1,036**
11-Test 6-next trip is opposite mismatch - swap around	0	109	8	1	109	8	1	0	**236**
12-Test 7-change to location type- all purposes in day shifted down 1	243	202	63	33	134	5	0	81	**761**
13-Test 8-last destination of day to Home loc with stay >3hr-change to home	105	623	0	0	0	0	0	0	**728**
14-Test 9-tests distance to usual location for purpose in case H/W/S have same loc	0	27	12	0	0	0	0	0	**39**
15-Test 10-change work-related to work if at work loc with no adj work purpose	0	0	173	0	0	0	0	0	**173**

[2] This table does not include trips with non-imputable trip purposes or trips on invalid travel days.

DESTINATION PURPOSE IMPUTATION CODE

16-Test 11-change to location type if it doesn't create consecutive trip w same purpose	0	407	327	559	0	0	0	0	**1,293**
19-No test passed - examine further	0	542	186	19	0	0	0	0	**747**
20-Imputed Escort purpose	0	0	0	0	530	19	44	0	**593**
21-Test 2A-within 200 m of usual location- no adj conflicts	0	0	0	0	1,294	2,082	484	0	**3,860**
22-Test 2B-within 300 m of usual location- no adj conflicts	0	0	0	0	652	0	0	0	**652**
23-Test 2C-within 500 m of usual location- no adj conflicts	0	0	0	0	614	0	0	0	**614**
24-Test 3-Is overnight stay away from home- out of region	0	0	0	0	778	10	0	0	**788**
25-Test 3-Is overnight stay away from home- within region	0	0	0	0	1,546	50	0	0	**1,596**
31-Imputed based on same persons stops within 50 m of same location	0	0	0	0	1,531	303	16	5,274	**7,124**
32-Imputed based on same persons stops within 100 m of same location	0	0	0	0	237	102	2	1,517	**1,858**
33-Imputed based on same persons stops within 200 m of same location	0	0	0	0	339	95	4	2,269	**2,707**
34-Imputed based on other persons stops within 50 m of same location	0	0	0	0	2,160	767	57	18,172	**21,156**
35-Imputed based on other persons stops within 100 m of same location	0	0	0	0	234	94	2	3,239	**3,569**
36-Imputed based on other persons stops within 200 m of same location	0	0	0	0	187	70	3	2,283	**2,543**
37-use Test 5-Work purpose not at primary work loc changed to work-related	0	0	0	0	0	84	0	0	**84**
38-Reported purpose is school, changed to school related	0	0	0	0	0	0	1	0	**1**
39-No other visits to location - impute other (no evidence)	0	0	0	0	390	0	0	3,616	**4,006**
Total	**263,839**	**10,792**	**2,651**	**878**	**11,257**	**3,727**	**614**	**36,451**	**330,209**

4.0 TRIP MODE IMPUTATION

4.1 INTRODUCTION

RSG developed a mode imputation model for the TBI study geography to be used as a validation of the trip survey reported mode. The sole purpose of this model is to be applied to person-days with all complete trips (no trips with unanswered surveys) and used to identify reported modes that are unlikely given the reported information. However, a secondary benefit is that the model can also be used to impute mode for trips where mode was not reported in the trip survey as discussed below.

This work involved two steps (model estimation and model application) which are discussed in more detail below.

4.2 STEP 1: MODEL ESTIMATION

In this step, trip data as well as household- and person-level data were used to estimate the parameters of a multinomial choice model. To be included in model estimation, the reported trip must occur on a complete person-day (all trip survey data was provided and not missing). Further filters were applied to remove trips that end or begin outside of the study region and trips directly added by the user (not captured by rMove).

The multinomial choice model included the following 4 alternatives:

- Walk trips
- Bike trips[3]
- Car trips
- Transit trips

These 4 alternatives were chosen based on their prevalence in the trip dataset and the ability for the model to differentiation them from other modes. Further disaggregation of the mode alternatives is not recommended (e.g., differentiating between household vehicle and friend/relative/colleague's car). All other modes were excluded from model estimation. For the purposes of model normalization, the car mode alternative served as the base alternative[4].

Independent variables in the model are listed in Table 4-1. These independent variables were selected after an extensive model-specification exercise in other geographies and were included because of their statistical significance in predicting the reported mode.

---

[3] Bike-share, scooter-share, and moped share were excluded from this alternative because of the unique nature of these modes.
[4] For categorical parameters, the estimates should be interpreted relative to this base alternative.

TABLE 4-1: INDEPENDENT VARIABLES IN IMPUTATION MODEL

TRIP VARIABLES		HOUSEHOLD VARIABLES	PERSON VARIABLES
Average speed and % of trip at speed bins	Trip purpose	Number of household vehicles	Disability
Distance	Percent of all trips made by mode	Household income	Frequency of transit usage
Day of week	Unlinking flags		Age
Hour of arrival	Loop trip		
Mode used on previous and next trip	Distance from origin/destination to nearest transit access point		
Destination distance to home / work	# of GPS points collected		

The estimated model coefficients and model fits can be found in Table 4-4 in Section 4.4.

4.3 STEP 2: MODEL APPLICATION

In this step, the estimated model was applied back to the data to generate a model-based probability of seeing the observed mode for the trip. The model was applied to all trips for person-days that have at least one complete trip and no more than 10 incomplete trips, with the following exceptions:

- trips where the reported mode does not correspond to one of the four modes used in the model;
- trips that start or end outside the region;
- trips that were user-added (not captured by rMove).

By applying it to a broader set of person-days, the model can be used to impute mode for many trips with incomplete survey records from the set of 4 mode alternatives modeled.

For the trips included in the model application, the following additional variables were included in the trip file:

- **mode_type_predicted:** The mode with the highest probability from the model.
- **predicted_mode_probability:** The model probability for mode_type_predicted.

- **mode_type_p_1, mode_type_p_2, mode_type_p_3, mode_type_p_4:** The modeled probabilities for each of the 4 alternatives in the model.

These variables were comprehensively included in dataset to allow the freedom to subsequently further review or edit the dataset. For example, Table 4-2 provides the predicted mode (the mode with the highest probability) by reported mode type while

Table 4-3 provides the mean predicted probabilities for each of the 4 mode alternatives for each reported mode type. These tables highlight the use case for a model like this. First off, there can be strong correlation between the predicted modes and the reported modes (e.g., the model predicts “walking” for 29,806 reported walk trips with a mean probability of 0.81) which serves as data validation. In cases where there is not strong validation, the model can provide guidance on what the likely mode could be. The data user can then make some assumptions based on the model’s probability to accept or reject the reported mode (e.g., reject the reported mode if the probability is below 0.02 and accept the predicted mode if is above 0.95). Finally, the model can serve as an imputation technique for trips were mode was not reported (the last row in Table 4-2 and Table 4-3).

TABLE 4-2: MODEL PREDICTIONS BY REPORTED MODE TYPE

REPORTED MODE TYPE	PREDICTED MODE			
	WALK	BIKE	VEHICLE	TRANSIT
Rail	230	40	657	1,363
School bus	159	3	1,644	19
Public bus	450	23	663	4,594
Other bus	193	33	1,106	180
Long distance passenger mode	5	0	368	326
Smartphone ride-hailing service	52	9	1,055	104
For-hire vehicle	18	2	207	8
Household vehicle	1,599	294	245,173	571
Other vehicle	229	29	18,756	134
Micromobility	297	3,769	500	87
Other	597	62	2,606	50
Walk	29,806	333	5,010	343
Missing: Non-response	1,735	151	8,006	237

TABLE 4-3: AVERAGE MODEL PROBABILITIES BY REPORTED MODE TYPE

REPORTED MODE TYPE	MEAN PROBABILITY			
	WALK	BIKE	VEHICLE	TRANSIT
Rail	0.11	0.04	0.32	0.54
School bus	0.13	0.07	0.77	0.03
Public bus	0.09	0.01	0.13	0.77
Other bus	0.13	0.03	0.7	0.13
Long distance passenger mode	0.01	0.00	0.52	0.47
Smartphone ride-hailing service	0.05	0.02	0.82	0.11
For-hire vehicle	0.07	0.01	0.87	0.04
Household vehicle	0.01	0.00	0.98	0.01
Other vehicle	0.02	0.00	0.97	0.01
Micromobility	0.07	0.78	0.12	0.03
Other	0.18	0.03	0.75	0.03
Walk	0.81	0.01	0.16	0.02
Missing: Non-response	0.17	0.02	0.77	0.04

4.4 MODEL COEFFICIENTS

TABLE 4-4: MODEL COEFFICIENTS FOR MODE IMPUTATION

COEFFICIENT	DESCRIPTION	ALTERNATIVE	ESTIMATE	STD ERROR	T-TEST
asc_1	alternative specific constants	walk	5.232	0.36	14.4
asc_2	alternative specific constants	bike	0.125	0.82	0.2
asc_3	alternative specific constants	car	0.000		
asc_4	alternative specific constants	transit	-1.030	0.45	-2.3
b_speed_1	linear speed in mph (all trips)	walk	-0.031	0.01	-4.0
b_speed_2	linear speed in mph (all trips)	bike	-0.100	0.02	-4.0
b_speed_3	linear speed in mph (all trips)	car	0.000		
b_speed_4	linear speed in mph (all trips)	transit	-0.036	0.01	-5.0
b_speed_log_1	logarithm of speed in mph (all trips)	walk	-0.444	0.11	-4.0
b_speed_log_2	logarithm of speed in mph (all trips)	bike	1.500	0.25	5.9
b_speed_log_3	logarithm of speed in mph (all trips)	car	0.000		
b_speed_log_4	logarithm of speed in mph (all trips)	transit	-0.657	0.11	-5.9
b_speed_05_1	speed > 5 mph (all trips)	walk	-0.455	0.10	-4.7
b_speed_15_2	speed > 15 mph (all trips)	bike	-0.489	0.19	-2.6
b_speed_3points_1	linear speed in mph (< 4 gps points)	walk	0.111	0.01	11.8
b_speed_3points_2	linear speed in mph (< 4 gps points)	bike	0.100	0.03	3.0
b_speed_3points_3	linear speed in mph (< 4 gps points)	car	0.000		
b_speed_3points_4	linear speed in mph (< 4 gps points)	transit	0.039	0.01	3.7
b_speed_log_3points_1	log of speed in mph (< 4 gps points)	walk	-1.520	0.13	-11.3
b_speed_log_3points_2	log of speed in mph (< 4 gps points)	bike	-2.638	0.32	-8.4
b_speed_log_3points_3	log of speed in mph (< 4 gps points)	car	0.000		
b_speed_log_3points_4	log of speed in mph (< 4 gps points)	transit	0.006	0.16	0.0
b_speed_05_3points_1	speed > 5 mph (< 4 gps points)	walk	0.054	0.16	0.3
b_speed_15_3points_2	speed > 15 mph (< 4 gps points)	bike	0.686	0.61	1.1
b_perc_speed_0_2_1	% of trip spent between 0 - 2 mph	walk	-0.141	0.07	-2.1
b_perc_speed_0_2_2	% of trip spent between 0 - 2 mph	bike	-0.668	0.22	-3.1
b_perc_speed_0_2_3	% of trip spent between 0 - 2 mph	car	0.000		
b_perc_speed_0_2_4	% of trip spent between 0 - 2 mph	transit	-0.731	0.18	-4.2
b_perc_speed_2_4_1	% of trip spent between 2 - 4 mph	walk	1.674	0.08	20.7
b_perc_speed_2_4_2	% of trip spent between 2 - 4 mph	bike	-0.601	0.28	-2.2
b_perc_speed_2_4_3	% of trip spent between 2 - 4 mph	car	0.000		
b_perc_speed_2_4_4	% of trip spent between 2 - 4 mph	transit	1.143	0.18	6.5
b_perc_speed_4_15_1	% of trip spent between 4 - 15 mph	walk	-0.983	0.10	-10.3
b_perc_speed_4_15_2	% of trip spent between 4 - 15 mph	bike	1.063	0.24	4.5
b_perc_speed_4_15_3	% of trip spent between 4 - 15 mph	car	0.000		
b_perc_speed_4_15_4	% of trip spent between 4 - 15 mph	transit	0.782	0.17	4.6
b_perc_speed_15_40_1	% of trip spent between 15 - 40 mph	walk	-3.184	0.13	-24.6
b_perc_speed_15_40_2	% of trip spent between 15 - 40 mph	bike	-4.383	0.32	-13.9
b_perc_speed_15_40_3	% of trip spent between 15 - 40 mph	car	0.000		
b_perc_speed_15_40_4	% of trip spent between 15 - 40 mph	transit	-0.216	0.18	-1.2
b_perc_speed_40_plus_1	% of trip spent greater than 40 mph	walk	-1.579	0.19	-8.4
b_perc_speed_40_plus_2	% of trip spent greater than 40 mph	bike	-7.412	0.60	-12.3
b_perc_speed_40_plus_3	% of trip spent greater than 40 mph	car	0.000		
b_perc_speed_40_plus_4	% of trip spent greater than 40 mph	transit	-1.511	0.22	-6.8
b_has_vehicle_1	hh has 1 car	car	0.639	0.08	8.3
b_has_vehicle_2	hh has 2 cars	car	0.750	0.08	9.2
b_has_vehicle_3	hh has 3+ cars	car	0.783	0.09	8.7
b_distance_1	linear distance in miles	walk	-1.307	0.03	-46.0
b_distance_2	linear distance in miles	bike	-0.008	0.02	-0.4
b_distance_3	linear distance in miles	car	0.000		
b_distance_4	linear distance in miles	transit	-0.071	0.01	-7.4
b_distance_log_1	log of distance in miles	walk	0.000		
b_distance_log_2	log of distance in miles	bike	0.571	0.13	4.3
b_distance_log_3	log of distance in miles	car	0.000		
b_distance_log_4	log of distance in miles	transit	1.838	0.08	22.1
b_has_disability_1	has disability	walk	-0.150	0.13	-1.1
b_has_disability_pnta_1	disability prefer not to answer	walk	-0.085	0.22	-0.4

