|  |  |
| --- | --- |
| **TO:** | TransLink Modelling Staff |
|  | |
| **FROM:** | David Hensle |
|  | |
| **CC:** | Joel Freedman, Ben Stabler |
|  | |
| **DATE:** | October 28, 2021 |
|  | |
| **SUBJECT:** | Survey Processing for SPA Input |
|  |  |

## Introduction

The purpose of this memo is to provide guidance to the TransLink staff on how to prepare their raw household travel survey (HTS) to be compatible with RSG’s Survey Processing Application (SPA).

Background information on the SPA tool and what it does will be presented first followed by specific data formatting and processing steps for TransLink staff on getting their raw survey data processed for SPA input.

## Survey Processing Application

The SPA tool processes the raw survey files into data structures consistent with activity-based modeling frameworks such as ActivitySim[[1]](#footnote-1). Input to the SPA tool consists of household, person, and place files, and the SPA tool’s primary function is to group trips for each person into tours and determine tour level variables in accordance with the region’s specific activity-based model design. Each of these input files need specific variables names to match the variable definitions in the SPA tool.

The SPA tool first computes person types according to Table 1. It also checks for and resolves any inconsistencies between the reported purpose and location for work and work-related activities. Then the process loops through place records to identify tours and linked trips.

Table 1: Person Type definitions

|  |  |  |  |
| --- | --- | --- | --- |
| Person Type | Age | Employment Category | Student Category |
| 1: Full-time worker | >=16 | Full-time (1) | Any |
| 2. Part-time workers | >=16 | Part-time (2) | Not attending (3) |
| 3: University student | >=17 | Not full-time (2,3) | College+ (2) |
| 4: Non-worker | >=16 and <=64 | No (3) | No (3) |
| 5: Retired | >=65 | No (3) | No (3) |
| 6: Student of driving age | >=16 and <=19 | Not full-time (2,3) | K-12 (1) |
| 7: Student of non-driving age | >=6 and <=15 | Not in labor force (4) | K-12 (1) |
| 8: Child too young for school | >=0 and <=5 | Not in labor force (4) | K-12, Not attending (1,3) |

A home-based tour is created for each sequence of place records starting and ending with “Home” as the activity purpose[[2]](#footnote-2). Work-based tours are created for each sequence of place records starting and ending with "Work" as the activity purpose. Within each tour, a trip is created for each pair of place records corresponding to an origin and a destination. A "trip linking" procedure links change mode activities with previous and subsequent places. Travelers typically report changing mode at transit stops (from walk or drive to transit, or vice-versa). For example, an individual that drives their car to a parking lot where they catch a bus to work has made two trips – the drive from home to the parking lot, and the bus from the parking lot to work. These two trips are combined to a single linked bus trip with park-and-ride access.

Once all place records are processed and corresponding trips and tours are created, the SPA checks for tours that start before or end after the designated survey period (3am to 3am). Such tours are tagged as partial tours because some of their attributes either cannot be determined or need to be computed in a manner different from non-partial tours. Additional attributes are then derived from their constituting trips.

To determine the tour purpose, a scoring system is used based on the activities within the tour, the duration of each activity, and distance of each activity from home. Table 2 shows the scores that are applied to each trip purpose by duration. The final score is the sum of the score from the table plus an additional distance term

where the distance from home is given in miles. The purpose of the trip with the lowest score is selected as the tour purpose. The idea behind this logic is to choose the purpose of the tour based on a hierarchy of activities, with mandatory (work and school) activities at the top of the hierarchy, followed by maintenance (escorting, shopping, and other maintenance) and finally by discretionary (eating out. social/recreation, and other discretionary) activities. However, the procedure recognizes that some travelers might construct their activity schedules according to their own hierarchy of needs, so we provide fuzzy logic rules that consider activity duration and distance as additional purpose ranking variables, assuming that longer and more distant activities are more important than shorter and closer activities.