COEFFICIENT	DESCRIPTION	ALTERNATIVE	ESTIMATE	STD ERROR	T-TEST
b_has_disability_missing_1	disability missing	walk	0.077	0.27	0.3
b_has_disability_2	has disability	bike	-0.787	0.39	-2.0
b_has_disability_pnta_2	disability prefer not to answer	bike	-0.697	0.60	-1.2
b_has_disability_missing_2	disability missing	bike	-0.842	0.98	-0.9
b_has_disability_3	has disability	car	0.000		
b_has_disability_pnta_3	disability prefer not to answer	car	0.000		
b_has_disability_missing_3	disability missing	car	0.000		
b_has_disability_4	has disability	transit	0.316	0.13	2.4
b_has_disability_pnta_4	disability prefer not to answer	transit	-0.535	0.27	-2.0
b_has_disability_missing_4	disability missing	transit	-1.113	0.38	-2.9
b_weekday_1	Trip occurred on Mon thru Thurs	walk	0.041	0.04	1.1
b_weekday_2	Trip occurred on Mon thru Thurs	bike	0.076	0.08	0.9
b_weekday_3	Trip occurred on Mon thru Thurs	car	0.000		
b_weekday_4	Trip occurred on Mon thru Thurs	transit	0.294	0.05	5.7
b_age67_1	Age of 35-55	walk	0.082	0.04	2.0
b_age67_2	Age of 35-55	bike	-0.164	0.09	-1.9
b_age67_3	Age of 35-55	car	0.000		
b_age67_4	Age of 35-55	transit	-0.069	0.05	-1.3
b_age8plus_1	Age of 55+	walk	-0.260	0.05	-5.6
b_age8plus_2	Age of 55+	bike	-0.348	0.11	-3.2
b_age8plus_3	Age of 55+	car	0.000		
b_age8plus_4	Age of 55+	transit	-0.437	0.07	-6.4
b_arrive_before_6_1	arrives before 6am	walk	-0.337	0.10	-3.4
b_arrive_7thru9_1	arrives between 7 and 9am	walk	-0.289	0.05	-5.4
b_arrive_16thru19_1	arrives between 4 and 7pm	walk	-0.073	0.04	-1.7
b_arrive_after_20_1	arrives after 8pm	walk	-0.301	0.07	-4.1
b_arrive_before_6_2	arrives before 6am	bike	0.034	0.21	0.2
b_arrive_7thru9_2	arrives between 7 and 9am	bike	0.336	0.12	2.8
b_arrive_16thru19_2	arrives between 4 and 7pm	bike	0.405	0.10	4.1
b_arrive_after_20_2	arrives after 8pm	bike	0.158	0.16	1.0
b_arrive_7thru9_3	arrives before 6am	car	0.000		
b_arrive_16thru19_3	arrives between 7 and 9am	car	0.000		
b_arrive_after_20_3	arrives between 4 and 7pm	car	0.000		
b_arrive_before_6_5	arrives after 8pm	car	0.000		
b_arrive_before_6_4	arrives before 6am	transit	-0.783	0.12	-6.5
b_arrive_7thru9_4	arrives between 7 and 9am	transit	-0.505	0.07	-6.9
b_arrive_16thru19_4	arrives between 4 and 7pm	transit	0.137	0.06	2.2
b_arrive_after_20_4	arrives after 8pm	transit	-0.215	0.10	-2.2
b_income_1	linear income in $10k	walk	0.005	0.00	2.0
b_income_2	linear income in $10k	bike	0.035	0.01	6.3
b_income_3	linear income in $10k	car	0.000		
b_income_4	linear income in $10k	transit	0.008	0.00	2.5
b_log_income_1	log of income in $10k	walk	-0.186	0.12	-1.5
b_log_income_2	log of income in $10k	bike	-1.516	0.26	-5.8
b_log_income_3	log of income in $10k	car	0.000		
b_log_income_4	log of income in $10k	transit	-0.423	0.14	-3.0
b_transit_weekly_4plus	Uses 4+ day/week	transit	2.704	0.08	35.6
b_transit_weekly_1to3	Uses 1 to 3 days /week	transit	2.395	0.09	27.9
b_transit_monthly	Uses monthly	transit	0.736	0.09	8.1
b_prev_inertia_1	Previous trip is same mode	walk	1.806	0.05	33.5
b_prev_inertia_2	Previous trip is same mode	bike	5.054	0.12	42.0
b_prev_inertia_3	Previous trip is same mode	car	2.305	0.04	55.8
b_prev_inertia_4	Previous trip is same mode	transit	1.291	0.08	17.1
b_prev_inertia_home_1	Previous trip is same mode, stop at home in between	walk	-1.789	0.13	-14.3
b_prev_inertia_home_2	Previous trip is same mode, stop at home in between	bike	-4.140	0.31	-13.3
b_prev_inertia_home_3	Previous trip is same mode, stop at home in between	car	-1.404	0.06	-21.8
b_prev_inertia_home_4	Previous trip is same mode, stop at home in between	transit	-1.512	0.24	-6.3

COEFFICIENT	DESCRIPTION	ALTERNATIVE	ESTIMATE	STD ERROR	T-TEST
b_prev_inertia_change_mode_1	Previous trip is same mode, change mode purpose in between	walk	-3.391	0.15	-22.7
b_prev_inertia_change_mode_2	Previous trip is same mode, change mode purpose in between	bike	-4.403	0.59	-7.5
b_prev_inertia_change_mode_3	Previous trip is same mode, change mode purpose in between	car	-2.344	0.12	-19.7
b_prev_inertia_change_mode_4	Previous trip is same mode, change mode purpose in between	transit	-1.006	0.11	-9.1
b_next_inertia_1	Next trip is same mode	walk	1.083	0.05	21.1
b_next_inertia_2	Next trip is same mode	bike	5.130	0.13	40.9
b_next_inertia_3	Next trip is same mode	car	2.734	0.04	63.8
b_next_inertia_4	Next trip is same mode	transit	1.411	0.08	17.5
b_next_inertia_home_1	Next trip is same mode, stop at home in between	walk	-0.767	0.13	-5.9
b_next_inertia_home_2	Next trip is same mode, stop at home in between	bike	-4.166	0.34	-12.3
b_next_inertia_home_3	Next trip is same mode, stop at home in between	car	-2.437	0.07	-35.6
b_next_inertia_home_4	Next trip is same mode, stop at home in between	transit	0.072	0.32	0.2
b_next_inertia_change_mode_1	Next trip is same mode, change mode purpose in between	walk	-2.522	0.15	-16.4
b_next_inertia_change_mode_2	Next trip is same mode, change mode purpose in between	bike	-2.119	0.65	-3.2
b_next_inertia_change_mode_3	Next trip is same mode, change mode purpose in between	car	-2.497	0.12	-21.3
b_next_inertia_change_mode_4	Next trip is same mode, change mode purpose in between	transit	-1.173	0.10	-11.8
b_next_inertia_change_mode_5	Next trip is same mode, change mode purpose in between	tnc	0.000		
b_close_to_home_1_1	Trip destination is < 1 mile of home	walk	-1.833	0.05	-34.2
b_close_to_home_2_1	Trip destination is < 5 miles of home	walk	-0.765	0.05	-14.9
b_close_to_home_1_2	Trip destination is < 1 mile of home	bike	-0.217	0.14	-1.6
b_close_to_home_2_2	Trip destination is < 5 miles of home	bike	-0.379	0.14	-2.7
b_close_to_home_1_3	Trip destination is < 1 mile of home	car	0.000		
b_close_to_home_2_3	Trip destination is < 5 miles of home	car	0.000		
b_close_to_home_1_4	Trip destination is < 1 mile of home	transit	-1.117	0.07	-15.2
b_close_to_home_2_4	Trip destination is < 5 miles of home	transit	-0.253	0.07	-3.8
b_close_to_work_1_1	Trip destination is < 1 mile of work	walk	0.487	0.05	10.1
b_close_to_work_2_1	Trip destination is < 5 miles of work	walk	0.166	0.05	3.4
b_close_to_work_1_2	Trip destination is < 1 mile of work	bike	0.636	0.11	5.7
b_close_to_work_2_2	Trip destination is < 5 miles of work	bike	0.108	0.10	1.1
b_close_to_work_1_3	Trip destination is < 1 mile of work	car	0.000		
b_close_to_work_2_3	Trip destination is < 5 miles of work	car	0.000		
b_close_to_work_1_4	Trip destination is < 1 mile of work	transit	0.584	0.07	8.7
b_close_to_work_2_4	Trip destination is < 5 miles of work	transit	0.110	0.06	1.8
b_perc_trips_1	Mode represents less than 25% of trips made by participants	All	0.000		
b_perc_trips_2	Mode represents between 25% and 50% of trips made by participants	All	1.099	0.03	33.1
b_perc_trips_3	Mode represents between 50% and 75% of trips made by participants	All	1.525	0.06	27.1
b_perc_trips_4	Mode represents greater than 75% of trips made by participants	All	2.918	0.17	16.7
b_shopping_3	Destination purpose is shopping	car	0.144	0.17	0.9
b_exercise_1	Destination purpose is exercise	walk	2.206	0.06	39.6
b_exercise_2	Destination purpose is exercise	bike	1.266	0.13	9.9
b_unlinked_1	Unlinked transit trip flag set to 1	walk	5.187	0.25	20.9
b_unlinked_4	Unlinked transit trip flag set to 1	transit	6.713	0.24	28.2
b_unlinked_o_home_1	Split unlinked transit trip flag set to 1 and origin is at home	walk	-0.410	0.23	-1.8

COEFFICIENT	DESCRIPTION	ALTERNATIVE	ESTIMATE	STD ERROR	T-TEST
b_unlinked_d_home_1	Split unlinked transit trip flag set to 1 and destination is at home	walk	0.964	0.18	5.5
b_split_loop_1	Split loop trip flag set to 1	walk	0.919	0.06	14.2
b_split_loop_2	Split loop trip flag set to 1	bike	-0.308	0.14	-2.3
b_o_escort_1	Origin purpose is escort	walk	-0.595	0.09	-6.3
b_o_escort_2	Origin purpose is escort	bike	-1.336	0.26	-5.1
b_o_escort_3	Origin purpose is escort	car	0.000		
b_o_escort_4	Origin purpose is escort	transit	-0.391	0.15	-2.6
b_d_escort_1	Destination purpose is escort	walk	-0.360	0.09	-3.9
b_d_escort_2	Destination purpose is escort	bike	-0.368	0.22	-1.6
b_d_escort_3	Destination purpose is escort	car	0.000		
b_d_escort_4	Destination purpose is escort	transit	-0.609	0.15	-4.0
b_has_3_points_or_fewer_1	Trip has 3 or fewer GPS points	walk	0.665	0.14	4.6
b_has_3_points_or_fewer_2	Trip has 3 or fewer GPS points	bike	2.611	0.42	6.2
b_has_3_points_or_fewer_3	Trip has 3 or fewer GPS points	car	0.000		
b_has_3_points_or_fewer_4	Trip has 3 or fewer GPS points	transit	-0.811	0.24	-3.3
b_o_got_gas_3	Origin purpose is "get gas"	car	1.279	0.20	6.3
b_d_got_gas_3	Destination purpose is "get gas"	car	1.356	0.22	6.1
b_o_distance_to_transit_1	Distance from trip origin to nearest transit access point	walk	-0.018	0.02	-1.0
b_o_distance_to_transit_2	Distance from trip origin to nearest transit access point	bike	-0.055	0.03	-1.7
b_o_distance_to_transit_3	Distance from trip origin to nearest transit access point	car	0.000		
b_o_distance_to_transit_4	Distance from trip origin to nearest transit access point	transit	-0.112	0.01	-7.6
b_d_distance_to_transit_1	Distance from trip destination to nearest transit access point	walk	-0.011	0.02	-0.6
b_d_distance_to_transit_2	Distance from trip destination to nearest transit access point	bike	0.027	0.03	0.9
b_d_distance_to_transit_3	Distance from trip destination to nearest transit access point	car	0.000		
b_d_distance_to_transit_4	Distance from trip destination to nearest transit access point	transit	0.070	0.01	5.5

McFadden's Rho-square: 0.941

5.0 WEIGHTING METHODOLOGY

5.1 INTRODUCTION AND SCOPE

This section describes the analysis, recommendations, and methodology used to expand[5] the data collected in the 2018-2019 TBI to the 2018 American Community Survey Public Use Microdata Sample (ACS PUMS) data. The weighting methodology applied adjusts for survey non-response, survey participation mode, and geographic bias due to oversampling and other factors. In addition, RSG adjusted trip rates between the three participation methods offered for the survey: 1) online, 2) call center, and 3) smartphone app.

The applied weighting process included four primary steps:

1. **Initial Expansion**: Calculating an “initial weight” based on the probability of selection in the sample design. This essentially “reverses” the sample plan, providing higher initial weights to areas where less sampling occurred.
2. **Reweighting to account for non-response bias**: Performing an iterative proportional fit (IPF) routine to several key household and person dimensions to ensure the weighted data accurately represent the entire survey region (and reduce sampling biases).
    a. To do this step, a few missing data elements (income, gender, and race) need to be imputed for those who did not provide that information.
3. **Creating day-level weights to account for multi-day survey data:** Adjusting the day-level and trip-level data to account for the fact that smartphone respondents provided multi-day travel diaries, while online and call center respondents provided a single-day travel diary (this is the “multi-day adjustment”). These relatively simple adjustments ensure that travel analyses accurately reflect the entire survey region and do not over-represent smartphone respondents with multiple travel days.
4. **Adjusting for non-response bias in day-pattern and trip rates**: Adjusting the trip-level weights by data collection method (smartphone, online, call center) to account for reporting biases that RSG has detected in this survey and prior travel surveys. These adjustments help make the day and trip-level data more consistent and increase the accuracy of trip rates across survey participation methods.

The following sections describe this process and the results in detail. The overall goal is to make the survey sample representative of the entire survey area across several key dimensions related to travel behavior.

[5] For the purposes of this report, the terms expansion, expansion factors, and weights are used interchangeably and are synonymous. They all represent the concept of an expansion weight.

5.2 INITIAL EXPANSION FACTORS

The purpose of the initial expansion is to expand each complete survey record to the population that was eligible to participate in the survey. The initial expansion weights are based on the relative probabilities of each respondent has of being in the sample, as a function of the sampling plan and the number of invitations sent to specific sampling segments.

Selection of Respondents for Weighting

After the data processing is complete and any invalid person-days and household-days have been flagged as incomplete, any household which has at least one complete and valid weekday travel day will be included in the weighting. For this purpose, a complete weekday is any complete Monday, Tuesday, Wednesday, or Thursday. The selection of “weekdays” essentially assumes that trip rates and behavior on those days are similar enough to consider them interchangeable, with an average weekday being the average of travel across those days. Only those weekdays will be given person-day weights for analysis.

RSG did not weight travel data for Friday, Saturday, or Sunday because (a) data was only collected from smartphone-participating households on those days, (b) the travel behavior for those days is not assumed to be interchangeable with the behavior for Monday-Thursday, and (c) the data is used primarily to analyze and model typical weekday travel.

Calculation of Initial Expansion Weights

To begin expanding the complete households, separate initial weights are calculated for each sampling segment. To calculate the initial expansion factors for each stratum, the ratio of population household counts to sampled households is calculated.

The initial expansion weights are used as the starting weights for further re-weighting to correct for non-response biases in the data, which is described in the following section. Table 5-1 provides a summary of the initial expansion factors by sample segment.

TABLE 5-1: INITIAL EXPANSION FACTORS

SAMPLE SEGMENT	SAMPLED HOUSEHOLDS	ACS HOUSEHOLDS	INITIAL EXPANSION FACTOR
Core-Rural	557	114,240	205.10
Core-Urban	4,530	879,673	194.19
Rural Ring	1,215	239,753	197.33
Hard-to-Survey	1,446	198,713	137.42
Total	**7,748**	**1,432,379**	**-**

5.3 REWEIGHTING TO ACCOUNT FOR NONRESPONSE BIAS

The 2018 American Community Survey Public Use Microdata Sample (ACS PUMS) data served as the target data for weighting this dataset. An iterative proportional fit algorithm (IPF) was used to adjust the initial weights so that the sum of the weights matched various household-level and person-level marginal targets within each of the defined weighting geographies. The IPF routine was seeded with the initial expansion weights. Then, the algorithm was completed in a way to minimize deviation from the initial weights while matching the control targets as closely as possible.

Weighting Geography

Using ACS PUMS data, separate sets of weighting controls were generated for each of the PUMAs in study areas. (PUMAs have populations in the range of 100,000–200,000). As survey sample sizes are often too sparse to weight to all distributions within each PUMA, another option is to use county-level targets instead (unless the counties are smaller than PUMAs, in which case PUMAs would be used). However, there can be a wide variation in the level of urbanization across PUMAs within a county.

This weighting process used the following geographies:

- **Anoka County, MN** (PUMAs: 01101, 01102, 01103)
- **Washington County, MN** (PUMAs: 01201, 01202)
- **Ramsey County, MN – St. Paul (**PUMAs: 01303, 01304)
- **Ramsey County, MN – Other** (PUMAs: 01301, 01302)
- **Hennepin County, MN – Minneapolis** (PUMAs: 01405, 01406, 01407)
- **Hennepin, County MN – Other** (PUMAs: 01401, 01402, 01403, 01404, 01408, 01409, 01410)
- **Dakota County, MN** (PUMAs: 01501, 01502, 01503)
- **Scott & Carver Counties, MN** (PUMAs: 01600, 01700)
- **Chisago & Isanti Counties, MN** (PUMAs: 00600)
- **Sherburne, Wright, McLeod, & Sibley Counties, MN** (PUMAs: 01000, 01800, 01900)
- **Goodhue, Rice, & Le Sueur Counties, MN** (PUMAs: 02300)
- **Pierce, Polk, & St. Croix Counties, WI** (PUMAs: 00700, 55101, 55102)

Household and Person Weighting Targets

There are a variety of person-level and household-level target categories. The person-level targets are designed to identify the person types that are typically used in activity-based modeling software such as CT-RAMP and DaySim. The weighting targets were derived from PUMS data using the person-level weights. PUMS allows definition of full-time vs. part-time workers in a way consistent with the survey, while ACS tables do not provide consistent information. (For example, in the ACS tables, "part-time" includes people who only worked part of the previous year.) The PUMA geography identified in the PUMS data is sufficient for setting weighting targets, even using the latest PUMS (2018). Table 5-2 and Table 5-3 provide the household and person-level variables used in the iterative proportional fit exercise.

TABLE 5-2: HOUSEHOLD-LEVEL TARGET VARIABLES

VARIABLE	CATEGORIES
Household Size	1-person 2-person 3-person 4-person 5-person or more
Income ***(Imputed if non-response)***	Under $25,000 $25,000 - $49,999 $50,000 - $74,999 $75,000 - $99,999 $100,000 - $149,999 $150,000 or more
Workers	0 workers 1 worker 2 workers 3 workers or more
Vehicles	0 vehicles or 1 vehicle 2 vehicles 3 vehicles or more
Age of Head of Household	Under 35 years 35 – 64 years 65 years or older
Presence of Kids	0 kids 1 or more kids
Total Households	-

TABLE 5-3: PERSON-LEVEL TARGET VARIABLES

VARIABLE	CATEGORIES
Gender *(Imputed if non-response)*	Male Female
Age	Under 5 years 5 – 15 years 16 – 17 years 18 – 24 years 25 – 44 years 45 – 64 years 65 years or older
Worker Status	Worker Non-worker
University Student Status	University student Non-university student
Race *(Imputed if non-response)*	White Black or African American Asian Other Two or more races
Typical Commute Mode	Walk or bicycle Car (drive alone) Carpool Other No commute
Total Persons	-

5.4 IMPUTATION OF MISSING VALUES

The income, gender, and race questions in the survey allowed participants to respond with "prefer not to answer." To facilitate data weighting missing values were imputed for these variables.

Income

Income was imputed using a model-based approach where missing income was predicted based on a set of independent variables including:

- Income distribution of the block group
- Number of working adults in the household
- Educational attainment of the household
- Number of children in the household
- Age of the primary survey respondent
- Homeownership
- Single-family home residence type

This model has been tested across many travel survey projects and adequately matches the income values that were reported, indicating it is reliable to predict the missing income values.

An assignment of imputed income was made based on the predicted probabilities generated by the imputation model. Model specification and coefficients are shown in Table 5-4.

TABLE 5-4: INCOME IMPUTATION MODEL SUMMARY

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
I (finc_0k_25k + finc_25k_50k + finc_50k_75k)	Fraction of people in block group with incomes 0k-75k	-0.353	0.363	-0.974
finc_100k_150k	Fraction of people in block group with incomes 100k-150k	0.754	0.477	1.581
finc_150k_plus	Fraction of people in block group with incomes more than 150k	2.848	0.379	7.512
nonworking_adult_n	Number of nonworking adults in household	0.748	0.049	15.398
child_n	Number of children in household	-0.010	0.027	-0.376
full_time_graduate_degree_n	Number of full-time workers with graduate degrees in household	2.612	0.066	39.303
part_time_graduate_degree_n	Number of part-time workers with graduate degrees in household	1.318	0.131	10.033
full_time_bachelor_degree_n	Number of full-time workers with bachelor's degrees in household	2.253	0.060	37.336
part_time_bachelor_degree_n	Number of part-time workers with bachelor's degrees in household	0.975	0.100	9.734
full_time_low_education_n	Number of full-time workers with no advanced degrees in household	1.490	0.056	26.692
part_time_low_education_n	Number of part time workers with no advanced degrees in household	0.246	0.079	3.114
head_under_35_n	Head of household under 35 years	-0.390	0.062	-6.309

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
head_over_65_n	Head of household over 65 years	0.044	0.069	0.646
own_home	Household owns home	0.937	0.070	13.460
single_family_home	Household lives in single family home	0.327	0.061	5.343

McFadden's rho-squared: 0.192

Gender

Missing gender was probabilistically assigned based on the sample data's gender distribution within the respondent's age category.

Race

Race was also imputed using a model-based approach. Table 5-5 shows the race distribution in the TBI sample before and after imputation alongside the ACS distribution.

In this case, two models were used depending on what was known about the respondent's household. In households where race was known for at least one adult, race was imputed using a model that depended on the percentage of the household with that category (Table 5-6). For households where race was not known for any adults, a more general model was used, which is described in Table 5-7.

TABLE 5-5: RACE DISTRIBUTION

RACE RECODE FOR IMPUTATION	SAMPLE COUNT	PERCENT	VALID PERCENT	ACS PERCENT	PERCENT WHERE IMPUTED	PERCENT FOR WEIGHTING
White	13,258	83.2%	89.9%	80.2%	87.2%	89.7%
Black or African American	373	2.3%	2.5%	7.8%	3.3%	2.6%
Asian	479	3.0%	3.2%	6.2%	3.6%	3.3%
Other	241	1.5%	1.6%	2.7%	3.0%	1.7%
Two or more races	397	2.5%	2.7%	3.2%	2.9%	2.7%
Missing	1,195	7.5%	-	-	-	-

TABLE 5-6: RACE IMPUTATION MODEL FOR HOUSEHOLDS WITH AT LEAST ONE ADULT REPORTING RACE

ALTERNATIVE[6]	PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
Black or African American	(Intercept)		-2.925	0.904	-3.234
	perc_race_white	% of HH who is white	-5.153	1.034	-4.981
	perc_race_afam	% of HH who is Black or African American	11.001	1.050	10.473
	perc_race_asian	% of HH who is Asian	-0.703	2.076	-0.339
	perc_race_other	% of HH who is other	2.141	1.931	1.109
	perc_race_multiple	% of HH who is two or more races	2.293	1.115	2.056
Asian	(Intercept)		-2.630	0.744	-3.536
	perc_race_white	% of HH who is white	-5.034	0.849	-5.933
	perc_race_afam	% of HH who is Black or African American	-0.698	2.049	-0.341
	perc_race_asian	% of HH who is Asian	10.625	0.871	12.202
	perc_race_other	% of HH who is other	1.550	1.646	0.942
	perc_race_multiple	% of HH who is two or more races	1.330	0.992	1.341
Other	(Intercept)		-1.941	0.678	-2.862
	perc_race_white	% of HH who is white	-6.187	0.845	-7.317
	perc_race_afam	% of HH who is Black or African American	0.783	1.816	0.431
	perc_race_asian	% of HH who is Asian	-2.188	2.126	-1.029
	perc_race_other	% of HH who is other	10.298	0.892	11.539
	perc_race_multiple	% of HH who is two or more races	-0.369	1.224	-0.301

[6] White was used as the base alternative in this model.

ALTERNATIVE[6]	PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
Two or more races	(Intercept)		-3.426	0.832	-4.117
	perc_race_white	% of HH who is white	-3.217	0.877	-3.668
	perc_race_afam	% of HH who is Black or African American	3.764	1.151	3.271
	perc_race_asian	% of HH who is Asian	2.760	1.137	2.427
	perc_race_other	% of HH who is other	2.156	1.468	1.468
	perc_race_multiple	% of HH who is two or more races	9.200	0.874	10.523

McFadden's rho-squared: 0.854

TABLE 5-7: RACE IMPUTATION MODEL FOR HOUSEHOLDS WITH NO ADULTS REPORTING RACE

ALTERNATIVE[7]	PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
Black or African American	(Intercept)		-0.308	0.231	-1.334
	frace_white_only	Fraction of people in block group who are white	-3.082	0.298	-10.347
	frace_afam_only	Fraction of people in block group who are Black or African American	1.544	0.385	4.010
	frace_asian_only	Fraction of people in block group who are Asian	0.057	0.493	0.117
	frace_other	Fraction of people in block group who are other	-0.523	0.726	-0.721
	frace_multiple	Fraction of people in block group who are two or more races	1.696	1.144	1.483
	factor(college_educated)1	Has Associate Degree	0.254	0.226	1.121
	factor(college_educated)2	Has Bachelor’s Degree	-0.763	0.183	-4.162
	factor(college_educated)3	Has Master/PhD	-0.177	0.200	-0.886
	factor(employed)1	Employed part-time	0.142	0.206	0.691
	factor(employed)2	Employed full-time	0.449	0.159	2.830
	factor(num_people_cat)2	HH size = 2	0.219	0.163	1.342
	factor(num_people_cat)3	HH size = 3	0.839	0.196	4.290
	factor(num_people_cat)4	HH size = 4	0.553	0.234	2.364
	factor(num_people_cat)5	HH size = 5+	0.714	0.232	3.082
	own_home	Owns home	-1.254	0.173	-7.258
	single_family_home	Lives in a single-family home	-0.436	0.173	-2.524
	is_student	Is a student (adult only)	0.045	0.161	0.280

[7] White was used as the base alternative in this model.