Table 2: Tour Purpose Scoring

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Purpose | | Score Function by Trip Distance | | | | | | | | |
| Code | Description | <2 mi | 2 mi | 4 mi | 6 mi | 8 mi | 10 mi | 12 mi | 14 mi | 16+ mi |
| 1-3, 10 | Work/ University/ School/ Work-related | 8 | 4 | 2 | 1.5 | 1.4 | 1.3 | 1.2 | 1.1 | 1 |
| 4 | Escorting | 8 | 6 | 4 | 3.5 | 3.4 | 3.3 | 3.2 | 3.1 | 3 |
| 5-6 | Shopping/ Maintenance/ Discretionary | 10 | 6 | 4 | 3.5 | 3.4 | 3.3 | 3.2 | 3.1 | 3 |
| 7 | Eat Out | 12 | 7 | 5 | 4.5 | 4.4 | 4.3 | 4.2 | 4.1 | 4 |
| 8 | Social/Visit | 14 | 8 | 6 | 5.5 | 5.4 | 5.3 | 5.2 | 5.1 | 5 |

Next, the SPA identifies joint travel episodes across household members. Due to the inconsistency often found in the data reported by different travel participants of a joint tour, rule-based intelligence is implemented in the SPA to identify and remove errors. The SPA further derives escort-related attributes (e.g., who is the chauffeur and whether the escorted individuals are being dropped off or picked up) for each travel group and tags the information to the associated trips.

Lastly, tours of different household members that contain identical series of joint travel are tagged as fully joint tours. Tours that contain both joint and non-joint trips are referred to as partially joint tours. As the last computational step of the SPA, escort related attributes at the trip level are aggregated to the tour level for partially joint tours.

Table 3 lists the six data tables and two log files produced by the SPA tool. Output formats most closely follow the output format of CT-RAMP (which the SPA tool was originally developed for), but ActivitySim outputs contain all the same information just formatted a bit differently.

Table 3: SPA Outputs

|  |  |
| --- | --- |
| File name | Description |
| Households.csv | Each row contains selected attributes of an observed household including income, workers, autos, size, children, and weight. |
| Persons.csv | Each row contains selected attributes of an observed person including age, gender, person type, employment status, student status, work and school location, and person weight. |
| Tours.csv | Each row contains attributes derived for an observed person tour, including tour purpose, mode, participants, origin and destination locations and times, and transit information if applicable. |
| Trips.csv | Each row contains attributes derived for an observed person linked trip. |
| Unique\_joint\_tours.csv | Each row contains attributes describing a fully joint tour made by multiple household members. |
| Unique\_joint\_trips.csv | Each row contains attributes describing a joint trip (unlinked) made by multiple household members. |
| Joint\_ultrips.csv | Optional output used for analyzing data errors relating to joint travel episodes. Each row contains selected attributes of an unlinked person trip that is part of a joint trip. |
| Error\_log.txt | Text file containing error and warning messages about unresolved data anomalies. |
| Recode\_log.txt | Text file describing assumptions and recoding made to resolve known data anomalies and inconsistencies. |

There are some input attributes, like person and household weights, that are not written to the SPA output but may be required for later analyses. Since the SPA tool preserves person, household, and place ID’s, these attributes can be joined later from the SPA input household, person, and place files.

Survey data processed with the SPA tool serves as the primary data source for comparing model outputs to the survey and creating model calibration targets.

## SPA Inputs

The SPA tool requires three input tables: households, persons, and places. This section will review the most important columns in each table and their values. Each of the fields in these tables should be able to be constructed from the TransLink survey data based on a review of the survey code book. Additional thoughts on the proposed trip and tour modes and purposes will also be included.

### Households

Table 4 shows the household level information for SPA input and specifies the column name, data type, description of that column, and additional notes about the values in each of the columns. If there are no specific value instructions, that field has been left blank. Each row of the household table corresponds to a single household.

Table 4: SPA Household Input Table FOrmat

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Field Name | | Type | Description | Values |
| SAMPN | INT | | Unique sample number / household ID | 1 to num HHs |
| HH\_ZONE\_ID | INT | | TAZ of household |  |
| HH\_WEIGHT | REAL | | Weight to expand data at household level |  |
| HH\_SIZE | INT | | Number of people in the household |  |
| HH\_VEH | INT | | Number of vehicles in the household |  |
| HH\_WORKERS | INT | | Number of workers in the household |  |

The only household variables that are absolutely required for the SPA tool is the SAMPN and HH\_ZONE\_ID. Other variables are used when creating summaries to compare to model outputs and including them in this file allows for downstream analysis to be set up faster. There are also no restrictions on including additional columns in the input tables that track additional information from the survey.