ALTERNATIVE[7]	PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
	factor(has_license)1	Has a driver license	-1.023	0.161	-6.354
	h_income_0k_25k	Income less than $25k	0.517	0.142	3.635
	h_income_25k_50k	Income between $25k and $50k	0.733	0.115	6.353
	h_income_50k_75k	Income between $50k and $75k	-0.329	0.148	-2.222
	h_income_75k_100k	Income between $75k and $100k	-0.174	0.160	-1.087
	h_income_100k_150k	Income between $100k and $150k	-0.529	0.187	-2.832
	h_income_150k_plus	Income greater than $150k	-0.526	0.212	-2.483
Asian	(Intercept)		-2.177	0.266	-8.168
	frace_white_only	Fraction of people in block group who are white	-1.673	0.306	-5.461
	frace_afam_only	Fraction of people in block group who are Black or African American	0.229	0.475	0.482
	frace_asian_only	Fraction of people in block group who are Asian	2.875	0.459	6.262
	frace_other	Fraction of people in block group who are other	-2.916	0.941	-3.098
	frace_multiple	Fraction of people in block group who are two or more races	-0.692	1.186	-0.583
	factor(college_educated)1	Has Associate Degree	0.232	0.255	0.908
	factor(college_educated)2	Has Bachelor's Degree	0.362	0.154	2.346
	factor(college_educated)3	Has Master/PhD	0.787	0.159	4.948
	factor(employed)1	Employed part-time	0.202	0.197	1.028

ALTERNATIVE[7]	PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
	factor(employed)2	Employed full-time	0.447	0.136	3.284
	factor(num_people_cat)2	HH size = 2	0.400	0.174	2.303
	factor(num_people_cat)3	HH size = 3	1.509	0.186	8.126
	factor(num_people_cat)4	HH size = 4	1.578	0.195	8.106
	factor(num_people_cat)5	HH size = 5+	1.452	0.211	6.866
	own_home	Owns home	-0.179	0.143	-1.247
	single_family_home	Lives in a single-family home	-1.158	0.135	-8.587
	is_student	Is a student (adult only)	0.383	0.158	2.425
	factor(has_license)1	Has a driver license	-0.423	0.187	-2.260
	h_income_0k_25k	Income less than $25k	-0.364	0.181	-2.007
	h_income_25k_50k	Income between $25k and $50k	-0.471	0.145	-3.241
	h_income_50k_75k	Income between $50k and $75k	-0.379	0.118	-3.209
	h_income_75k_100k	Income between $75k and $100k	-0.507	0.132	-3.829
	h_income_100k_150k	Income between $100k and $150k	-0.285	0.121	-2.347
	h_income_150k_plus	Income greater than $150k	-0.172	0.130	-1.323
Other	(Intercept)		-1.671	0.308	-5.426
	frace_white_only	Fraction of people in block group who are white	-1.641	0.367	-4.472
	frace_afam_only	Fraction of people in block group who are Black or African American	-1.096	0.653	-1.679
	frace_asian_only	Fraction of people in block group who are Asian	0.144	0.702	0.204

ALTERNATIVE[7]	PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
	frace_other	Fraction of people in block group who are other	0.134	0.969	0.138
	frace_multiple	Fraction of people in block group who are two or more races	0.788	1.496	0.527
	factor(college_educated)1	Has Associate Degree	-0.115	0.286	-0.403
	factor(college_educated)2	Has Bachelor's Degree	-0.367	0.182	-2.010
	factor(college_educated)3	Has Master/PhD	-0.203	0.204	-0.991
	factor(employed)1	Employed part-time	0.130	0.221	0.585
	factor(employed)2	Employed full-time	-0.294	0.171	-1.725
	factor(num_people_cat)2	HH size = 2	0.510	0.202	2.529
	factor(num_people_cat)3	HH size = 3	0.971	0.245	3.961
	factor(num_people_cat)4	HH size = 4	0.058	0.322	0.180
	factor(num_people_cat)5	HH size = 5+	1.105	0.275	4.023
	own_home	Owns home	0.028	0.195	0.145
	single_family_home	Lives in a single-family home	-0.713	0.176	-4.043
	is_student	Is a student (adult only)	-0.577	0.233	-2.472
	factor(has_license)1	Has a driver license	-0.455	0.219	-2.078
	h_income_0k_25k	Income less than $25k	0.277	0.188	1.469
	h_income_25k_50k	Income between $25k and $50k	-0.248	0.165	-1.501
	h_income_50k_75k	Income between $50k and $75k	-0.334	0.157	-2.129
	h_income_75k_100k	Income between $75k and $100k	-0.253	0.161	-1.568
	h_income_100k_150k	Income between $100k and $150k	-0.448	0.171	-2.622
	h_income_150k_plus	Income greater than $150k	-0.664	0.203	-3.274

ALTERNATIVE[7]	PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
Two or more races	(Intercept)		-1.848	0.260	-7.113
	frace_white_only	Fraction of people in block group who are white	-1.880	0.306	-6.149
	frace_afam_only	Fraction of people in block group who are Black or African American	0.185	0.476	0.388
	frace_asian_only	Fraction of people in block group who are Asian	1.186	0.500	2.373
	frace_other	Fraction of people in block group who are other	-1.717	0.871	-1.970
	frace_multiple	Fraction of people in block group who are two or more races	0.378	1.220	0.310
	factor(college_educated)1	Has Associate Degree	-0.293	0.269	-1.088
	factor(college_educated)2	Has Bachelor's Degree	-0.403	0.161	-2.501
	factor(college_educated)3	Has Master/PhD	-0.458	0.199	-2.305
	factor(employed)1	Employed part-time	0.201	0.195	1.031
	factor(employed)2	Employed full-time	0.453	0.155	2.931
	factor(num_people_cat)2	HH size = 2	0.738	0.181	4.087
	factor(num_people_cat)3	HH size = 3	1.180	0.207	5.696
	factor(num_people_cat)4	HH size = 4	0.965	0.224	4.317
	factor(num_people_cat)5	HH size = 5+	0.669	0.240	2.790
	own_home	Owns home	-0.553	0.164	-3.371
	single_family_home	Lives in a single-family home	-0.229	0.159	-1.441
	is_student	Is a student (adult only)	0.927	0.152	6.105
	factor(has_license)1	Has a driver license	-0.520	0.171	-3.048

ALTERNATIVE[7]	PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
	h_income_0k_25k	Income less than $25k	0.473	0.143	3.299
	h_income_25k_50k	Income between $25k and $50k	0.103	0.122	0.842
	h_income_50k_75k	Income between $50k and $75k	-0.540	0.132	-4.100
	h_income_75k_100k	Income between $75k and $100k	-0.472	0.137	-3.431
	h_income_100k_150k	Income between $100k and $150k	-0.619	0.136	-4.543
	h_income_150k_plus	Income greater than $150k	-0.794	0.156	-5.077

McFadden's rho-squared: 0.144

Expansion of Household and Person Data

Table 5-8 provides the distribution of weights that were calculated for each weighting geography in the sample.

Table 5-9 summarizes the ratio of the final weight against the initial expansion factor (the weight derived based on the probability of being sampled). In the weighting process, the ratio of the final weight to the initial weight was constrained to be in the range of 0.25 to 5.0 for each household. Allowing the weights to be outside this range would enable the process to match the ACS PUMS targets more exactly, but at the cost of having more extremely high or low weights and the introduction of more variance. Considering that the PUMS targets are estimates based on census survey data, it is not good practice to try to match the targets too precisely at the expense of allowing the survey weights to vary too widely. The range of 0.25 to 5.0 was arrived at after testing alternative limits and judging the best trade-off between accuracy and variability. With these weights, the ratios are near one which suggests that the final weights (on average) have not deviated significantly from the initial expansion factors.

TABLE 5-8: SUMMARY STATISTICS OF THE FINAL WEIGHTS

WEIGHTING GEOGRAPHY	MIN	MEAN	MEDIAN	MAX
Anoka MN	34.418	236.522	89.232	1005.067
Chisago & Isanti MN	49.337	357.586	198.226	974.328
Dakota MN	34.374	217.856	49.158	1009.635
Goodhue, Rice, & Le Sueur MN	34.488	210.026	82.680	986.515
Hennepin MN – Minneapolis	34.359	126.548	48.566	967.213
Hennepin MN – Other	34.568	219.003	50.620	1010.625
Pierce, Polk, & St. Croix WI	50.186	556.649	543.976	977.228
Ramsey MN – Other	34.402	172.139	50.389	949.100
Ramsey MN - St. Paul	34.405	139.643	48.754	967.448
Scott & Carver MN	35.369	258.948	107.396	1025.494
Sherburne, Wright, McLeod, & Sibley MN	35.575	313.934	236.660	962.531
Washington MN	34.498	223.144	65.227	1014.186

TABLE 5-9: SUMMARY STATISTICS FOR THE RATIO OF FINAL TO INITIAL WEIGHTS

WEIGHTING GEOGRAPHY	MIN	MEAN	MEDIAN	MAX
Anoka MN	0.250	1.255	0.469	4.900
Chisago & Isanti MN	0.250	1.813	1.005	4.938
Dakota MN	0.250	1.156	0.253	4.931
Goodhue, Rice, & Le Sueur MN	0.251	1.080	0.427	4.999
Hennepin MN – Minneapolis	0.250	0.753	0.253	4.981
Hennepin MN – Other	0.252	1.188	0.260	4.995
Pierce, Polk, & St. Croix WI	0.254	2.839	2.757	4.952
Ramsey MN – Other	0.250	0.920	0.260	4.889
Ramsey MN - St. Paul	0.250	0.842	0.255	4.982
Scott & Carver MN	0.251	1.323	0.556	5.000
Sherburne, Wright, McLeod, & Sibley MN	0.253	1.622	1.199	4.878
Washington MN	0.251	1.134	0.318	4.945

Final Household and Person Weights

The final weights are effective in facilitating close matches to the regional totals for people, households, persons-in-households, and vehicles-in-households when using this dataset. The expanded and weighted survey values match the targets well, with nearly all household and person categories within 5%, while keeping the weights relatively constrained. Race is a notable exception in several weighting geographies due to small sample sizes for non-white participants. For this reason, data users should use caution when conducting any analysis by race, particularly for the more rural counties.

Table 5-17 provides a comparison between the resulting weights and the targets. Figure 5-1, Figure 5-2, Figure 5-3, and Figure 5-4 provide graphical comparisons between the resulting weights and the targets. As mentioned previously, matching the survey data to the target data even more closely can be achieved by relaxing the constraints on the ratio of the final to initial weights. However, this introduces more variance in the final weights and thereby increases the statistical error in any estimates. Allowing for more extreme weights also increases the likelihood of travel behavior analyses being impacted by extreme or outlier weights, which could unknowingly bias an estimate.

This data will be used in an activity-based model, so priority was given to matching the person targets. As noted above, the PUMS targets are in fact just estimates themselves, so matching the targets perfectly at the expense of increased statistical error is generally not recommended.

5.5 CREATING DAY WEIGHTS WITH MULTIDAY SURVEY DATA

With the shift to data collection using smartphone applications such as rMove, it has become cost effective to capture multiple days of data for each respondent. The question then is how to combine the multi-day smartphone-based data with the single-day data from online and call center participants using a consistent weighting method.

RSG's usual approach to create an "average weekday" day-level weight for multi-day smartphone data has been as follows:

- Weight to regional targets to obtain the household- and person-level weights for the included respondents.
- Define weekdays as Monday through Thursday as discussed previously.
- For each respondent, count the number of weekdays (N) for which the respondent provided complete and valid data. Set the person-day level weight equal to the person-level weight divided by N. In this way, when the data is weighted and aggregated, the sum of the person-day weights across days for each person is equal to the person weight, and the weighted results will reflect an average day for each respondent.

This method results in an "average weekday" for each respondent regardless of the number of days of data provided making the multi-day smartphone-based data compatible with the single-day online and call center-based data.

5.6 ADJUSTING FOR NONRESPONSE BIAS IN DAY-PATTERN AND TRIP RATES

It has been found in previous surveys that the trip rates from the smartphone-based survey data are 15–20% higher than those from online and call center-based survey data. There are three main reasons for this:

- Smartphone-owning households have different socio-demographic characteristics than non-smartphone households and tend to make more trips.
- There are about twice as many "stay at home" days with no reported trips in the online and call center-based data in comparison to the smartphone-based data.
- Even on days with one or more reported trips, there are more trips per day reported on average in the smartphone-based data than in the online and call center-based data.

All three of these factors are interrelated and need to be isolated from each other through careful analysis and a series of weighting adjustments, as described in the sections below.

A typical method for adjusting the trip rates for online and call center-based data to match smartphone-based data is to adjust the weights at the trip-level. However, RSG employs a two-stage approach, first adjusting weights at the person-day level to adjust for biases in day-pattern types, and then a second stage to adjust weights at the trip-level. There are two key reasons for this:

1. First, as noted above, one of the key reasons that trip rates are different between the methods is the higher proportion of “stay at home” days with no trips reported in the online and call center-based data. While some of this difference is likely legitimate due to differences in demographics, some of it is also likely due to so-called “soft refusal,” whereby it is easy for respondents using the online or call center diary recall method to state that they did not make any trips when in fact they did. It is important to identify the extent of such bias and correct for it at the person-day level, because the “stay at home” cases have no trip records in the data, so the correction cannot be made by factoring weights at the trip-level.
2. Second, most activity-based models include a model component to predict the day-pattern type, e.g. stay at home, make mandatory (work or school) trips (and possibly other trips), or make non-mandatory trips only. If the data is used to calibrate such a model at the person-day and household-day levels, it is important to correct any biases that distort the day-pattern types in the data.

Day-Pattern Adjustments

RSG has developed a method for identifying biases in day-patterns and adjusting for them in the weighting process. The following steps were taken to adjust for biases in day-patterns:

1. A multinomial choice model was estimated at the person-day level. There were three day-pattern choices that were modeled: (1) participant made no trips, (2) participant made mandatory (work or school) trips (and possibly other trips) and (3) participant made non-mandatory trips only. The model included the following variables as independent variables:
   - Income
   - Presence of vehicles in the household
   - Worker status
   - Student status
   - Age

   The model also included additional bias variables for adults proxied via online and call center diary and smartphone ownership, which capture the trip reporting bias after accounting for the variables listed above. The day pattern model specification and coefficients are shown in Table 5-10.
2. The estimated model was applied to each person-day to calculate the probabilities of each of the three-day-pattern alternatives. Then the weighted probabilities were added across the sample within the categories of person-days—(a) those provided by respondents’ own smartphones, and (b) those provided by online and call center-diary. The aggregate choice shares from applying the model should match the actual choice

shares in the data. This provides a check that the model is being applied correctly to the data.

3. Step 2 was repeated, but this time, any bias coefficients in the model were set to 0. None of the bias coefficients apply to smartphone respondents, so the results for this category were unchanged. For the last two categories (online/call center and proxied participants) the new predictions were what the choice shares would be if any biases did not exist (but all socio-demographic factors still apply). Table 5-11 shows the percent of weighted days in each category before and after removing the bias, by household group type and smartphone participation status.
4. The modified aggregate choice predictions (segmented by weighting geography) were added as a new set of targets in the household/person weighting process described in previous sections. Then the number of person-days for each day-pattern type for each person were counted and used as the corresponding input for weighting at the person-level.
5. The IPF weighting procedure was then rerun with this new added target. The result was that the online/call center households with no trips tended to have their weights reduced, while those with trips (and particularly with non-mandatory trips only) tended to have their weights increased to match the adjusted targets. The weights for smartphone respondents remained essentially unchanged. The advantage of adding these new targets into the household- and person-level weighting process and using all of the targets simultaneously is that all of the household- and person- level weighting targets were still matched as well, which would not be the case if the adjustment was made to the new day-pattern targets in isolation.

TABLE 5-10: DAY PATTERN MODEL SUMMARY

ALTERNATIVE[8]	PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
Makes mandatory trips	(Intercept)		-2.676	0.109	-24.651
Makes mandatory trips	online_data	Online/call center diary data	-0.164	0.060	-2.720
Makes mandatory trips	zero_vehicle	No vehicles in household	-0.481	0.106	-4.554
Makes mandatory trips	income_aggregate2	Income 25k-50k	0.369	0.083	4.465
Makes mandatory trips	income_aggregate3	Income 50k-75k	0.670	0.086	7.805
Makes mandatory trips	income_aggregate4	Income 75k-100k	0.714	0.088	8.134
Makes mandatory trips	income_aggregate5	Income > 100k	0.665	0.078	8.576
Makes mandatory trips	income_aggregate999	Prefer not to answer income	0.406	0.090	4.519
Makes mandatory trips	age_under_35	Age < 35 years	0.867	0.079	10.959
Makes mandatory trips	age_35_65	Age between 35-65 years	0.781	0.062	12.689
Makes mandatory trips	employed	Employed full/part/self	3.304	0.058	56.597
Makes mandatory trips	is_student	Full or part-time student	0.849	0.082	10.340
Makes mandatory trips	was_proxied	Adult proxied	-0.516	0.060	-8.899
Makes mandatory trips	owns_phone	Adult owns smartphone	0.428	0.065	6.561
Makes mandatory trips	online_data_age_under_35	Online/call center diary data x Age	-0.925	0.094	-9.866
Makes non-mandatory trips only	(Intercept)		0.665	0.081	8.234
Makes non-mandatory trips only	online_data	Online/call center diary data	-0.031	0.056	-0.556
Makes non-mandatory trips only	zero_vehicle	No vehicles in household	-0.207	0.079	-2.628
Makes non-mandatory trips only	income_aggregate2	Income 25k-50k	0.091	0.066	1.388
Makes non-mandatory trips only	income_aggregate3	Income 50k-75k	0.185	0.072	2.566

[8] The no-travel day alternative was used as the base alternative in this model.

Makes non-mandatory trips only	income_aggregate4	Income 75k-100k	0.214	0.075	2.850
Makes non-mandatory trips only	income_aggregate5	Income > 100k	0.179	0.065	2.758
Makes non-mandatory trips only	income_aggregate999	Prefer not to answer income	-0.147	0.073	-2.015
Makes non-mandatory trips only	age_under_35	Age < 35 years	-0.193	0.071	-2.729
Makes non-mandatory trips only	age_35_65	Age between 35-65 years	-0.019	0.047	-0.398
Makes non-mandatory trips only	employed	Employed full/part/self	0.017	0.043	0.389
Makes non-mandatory trips only	is_student	Full or part-time student	0.070	0.084	0.831
Makes non-mandatory trips only	was_proxied	Adult proxied	-0.666	0.049	-13.455
Makes non-mandatory trips only	owns_phone	Adult owns smartphone	0.344	0.055	6.218
Makes non-mandatory trips only	online_data_age_under_35	Online/call center diary data x Age	-0.927	0.098	-9.491

McFadden's rho-squared: 0.235

TABLE 5-11: DAY CATEGORY BY HOUSEHOLD GROUP & SMARTPHONE PARTICIPATION, WITH AND WITHOUT BIAS REMOVED

		WITH BIAS			*BIAS REMOVED*		
HOUSEHOLD GROUP TYPE	SMARTPHONE PARTICIPANT	NO-TRAVEL DAYS	MANDATORY TRIP DAYS	NON-MANDATORY TRIP DAYS	NO-TRAVEL DAYS	MANDATORY TRIP DAYS	NON-MANDATORY TRIP DAYS
All adults use own phone	Yes	9%	63%	28%	9%	63%	28%
Online diary	No	23%	37%	41%	19%	40%	41%

Trip-Rate Adjustments

After the first stage of adjustment described above, the new person-day weights were applied to compare the trip rates for the different survey participation methods. Adjusting the weights for day-pattern biases reduced the discrepancy in trip rates between methods, but it did not eliminate it altogether. In practice, the difference in trip rates tends to be higher for non-mandatory trips than for mandatory trips, as respondents are less likely to omit their work and school trips in recall-based diary methods. The differences can also be large for non-home-based trips, since online/call center and by-proxy respondents often tend to omit intermediate stops on multi-stop tours.

The process for adjusting the trip-level weights was relatively analogous to that described above for day-pattern types but was somewhat simpler. The starting point for the two-stage trip-rate bias correction was the person-day weights. The following steps were then taken to adjust trip rates:

1. Trips were segmented into the following four trip types that have different levels of underreporting. Then for each person-day in the sample, the number of trips were counted by type.
    a. Home-based work/school trips
    b. Home-based other trips
    c. Non-home-based work/school trips
    d. Non-home-based other trips
2. For each trip type, a Poisson regression model was estimated where the dependent variable was the number of trips of that type for the person-day. The independent variables were the same set of household and person variables listed above for the day-pattern models, plus dummy variables for online and call center-based person-days.

For each person-day and for each trip type, the estimated regression model was applied with and without the bias coefficients. The ratio of the two estimates resulted in a factor to apply to the trip weight for that person-day. For example, if the model predicted 1.10 trips with the estimated model and 1.32 trips with the bias parameters set to 0 for an online or call center-based person-day, then a factor of 1.32/1.10 = 1.2 was used to multiply the person-day weight to get an adjusted trip weight. For smartphone respondents, the bias coefficients do not apply, so the factor was always 1.0 and the trip weight equaled the person-day weight. A lower bound of 1.0 and an upper bound of 2.0 were placed on ratios to avoid extreme adjustment to the weights. The specifications for each of the four regression models are shown in Table 5-12, Table 5-13, Table 5-14, and Table 5-15. The resulting trip adjustment factors by diary method and trip type are shown in Table 5-16. Non-home-based trips have rather high adjustment factors for online and call center-based diary participants, which is likely due to poor recall of intermediate stops between home and another location. As smartphone ownership increases among adults, the need to assign adults to proxy for other adults via smartphone will decrease.

TABLE 5-12: HOME-BASED WORK TRIP MODEL

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
(Intercept)		-2.624	0.045	-58.788
online_data	Online/call center diary data	0.074	0.017	4.277
age_under_25	Under age 25	0.548	0.038	14.420
age_25_45	Age 25 to 45	0.286	0.034	8.505
age_45_65	Age 45 to 65	0.363	0.033	11.159
employed_ft	Employed full-time	2.253	0.037	61.607
employed_pt	Employed part-time	1.969	0.039	50.525
employed_self	Self-employed	1.908	0.046	41.469
Bachelors	Has bachelor's degree	0.039	0.015	2.623
graduate_degree	Has masters/PhD	0.063	0.017	3.622
is_student	Is student	0.184	0.024	7.656
work_loc_varies	Work location varies	0.025	0.018	1.366
has_kids	HH has children	-0.067	0.014	-4.769
two_plus_jobs	Works 2+ jobs	0.213	0.019	10.962
sf_home	Lives in single family home	0.035	0.015	2.343
owns_phone	Owns phone	0.035	0.026	1.324

McFadden's rho-squared: 0.146

TABLE 5-13: HOME-BASED OTHER TRIP MODEL

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
(Intercept)		0.531	0.019	27.341
online_data	Online/call center diary data	-0.162	0.013	-12.804
age_under_25	Under age 25	-0.415	0.024	-17.065
age_25_45	Age 25 to 45	-0.084	0.016	-5.337
age_45_65	Age 45 to 65	-0.059	0.014	-4.206
employed_ft	Employed full-time	-0.417	0.012	-34.037
employed_pt	Employed part-time	-0.130	0.016	-8.266
employed_self	Self-employed	-0.182	0.022	-8.294
Bachelors	Has bachelor's degree	0.205	0.010	19.919
graduate_degree	Has masters/PhD	0.252	0.012	21.529
is_student	Is student	-0.091	0.020	-4.633
work_loc_varies	Work location varies	-0.018	0.015	-1.196
has_kids	HH has children	0.274	0.010	27.193
two_plus_jobs	Works 2+ jobs	0.103	0.016	6.510
sf_home	Lives in single family home	0.014	0.010	1.395
owns_phone	Owns phone	0.096	0.016	5.910

McFadden's rho-squared: 0.042

TABLE 5-14: NON-HOME-BASED WORK TRIP MODEL

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
(Intercept)		-2.338	0.051	-45.665
online_data	Online/call center diary data	-0.760	0.022	-35.118
age_under_25	Under age 25	0.170	0.039	4.405
age_25_45	Age 25 to 45	0.127	0.034	3.777
age_45_65	Age 45 to 65	0.169	0.033	5.095
employed_ft	Employed full-time	2.146	0.036	59.636
employed_pt	Employed part-time	1.655	0.039	42.537
employed_self	Self-employed	1.787	0.045	40.096
Bachelors	Has bachelor’s degree	-0.022	0.014	-1.564
graduate_degree	Has masters/PhD	-0.049	0.016	-3.018
is_student	Is student	0.383	0.022	17.744
work_loc_varies	Work location varies	0.245	0.016	15.399
has_kids	HH has children	0.122	0.013	9.410
two_plus_jobs	Works 2+ jobs	0.230	0.018	12.709
sf_home	Lives in single family home	0.046	0.014	3.281
owns_phone	Owns phone	0.258	0.037	6.897

McFadden’s rho-squared: 0.219

TABLE 5-15: NON-HOME-BASED OTHER TRIP MODEL

PARAMETER	DESCRIPTION	ESTIMATE	STD ERROR	T-STATISTIC
(Intercept)		0.456	0.027	17.185
online_data	Online/call center diary data	-0.544	0.018	-30.379
age_under_25	Under age 25	-0.449	0.030	-15.128
age_25_45	Age 25 to 45	-0.173	0.020	-8.635
age_45_65	Age 45 to 65	-0.111	0.018	-6.218
employed_ft	Employed full-time	-0.631	0.016	-40.597
employed_pt	Employed part-time	-0.232	0.020	-11.812
employed_self	Self-employed	-0.389	0.029	-13.248
Bachelors	Has bachelor’s degree	0.021	0.013	1.593
graduate_degree	Has masters/PhD	0.023	0.015	1.463
is_student	Is student	0.074	0.023	3.163
work_loc_varies	Work location varies	0.071	0.019	3.748
has_kids	HH has children	0.180	0.013	13.554
two_plus_jobs	Works 2+ jobs	-0.009	0.021	-0.427
sf_home	Lives in single family home	-0.158	0.013	-12.565
owns_phone	Owns phone	0.208	0.023	8.942