### Persons

Table 5 shows the person level information for SPA input. As with the household table, each row contains information for a single person.

Table 5: SPA Person Input Table Format

|  |  |  |  |
| --- | --- | --- | --- |
| Field Name | Type | Description | Values |
| SAMPN | INT | Household ID | 1…num HHs |
| PERNO | INT | Person ID | 1…num PERs |
| STAZ | INT | School TAZ | -9 if not a student |
| WTAZ | INT | Work TAZ | -9 if not a worker |
| PER\_WEIGHT | REAL | Weight to expand data at person level |  |
| STUDE | INT | Student Category | 1: School 2: University 3: Non-student |
| SCHOL | INT | Level of School | 1: Daycare 2: Pre-school 3: K-8 4: 9-12 5: Vocational / Tech School 6: Community College 7: 4-year College or University 8: Graduate / Professional School 0: Not a student |
| EMPLY | INT | Employment Category | 1: Full-time 2: Part-time 3: Unemployed 4: Non-worker |
| AGE | INT | Age of person | 1…99 |
| EMPLY\_LOC\_TYPE | INT | Employment location type | 1: fixed  2: works from home  3: varies 0: Not a worker |

Using student status, employment status, and age from the person table, the SPA tool calculates person type as determined by Table 1. For most person types, all of these variables are required. As such, it is suggested that some assumptions be made to impute missing data for these categories.

A good estimate for missing age is student or employment status. Exact ages are not required, but just need to fall into the appropriate categories used for determining person type. For example, if the survey has the person taking a work trip and no school trips, any value of adult below the retirement age can be used.

The existence of work or school trips can be used to inform missing employment or student status. If a person states they are a worker or student, but do not supply a work or school TAZ, the missing values can be imputed by using the location of the trip destinations for a reported work or school trip.

Ideally the level of school is known to the fidelity listed in Table 5, but often this is not the case. This field is usually used to help determine person status when age or other variables may be missing or inconsistent. As such, the next best scenario would be splitting the level of school into primary, secondary, and college.

### Places

Table 6 shows the trip level information for SPA input. Each row in the places table consists of a place that each person went to during the day. If a person starts their day at home, that home location should be listed at the start of the person’s day. (For this reason, the verbiage “places” is used instead of “trips”, but they are often used interchangeably.)

Table 6: SPA Places Input Table Format

| Field Name | Type | Description | Values |
| --- | --- | --- | --- |
| SAMPN | INT | Household ID | 1…num HHs |
| PERNO | INT | Person ID | 1…num PERs |
| PLANO | INT | Place number | 1… # of Places For That Person |
| TRIP\_WEIGHT | REAL | Weight to expand data at trip level |  |
| TAZ | INT | Place TAZ |  |
| TPURP | INT | Place purpose – what are they doing at this location? | 0: Home 1: Work 2: University 3: School 4: Escorting 5: Shopping 6: Maintenance 7: Eat Out 8: Social / Visit 9: Discretionary 10: Work Related 11: Loop (Home -> Home) 12: Change Mode 13: Other |
| MODE | INT | Surveyed Trip Mode – how did they get to this place? | 1: SOV 2: HOV2 3: HOV3+ 4: Walk 5: Bike 6: TNC / Taxi 7: Transit-Bus 8: Transit-Rail 9: Transit-WCE 10: School Bus 11: Other  0: NA |
| DEP\_HR | INT | Place Departure Hour | 0…23 |
| DEP\_MIN | INT | Place Departure Minute of Hour | 0…59 |
| ARR\_HR | INT | Place Arrival Hour | 0…23 |
| ARR\_MIN | INT | Place Arrival Minute of Hour | 0…59 |
| DRIVER | INT | Flag to indicate whether the person was driving | 1: Driver 2: Passenger / NA |
| TOLL\_NO | INT | Toll paid on Trip? | 1: Yes 2: No |
| TOTTR | INT | Total travelers on Trip |  |
| TOTTR\_NEXT | INT | Total travelers on subsequent trip | 0 if last place of the day |
| HHMEM | INT | Number of household members on the trip |  |
| PER1 | INT | Person ID of first additional household member on trip |  |
| PER2 | INT | Person ID of second additional household member on trip |  |
| PER3 | INT | Person ID of third additional household member on trip |  |
| PER4 | INT | Person ID of fourth additional household member on trip |  |
| PER5 | INT | Person ID of fifth additional household member on trip |  |
| PER6 | INT | Person ID of sixth additional household member on trip |  |