McFadden’s rho-squared: 0.043

TABLE 5-16: TRIP ADJUSTMENT FACTORS

DIARY METHOD	HOME-BASED WORK	HOME-BASED OTHER	NON-HOME-BASED WORK	NON-HOME-BASED OTHER
Smartphone participant	1.00	1.00	1.00	1.00
Online/call center diary	1.00	1.18	2.00	1.72

5.7 FINAL WEIGHTS AND RECOMMENDED USE

The final weights provided with the dataset are described below:

- **hh_weight**: The resulting weights from expanding to the PUMS data. The sum of the hh_weight in the household table reflects the total number of households in the survey region.
    - ***Uses**: This weight should be used for household-level and vehicle-level analyses.*
- **person_weight**: The resulting weights from expanding to the PUMS data. The sum of the person_weight in the person table reflects the total number of persons in the survey region.
    - ***Uses**: This weight should be used for person-level analyses.*
- **day_weight**: The adjusted day-level weights, which are the hh_weight divided by the number of complete days and adjusted based on the day category (no trips, mandatory trips, or non-mandatory trips only). The sum of the day_weight should match the sum of the person_weight in the person table.
    - ***Uses**: This weight should be used for day-level analyses.*
- **trip_weight**: The adjusted trip-level weights, which are the trip factor multiplied by the day_weight. The sum of trip_weight in the trip table equals the number of trips taken by residents of the survey region on a “typical day,” as estimated by this survey and weighting approach.
    - ***Uses**: This weight should be used for trip-level analyses.*

5.8 WEIGHTING VALIDATION

TABLE 5-17: DIFFERENCES BETWEEN HOUSEHOLD WEIGHTED SAMPLE AND TARGET PUMS DATA

VARIABLE	ANOKA COUNTY MN	CHISAGO & ISANTI COUNTIES MN	DAKOTA COUNTY MN	GOODHUE, RICE, & LE SUEUR COUNTIES MN	HENNEPIN COUNTY MN – MINNEAPOLIS	HENNEPIN COUNTY MN - OTHER	PIERCE, POLK, & ST. CROIX COUNTIES WI	RAMSEY COUNTY MN - OTHER	RAMSEY COUNTY MN - ST. PAUL	SCOTT & CARVER COUNTIES MN	SHERBURNE, WRIGHT, MCLEOD, & SIBLEY	WASHINGTON COUNTY MN
h_income_0k_25k	-105	-217	-1,964	-115	-2,701	-441	-12,285	-1,514	-1,235	-327	-554	-164
h_income_25k_50k	123	-266	-1,879	-479	2,843	808	3,547	682	-216	-104	119	134
h_income_50k_75k	152	255	-413	-107	1,631	1,069	4,282	368	159	63	179	58
h_income_75k_100k	93	48	-428	-180	2,149	2,316	3,864	522	2,224	391	428	153
h_income_100k_150k	391	135	3,146	146	2,249	2,527	5,156	634	2,690	300	211	211
h_income_150k_plus	-125	113	1,346	-78	1,885	922	3,907	727	753	496	85	289
h_size1	334	163	-739	918	5,886	2,145	-138	580	1,572	-129	687	453
h_size2	1,088	-46	666	-1,799	3,073	4,858	5,287	1,014	1,185	-185	377	368
h_size3	304	-170	108	-71	2,141	2,223	6,885	824	1,866	241	-33	-71
h_size4	-519	153	-1,096	-34	-1,859	-609	4,103	381	1,273	517	-245	-34
h_size5plus	-679	-33	867	174	-1,184	-1,415	-7,666	-1,379	-1,520	373	-320	-35
h_worker_0	-997	-848	-3,064	-736	2,752	-216	3,025	-2,115	794	222	-1,021	59
h_worker_1	684	304	873	439	2,672	3,326	2,423	1,231	1,027	302	665	371
h_worker_2	672	318	2,525	-545	4,085	3,962	4,892	2,502	4,107	384	701	104
h_worker_3plus	169	293	-528	28	-1,452	129	-1,870	-198	-1,553	-90	123	147
h_car_0_or_1	66	-400	-326	74	11,371	4,056	225	3,018	6,499	-114	-54	126
h_car_2	11	69	-2,467	-524	-2,202	2,475	3,375	-319	-68	694	31	291
h_car_3plus	452	399	2,599	-363	-1,112	672	4,871	-1,280	-2,056	238	491	263
h_head_under_35	245	55	-506	334	5,677	1,756	-270	-366	2,832	263	27	253
h_head_35_64	811	389	3,726	-31	3,836	7,716	6,738	3,806	1,972	972	2,346	609
h_head_over_65	-528	-377	-3,414	-1,116	-1,456	-2,269	2,002	-2,020	-429	-416	-1,905	-181
h_has_kids	-1,229	-50	2,689	-1,204	895	6,577	3,993	1,187	3,631	1,881	-482	1,023
h_has_no_kids	1,758	117	-2,883	391	7,163	625	4,478	232	744	-1,063	949	-342
h_total	529	67	-193	-813	8,057	7,202	8,471	1,420	4,376	818	468	681
p_male	-173	-1,232	-1,602	-5,170	-2,962	-1,617	-1,960	-1,906	-1,899	-3,124	-1,770	-1,417
p_female	173	1,232	1,602	5,170	2,962	1,617	1,960	1,906	1,899	3,123	1,770	1,417
p_age0_4	-157	-95	6,292	303	-336	1,518	-2,624	765	807	673	385	12
p_age5_15	870	260	-233	928	-8,859	-1,289	-2,961	1,223	-2,214	629	439	285
p_age16_17	531	62	-311	344	-364	59	-764	546	-2,449	416	204	-74

p_age18_24	202	443	-3,807	396	-1,758	-584	-9,672	-1,290	-3,768	-301	648	-56
p_age25_44	331	-131	7,464	-777	9,104	2,786	6,978	3,111	4,590	-398	-221	514
p_age45_64	166	586	-5,617	-229	2,555	287	3,762	-131	777	-883	356	165
p_age65plus	-1,943	-1,125	-3,788	-965	-342	-2,777	5,281	-4,223	2,256	-136	-1,811	-846
p_worker	2,249	1,630	3,755	-339	1,821	4,931	5,605	4,396	1,949	1,062	2,640	1,761
p_nonworker	-2,249	-1,630	-3,755	339	-1,821	-4,931	-5,605	-4,396	-1,949	-1,063	-2,640	-1,761
p_univstudent	-109	289	-1,414	212	-3,973	-1,343	-6,786	1,299	-2,171	-269	77	50
p_not_univstudent	109	-289	1,414	-212	3,973	1,343	6,786	-1,299	2,171	268	-77	-50
p_white	5,800	1,332	14,468	5,237	24,305	13,920	511	7,634	20,044	11,823	8,459	2,532
p_afam	-5,509	-248	-7,436	-4,868	-14,584	-10,990	9	-5,003	-4,481	-10,440	-6,192	-1,826
p_asian	-55	-436	-2,643	138	-2,678	-1,596	108	-1,805	-13,029	-2,189	-2,695	-823
p_other	-11	-30	-1,710	-502	-5,629	-554	83	-411	-1,363	692	270	76
p_multiple	-225	-618	-2,678	-4	-1,414	-780	-711	-415	-1,171	114	159	41
p_commute_none	3,853	1,356	3,054	947	1,143	2,767	925	3,441	1,355	847	4,133	1,804
p_commute_car_alone	-3,397	-1,152	-2,765	-1,435	-1,829	-2,749	1,788	-2,597	-2,249	-713	-3,743	-1,650
p_commute_carpool	-393	-183	-459	2	-306	-485	206	-948	-462	-284	-369	-115
p_commute_walkbike	-114	-22	-115	290	542	95	-2,977	22	540	116	-155	-20
p_commute_other	50	2	285	196	450	372	57	83	816	33	134	-19
p_made_no_trips	-181	-8	98	-3	-133	509	-630	-54	-108	13	-1,259	122
p_made_mandatory_trips	529	-2	-163	-1	123	-303	1,948	153	398	17	4,178	-255
p_made_nonmandatory_only	-204	6	59	1	1	-120	-825	-48	-178	-47	-1,724	86
p_made_not_applicable	-144	4	7	3	9	-86	-493	-51	-112	16	-1,195	46
p_total	0	0	-0	0	-0	-0	-0	0	-0	-1	0	-0

FIGURE 5-1: % DIFFERENCES BETWEEN HOUSEHOLD WEIGHTED SAMPLE AND TARGET PUMS DATA (ANOKA, WASHINGTON, AND RAMSEY COUNTIES, MN)

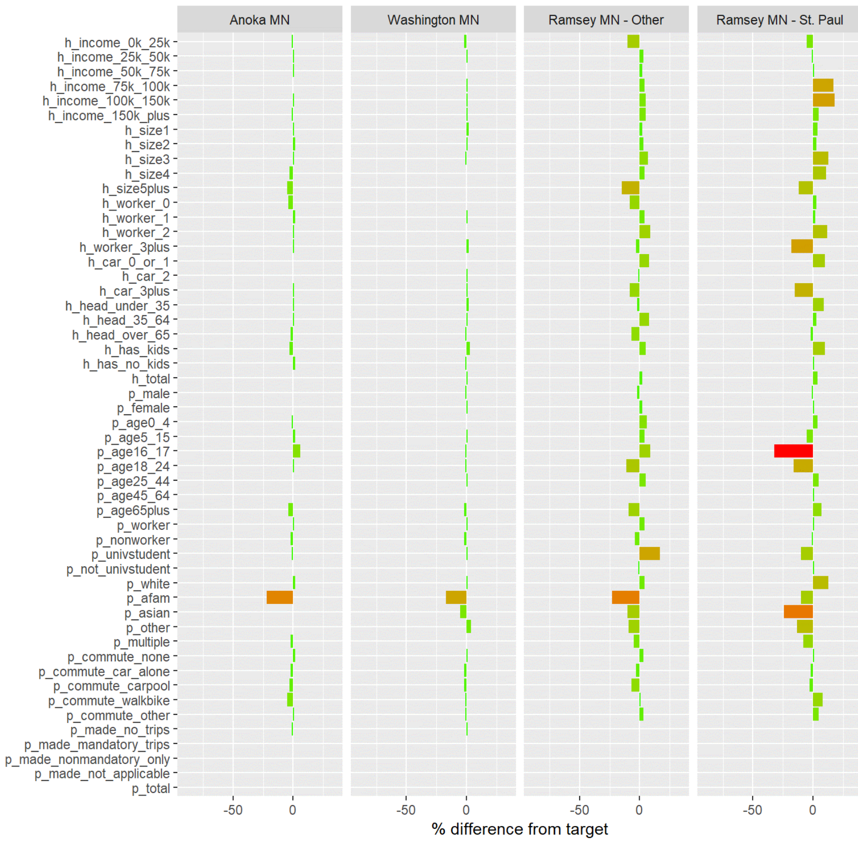


FIGURE 5-2: % DIFFERENCES BETWEEN HOUSEHOLD WEIGHTED SAMPLE AND TARGET PUMS DATA (HENNEPIN, DAKOTA, SCOTT, AND CARVER COUNTIES, MN)

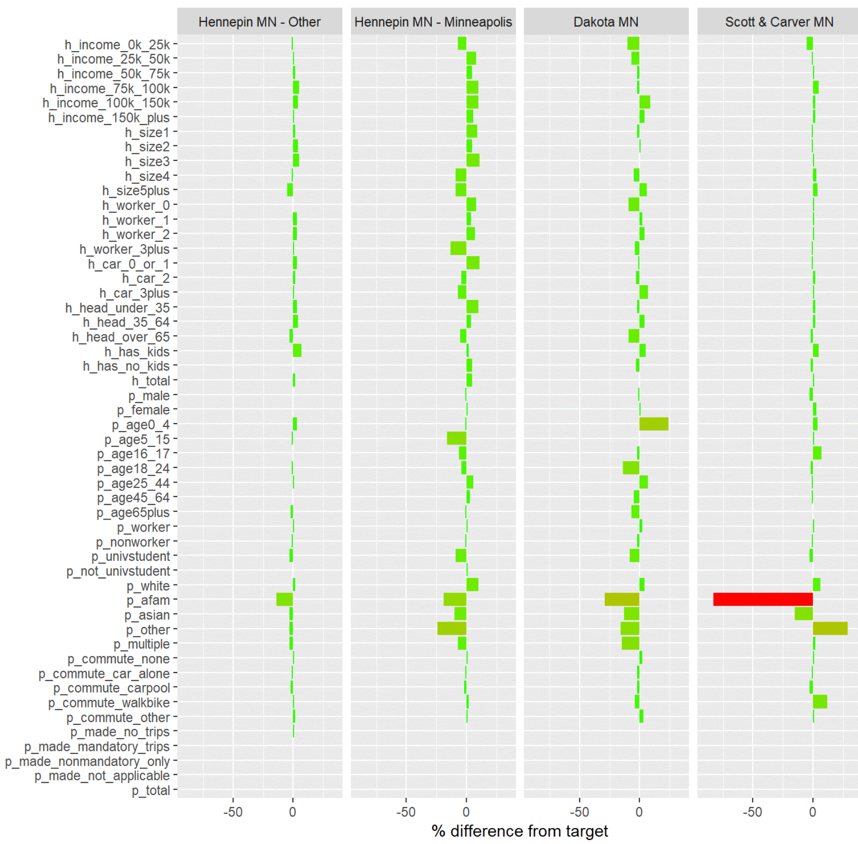


FIGURE 5-3: % DIFFERENCES BETWEEN HOUSEHOLD WEIGHTED SAMPLE AND TARGET PUMS DATA (CHISAGO, ISANTI, SHERBURNE, WRIGHT, MCLEOD, AND SIBLEY COUNTIES, MN)

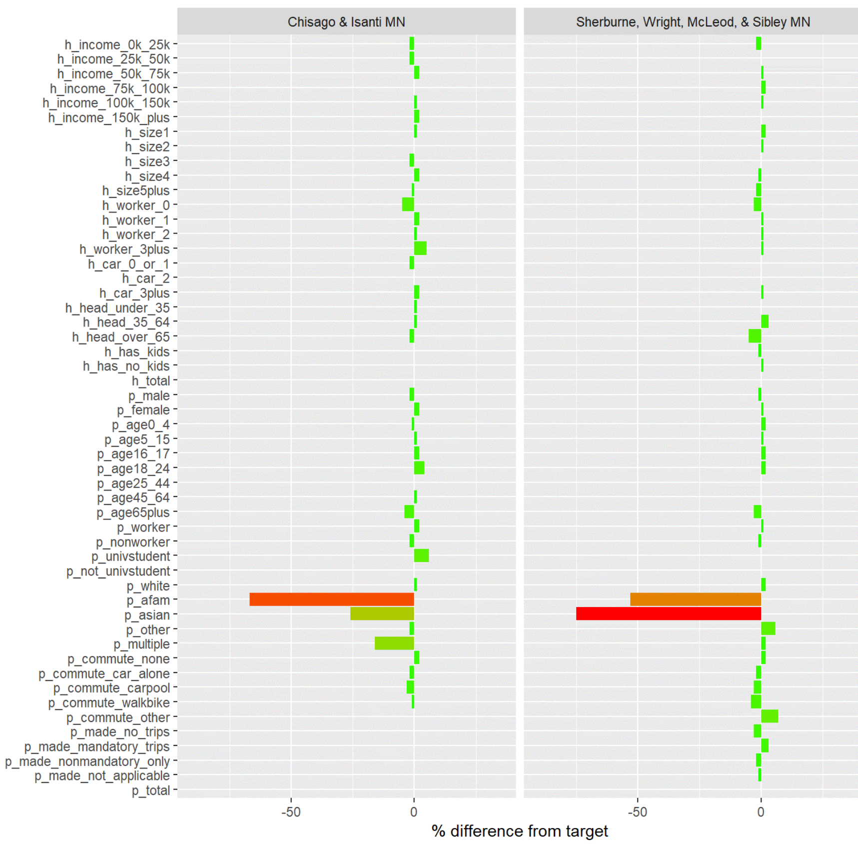
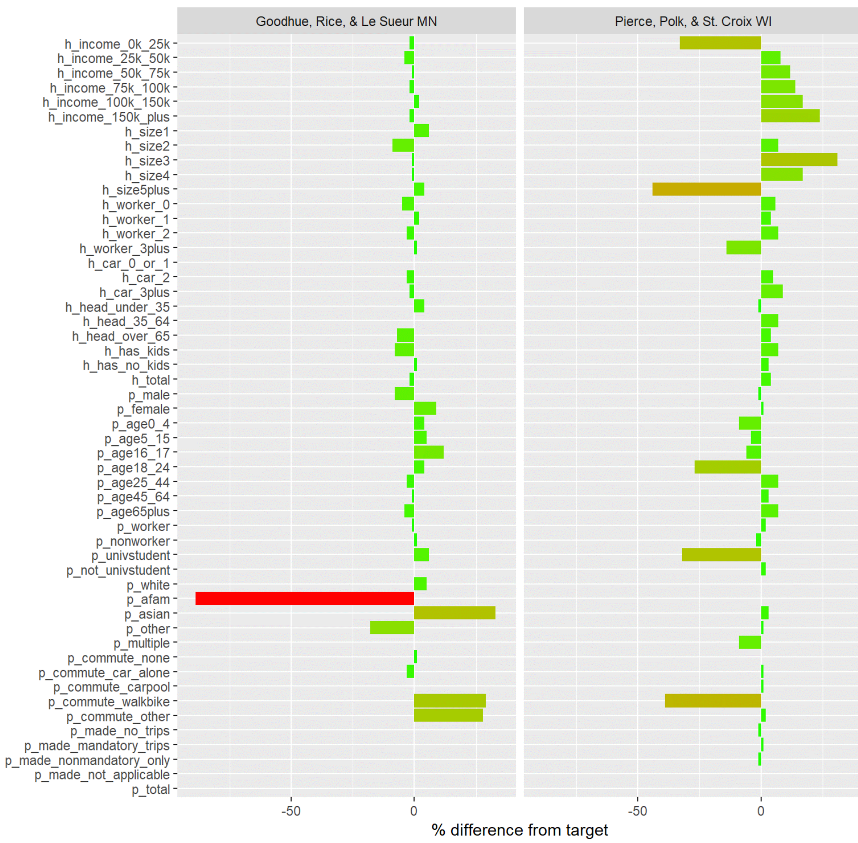


FIGURE 5-4: % DIFFERENCES BETWEEN HOUSEHOLD WEIGHTED SAMPLE AND TARGET PUMS DATA (GOODHUE, RICE, AND LE SUEUR COUNTIES, MN AND PIERCE, POLK, AND ST. CROIX COUNTIES, WI)



6.0 CHILD TRIP ADJUSTMENTS

Trip representation problems remained for children (members under age 18) in the dataset, particularly for school trips. Children's trips were reported by proxy by an adult in the household. Households who participated using smartphone-based methods in particular had school trip representation problems. The under-reporting bias was strong with less than 40% of school-age children having school trips on weekdays across all diary types. This under-reporting particularly for smartphone-based methods may be due to the extra burden of reporting proxy travel over multiple days or privacy concerns related to reporting children's travel in a survey app that tracks trips so accurately.

6.1 METHODOLOGY

Given the possibility for surveying on weekdays where there was no school in a given district (e.g., in-service days or snow days), a target of between 80 and 90% of child-days was used when considering which days would receive a synthesized school trip. Synthesis was only relevant for children in public K-12 schools (children attending daycare, preschool, or homeschool were not included in the synthesis). Public holidays, known days with weather events, and days where the main school districts had off (e.g., school vacations, parent-teacher conference days) were also excluded. Since data collection was year-round, school trip adjustments were only made on weekdays during the months of October through May.

Commented [AA1]: Why this target?

The following steps comprised the school trip synthesis:

- Identify the child-days that would receive a synthesized school tour. These are days where a school trip was expected but no trip occurred before 4pm.
  - Trips were only synthesized for children where a school location was known. This results in a smaller improvement for children whose parents recruited using the rMove application. In this case, school locations were imputed from trips to/from school but since these are under-reported, there was no school location identified for a large number of these children.
  - Since the online diary respondents had a school attendance rate within the target range, no trips were synthesized for these children.
- Generate a sample dataset of school tours (between home and school) from the sample of child-days that went to school (where imputed purpose is school).
- For each day needing a school tour:
  - Trips to school start in the morning between 6 and 8 with a time randomly selected from the reported school trips.
  - Trips from school start between 1 pm and 4 pm with a time randomly selected from the reported school trips.
  - Modes are the children's reported school mode (for both to/from school)

Commented [AA2]: This will be improved in future surveys

  - Speeds are assumed based on the reported speeds for these modes
  - Distance is the straight-line distance between school and home multiplied by a 1.25 circuity factor
  - Duration is simply distance divided by speed
  - Arrive time is the departure time plus the duration of the synthetic trip
- Insert the synthesized school trips into the rest of the trip data and clean up relevant fields (origin purpose and category, dwell time, and other fields that depend on surrounding trips).

This process resulted in an increase of child-days overall receiving a school trip, with some variation across household participation groups and decreased the number of children days with other types of travel during school hours. Table 18 provides the attendance rate pre- and post-imputation for the different participation methods. As mentioned previously, a number of children have missing school locations for those parents who recruited via the smartphone application. While the 70% post-imputation attendance rate is outside the target range, further improving this rate would involve synthesizing trips on school days were other trips exist. This would likely cause issues with tour formation and so has been avoided. The next section discusses improvements in data collection to avoid this issue going forward.

TABLE 18: CHILD-DAYS WITH A SCHOOL TRIP

RECRUIT METHOD	DIARY METHOD	PRE-IMPUTATION	POST-IMPUTATION
Online	App	36%	70%
Online	Online	86%	86%
App	App	34%	41%

6.2 FUTURE WORK TO ACCURATELY COLLECT CHILD TRIP DATA

In future survey years, RSG will consider adding questions to make trip imputation for children easier while perhaps reducing the amount of proxy reporting detail that adults need to provide for their children’s travel, which seems to be an issue of growing sensitivity with regard to data privacy, as well as a significant response burden for parents. One example would be to collect school location for all children as part of the recruit survey so it exists for all children. Another example of a possible change would be to only ask about children’s trips on one day instead of the full 7 days thereby reducing the burden on parents.

7.0 APPENDIX A: DETAILED DESCRIPTION OF PURPOSE IMPUTATION

The columns in Table 3-3 are based on the mismatch between the trip destination location type and the reported destination purpose category. Location types[9] are home, primary work, school, and other, and the reported destination purpose categories are home, primary work, school, other-non-missing, and missing. The eight columns in the table are:

- **No mismatch**: The destination location type (home, school, work, other) matches the reported destination purpose category.
- **Location = Home / Purpose = Not Home**: The destination location type is home, but the reported destination purpose category is not home and is not missing.
- **Location = Work / Purpose = Not Work**: The destination location type is primary work, but the reported destination purpose category is not primary work and is not missing.
- **Location = School / Purpose = Not School**: The destination location type is school, but the reported destination purpose category is not school and is not missing.
- **Location = Not Home / Purpose = Home**: The destination location type is not home, but the reported destination purpose category is home.
- **Location = Not Work / Purpose = Work**: The destination location type is not primary work, but the reported destination purpose category is primary work.
- **Location = Not School / Purpose = School**: The destination location type is not school, but the reported destination purpose category is school.
- **Purpose = Missing**: The reported destination purpose is missing (either an incomplete trips survey, or a reported purpose of “other” that did not have a text response that could be recoded to a more detailed purpose).

7.1 RULES APPLIED TO IMPUTE TRIP DESTINATION PURPOSE

A series of sequential rules (described below) were applied to the trips, with different rules applying to the different “mismatch” categories described above. In many cases, the rules also depend on other trips made during the person-day. The description of some rules refers to “iterations”. The set of rules was run over the dataset five times because changing the purpose in response to one rule can have cascading effects for other rules that compare against the previous and next trips. Running several iterations allows any cascading effects to be caught and evaluated for consistency. **The rows in** Table 3-3 **correspond to the following rules.**

[9] The location type variable is based on the destination location being within 100 meters of the primary home, work, or school location, respectively. Additionally, the dwell time factors into it the assignment of location type.

Rule 1

Rule 1: OK. Use as is: Generally, the imputed purpose for these trips is the same as the reported purpose, although it may be changed by one of the subsequent tests (Rule 12, below).

Rules 2–19

Rules 2–19 below apply to mismatches where Location = X / Purpose = Not X trips, where X = Home, Work or School. The rules are applied sequentially for X = Home, then X = Work, then X = School.

Rule 2: "change mode" purpose is ok: If the reported purpose is "change mode", but the location is one of Home, Work or School, check for reasonableness. A valid change mode stop satisfies one of the following two criteria:

- The location type for the previous destination (or the origin if it is the first trip of the day) is X, the mode is Walk or Bike, and the mode for the next trip is Transit, or
- The location type for the next destination is X, the mode is Transit, and the mode for the next trip is Walk or Bike.

If the trip passes one of these criteria, the imputed purpose is set to "Change mode" and the imputed location type to "Other".

Rule 3: Change purpose to match location type: These mismatches tend to be common, mainly for X=Home, which tend to be cases of people reporting what they did at home rather than stating "Home" as a purpose. (Changes are planned for future releases of rMove to prevent this issue.)

The criteria for applying this rule are that there is no location type/purpose mismatch for either the preceding trip or the next trip, and neither the previous destination nor the next destination is at location type X.

If the trip passes these criteria, the imputed purpose is set to match the location, whether Home, Work, or School.

Rule 4: for last trip of day, change trip purpose to home: This is a variation of Rule 3, applied only to X=Home trips that are the last trip of the day. These mismatches tend to be almost as common as those described in Rule 3, and again seem to be mostly cases of people reporting what they did when they got home rather than stating "Home" as a purpose. The criteria are:

- It is the last destination of the person-day,
- There is no location type/purpose mismatch for the preceding trip and the previous trip is not at location type Home.

If the trip passes these criteria, the imputed purpose is set to Home

Rule 5: Change purpose to match location type, modified for missing purposes: This is a variation of Rule 3 that relaxes the criteria for adjacent trips to allow missing destination purposes instead of considering them mismatches. Compared to Rule 3, fewer cases meet these criteria. The criterion is:

- For both the preceding and next trip there is no location type/purpose mismatch or else the purpose is missing, and neither the previous nor the next trip is at location type X.

If it passes this test, the imputed purpose is set to X.

Rule 6: For last trip of day, change purpose to home, modified for missing purposes: This is a variation of Rule 4, applied only to X=Home trips that are the last trip of the day, that does not count a preceding trip with missing destination purpose as a mismatch. The criteria are:

- It is the last destination of the person-day,
- There is no location type/purpose mismatch for the preceding destination, or the purpose for the preceding trip is missing, and the previous trip is not at location type Home.

If it passes these criteria, the imputed purpose is set to Home.

Rule 7: Keep purpose as is, set location to Other, if previous trip was to X: This rule is to account for the fact that there are several legitimate activity purposes that that could conceivably be performed very near (within 100 m) the home or work location (e.g. shop, meal, social/recreation, errand, escort, work-related) and also avoids assigning the same purpose to two adjacent trips. The criteria are:

- It is not the first destination of the day,
- The destination location of the previous trip is also X with no purpose/location type mismatch,
- The dwell time at the destination is not greater than 60 min, the dwell time at the previous destination is at least 90 min, and the dwell time at the previous destination is at least 3 times as long as the dwell time at the current destination. The idea is that people will tend to spend a long time at Home, Work, or School, but shorter times for other purposes.

If it passes these criteria, the reported purpose is retained as the imputed purpose and the location type is set to Other.

Rule 8: Keep purpose as is, set location to Other, if next trip was to X: This test is analogous to Rule 7 but comparing to the *next* destination instead of the previous one.

The criteria are:

- It is not the last destination of the day,
- The destination location of the next trip is also X with no purpose/location type mismatch,
- The dwell time at the destination is not greater than 60 min, the dwell time at the next destination is at least 90 min, and the dwell time at the next destination is at least 3 times as long as the dwell time at the current destination.

If it passes these criteria, the reported purpose is retained as the imputed purpose and the location type is set to Other.

Rule 9: Previous and next trips are opposite mismatch - swap purposes between adjacent trips: This is the potential case where a respondent mixes up trip ends and, e.g. reports a Home-Shop-Home tour as a Shop-Home-Shop tour. The criteria are:

- It is not the first or last destination of the day,
- Both the previous destination and the next destination have the opposite mismatch,
- If there are destinations in the day that are two earlier or two later than the current one, then those do not have a purpose/location mismatch.

If it passes these criteria, then the imputed purposes for the previous and next destinations are set to the reported purpose of the current destination, and the imputed purpose for the current destination is set to the reported purpose for the previous destination.

Rule 10: Previous trip is opposite mismatch - swap purposes between adjacent trips: This is a simpler version of the previous test where only two consecutive destinations have "swapped" purposes. The criteria are:

- It is not the first trip of the day,
- The previous destination has the opposite mismatch (reported purpose is X, but location type is not X),
- If it is at least the third destination of the day, there is not a purpose/location mismatch for the destination before the previous destination, and
- It is either the last destination of the day or the next destination has no purpose/location mismatch. (In other words, the surrounding destinations in the day, if there are any, have no purpose/location type mismatch.)

If these criteria are met, the imputed purpose for the previous destination is set to the reported purpose of the current destination, and the imputed purpose for the current destination is set to the reported purpose for the previous destination.

Rule 11: Next trip is opposite mismatch - swap purposes between adjacent trips: This is the same as Rule 10 but comparing to the next destination instead of the previous. The criteria are:

- It is not the last destination of the day,
- The next destination has the opposite mismatch (reported purpose is X, but location type is not X),
- If there are at least two more destinations in the day, then there is not a purpose/location mismatch for the destination after the next destination, and
- It is either the first destination of the day or the previous destination has no purpose/location mismatch. In other words, the surrounding destinations in the day, if there are any, have no purpose/location type mismatch.

If these criteria are met, the imputed purpose for the next destination is set to the reported purpose of the current destination, and the imputed purpose for the current destination is set to the reported purpose for the next destination.

Rule 12: Purpose to change to match location type- all subsequent purposes in day shifted down one trip: This test looks for situations where inserting a destination purpose and shifting all the purposes for subsequent destinations in the day to the next trip cause no mismatches. The criteria are:

- After the current location X / purpose Other mismatch trip, shifting the destination purposes for all subsequent destinations in the day to the next destination does not cause any purpose/location type mismatches for any of the subsequent trips (a “Missing” purpose is not treated as a mismatch)

If the test is passed, set the imputed purpose for the current destination to X and set the imputed purpose for each subsequent destination to the reported purpose for the preceding destination. This test was revised a number of times to make sure that a last trip of the day at the home location would be retained, and that certain types of purpose shifts were given lower priority—meaning that they were only changed on later iterations of these rules if other rules had not fixed the problem first. Note that this is one of the only tests that can change the imputed purpose for an “OK” trip that had no purpose/location type mismatch. Without having point of interest land use data to relate to the geo-locations, it is impossible to know if this recoding is correct, but in any event, it does not create a mismatch in type and purpose.

Rule 13: last destination of day at/near Home location with stay > 3hr- change purpose to home: After the preceding tests, there were still many cases remaining where the last destination of the day was an overnight dwell at/near the home location, but the reported purpose was not home. The criteria are:

- The location type is Home,
- It is the last destination of the person-day, and
- The dwell time as at least 3 hours.

If it passes the test, change the imputed purpose to Home.

Rule 14: One or more habitual locations are near each other; keep reported purpose: There are some cases in the data where the primary home, work, and/or school locations can be the same, with the most common being people who primarily work from home (or within 100 m of home). In those cases, the coded destination location type is ambiguous. The criterion for this test is:

- The reported purpose is home and the distance to home is less than 200 m, OR
- The reported purpose is primary work and the distance to the primary workplace is less than 200 m, OR
- The reported purpose is school and the distance to the school location is less than 200 m.

If it passes this test, there is assumed to be no purpose/location type mismatch, so set the imputed purpose to the reported purpose.

Rule 15: Change "work-related" purpose to "work" if at primary work location with no adjacent work purpose: This rule resolves mismatches between Work location and "work-related" purpose while avoiding the creation of a work-work trip.

The criteria are:

- The location type is Work,
- The reported purpose is "work-related", and
- Neither the preceding nor the following destination (if they exist) have the reported purpose of primary work.

If it passes, the imputed purpose is set to Work.

Rule 16: Change purpose to match location type if it doesn't create consecutive trips with same purpose: The last test for the Location X / Purpose Not-X trips accepts the location type as the best guess for the purpose, as long as it won't create a Home-Home, Work-Work or School-School trip. The criterion is:

- Neither the preceding destination nor the next destination in the day, if they exist, have the reported purpose of X.

If this test is passed, set the imputed purpose to X.

Code 19: Rules 2–16 not applied, keep reported purpose and examine further: If none of the tests above are passed, then the imputed purpose is kept as the reported purpose.

The "examine further" text is added to the label to indicate that these would be good cases for agencies to examine visually, one-by-one, to see if any particular imputed purpose would make sense.

Rules 20–37

Rules 20-37 below apply to mismatches where Location = Not-X / Purpose = X trips, where X = Home, Work or School. The rules are applied sequentially for X = Home, then X = Work, then X = School.

Rule 20: Impute Escort purpose: This test checks for likely "escort" trips dropping off or picking someone up from another person's home, work, or school. In these cases, it is likely to be someone else's home, work, or school location, not the respondent's. The criteria are:

- The destination is not the last one of the day,
- The mode for the trip is the same as the mode for the next trip,
- The number of travelers for the trip is different from the number of travelers for the next trip, and
- The duration of stay at the location is 30 minutes or less.

If these criteria are met, the imputed purpose is "pick up/drop off passenger".

Rule 21: Where purpose is go to X, location is not X but within 200 m of X, retain reported purpose: Given that the GPS locations are not always precise, and that respondents may delete very short spurious trips near their destination rather than merging them, the distance limit of 100 meters to impute location type may be too strict in some cases. This test relaxes that limit when the location is very near a habitual location, but not within the 100 m limit. The criteria are:

- The reported destination purpose is home, primary work or school, and
- The corresponding home, primary work or school location, respectively, is within 200 m of the destination coordinates,
- The destination location type for the preceding trip and the next trip are different from that of the current trip (so as not to create a home-home, work-work or school-school trip).

If it passes the criteria, the reported purpose is retained as the imputed purpose and the location type is set to X. Some work and school locations may cover an area larger than 100 m in themselves, and some people may make quick stops close to work that they delete.

Rule 22: Where purpose is go to X, location is not X but within 300 m of X, retain reported purpose: This is the same as Rule 21 above, but with the allowable distance extended to be in the range between 200 and 300 meters.

Rule 23: Where purpose is go to X, location is not X but within 500 m of X, retain reported purpose: This is the same as Rule 21 above, but with the allowable distance extended to be in the range between 300 and 500 meters. The prior recategorization of some work and school trips work-related and school-related trips has pre-empted this check in some cases. This means that for very large school or work campuses, some valid work or school cases may be categorized as work-related or school-related trips.

Rule 24: Overnight stay away from home - out of region: Many overnight stays are more than 500 m from the respondent's home location. The criteria are:

- The destination is the last of the respondent's day,
- With a dwell time of at least 3 hours, and
- A distance of more than 500 m from the home location, and
- The location is outside the study region.

If it passes this test, assign a new code for the imputed purpose "overnight at a location outside the region", and a new code for the imputed purpose category "other person's home". The location type is set to Other.

Rule 25: Overnight stay away from home - within the region: This is the same test as the preceding, but with the overnight location within the region.

If it passes this test, assign a new code for the imputed purpose "overnight at a non-home location within the region". The location type is set to Other

Rules 31–36

Rules 31–36: Imputation based on reported purposes of other trips to same location in the survey data: If a destination purpose for a trip of type Location=Not X, Purpose=X could not be imputed based on multiple iterations of the rules above, or if the destination purpose is missing or “other”, then a procedure was used based on other trips in the data. This procedure was applied only after multiple iterations of the rules above and works by locating other trips to locations nearby and using their reported purposes.

The following rules all use purposes determined from the trips meeting the following criteria:

- The distance from the current destination is less than D meters, where D can be 50, 100, or 200 meters,
- The reported destination purpose is not missing, is not “change mode”, and
- Is different from the reported purpose for the current destination (since that purpose is a mismatch with the location type)

Rule 31: Assign purpose from respondent’s trips within 50 m: If there are multiple trips for the respondent that meet the preceding criteria, then one trip is chosen at random using a Monte Carlo method. (The random number used is written to the trip file.). The imputed purpose is set to the reported purpose of the randomly selected trip. The random selection has the property that the distribution of imputed purposes at the location will tend towards the distribution of observed eligible purposes at that location. The imputed location type is set to Other.

Rule 32: Assign purpose from respondent’s trips within 100 m: If none of the respondent’s destinations meet the criteria above, then eligible trips are chosen from an expanded search radius from 50 m to 100 m.

Rule 33: Assign purpose from respondent’s trips within 200 m: If none of the respondent’s destinations meet the criteria above, then eligible trips are chosen from an expanded search radius from 100 m to 200 m.

Rule 34: Assign purpose from other respondent’s trips within 50 m: If no qualifying trips can be found from the respondent’s own trips, then a similar search is done across trips made by other respondents in the survey sample. The criteria are very similar to those described above for the respondent’s trips, with some variations:

- The distance from the current destination is less than D meters, where D can be 50, 100, or 200 meters,
- The reported destination purpose is not missing, is not “change mode”, and
- Trips with purpose = Work are not allowed unless the respondent has reported work trips of their own during the survey period,
- Trips with purpose = school are not allowed unless the person has reported school trips of their own during the survey period.

If there are multiple trips in the dataset that meet these criteria, then one is chosen at random using a Monte Carlo method The imputed purpose is set to the reported purpose of the randomly selected trip and the imputed location type is set to Other.

Rule 35: Assign purpose from other respondent's trips within 100 m: If none of the other respondents' destinations meet the criteria above, then the search radius is expanded from 50 m to 100 m.

Rule 36: Assign purpose from other respondent's trips within 200 m: Finally, if none of the other respondents' destinations meet the criteria above, then the search radius is expanded once again from 100 m to 200 m.

Rules 37–39

Rule 37: Work purpose not at primary work location, changed to work-related: This test is essentially redundant now that some work purposes have been recategorized to work-related based on location prior to purpose imputation.

Rule 38: School purpose not at school location, changed to school-related: This test is also redundant now that some school trips have been recategorized as school-related based on location prior to purpose imputation.

Rule 39: No other visits to location - impute purpose to Other (no evidence): The final imputation code is for the Location=Not X, Purpose=X trips that do not pass any of the tests above.

7.2 IMPUTING ORIGIN PURPOSES

Trip origin purpose is imputed based on the imputed destination purpose of the preceding trip where possible.

Rule 1: Copied from preceding destination in same day: For the second and subsequent destinations visited in each person-day, the imputed origin purpose is set equal to the imputed destination purpose for the preceding trip.

Rule 2: Copied from last destination in preceding day: If the preceding day has valid trips, the imputed origin purpose for the first trip of the day is set to the imputed destination purpose of the last destination from the preceding day.

Rules 3-6: First trip -- Imputed based on origin location type: If it is the first day the person participated in the survey, or if there were no valid trips on the preceding day, the imputed origin purpose for the first trip is based on the origin location type (**o_location_type**) for that trip, determined according to distance from reported habitual locations. The codes are 3 for Home, 4 for Work, 5 for School and 6 for Other.

RSG
the science of insight

55 Railroad Row
White River Junction, VT 05001
802.295.4999
www.rsginc.com

White River Junction & Burlington, VT | Arlington, VA | Chicago, IL | Evansville, IN | Portland, OR | Salt Lake City, UT | San Diego, CA