If the raw survey data lists actual trips instead of places, initial starting places need to be inserted into the table for each person at the start of their day. These “starting place” entries should have a MODE of 0, and a purpose to denote where they were at the start of the day, for most people this means a TPURP of 0 for home. Of course, people do not always start the day at home and as such the starting purpose should reflect that.

For the first place of the day, the place arrival hour should be the time that corresponds to a new model day. In most ActivitySim implementations, this is set to 3:00 AM. The place departure time then corresponds to the time that the person left their starting location for the first trip of the day.

Trip purposes in Table 6 follow most ActivitySim implementations with a few additional categories that fall outside the ActivitySim framework, namely loop, change mode, and other. These additional purposes are typically kept to help understand what travel might be missing from later model implementation. Trip purposes and modes can be changed to different segmentations as desired.

The SPA tool distinguishes people on joint tours as whether they are members of the same household. In ActivitySim, joint tours are defined as between multiple members of the same household. For this reason, data for each household member on the joint tour is required as well as the total number of people (household members + non-members + current person) on the trip. Listing the person IDs of the household members allows for person attributes to be joined for joint tour composition summaries.

The input place file should be sorted by person ID and place number where the place number restarts at one for each person and counts places attended by that person chronologically.

### Trip and Tour Mode Determination

Trip modes that are output from the SPA tool are linked trips compared to the unlinked trips listed in Table 6. Tour mode is the selected as the linked trip on that tour that is the highest in the hierarchy. A proposed tour mode hierarchy for TransLink, from top to bottom, is: School Bus, PNR-WCE, PNR-Rail, PNR-Bus, Walk-WCE, Walk-Rail, Walk-Bus, TNC / Taxi, HOV3, HOV2, SOV, Bike, Walk, and Other.

While linked trip and tour modes are not important for SPA input data preparation, having the hierarchy determined before SPA work begins can make that process faster and help decide what data needs to be extracted from the survey.

### Handling Multiple Days

The SPA tool outputs follow ActivitySim in modeling results for a single day. However, survey data can be gathered over multiple days, so some additional processing is required to handle this case.

For the TransLink survey data, the easiest way to handle the households and persons that have multiple surveyed days would be to duplicate those household and person records in the person and households table, assign new sample numbers (while keeping track of their surveyed household id), and divide the person and household weights by the number of duplicated rows. The trip records for each day would then be matched to one of the duplicated persons and households.

After the duplication, a check on the data summarizing the total weighted number of households and persons should be the same as before duplication and each new household and person should only have a single day of trip records. Trip weights would be updated to reflect the total number of trips in a single day in the region if not already weighted as such.

Duplicating rows like this prevents the need for additional input files and therefore any additional SPA processing and post-processing steps.

## Summary

The Survey Processing Application (SPA) is responsible for transforming survey data into a format comparable to activity-based modeling outputs. This includes the classification of person type, the grouping of trips into tours, and the determination of tour mode and purpose.

Input tables required for the SPA tool are listed and described. The household table contains a row for each household and includes the home zone ID and a unique home identifier. Each row of the persons table corresponds to a single person and lists attributes about that person’s employment, student status, and age, among other things. The places table lists a single row for each place a person visited in a survey day, how and when they got from one place to another, and the activity of each place. Additional discussion for trip and tour mode determination as well as the handling of data over multiple days was included.

Once the data preparation described in this memo has been completed, the SPA tool can be run and modified for the specific modes and purposes for the implementation region, which is the TransLink modeling region in this memo.

1. For more information on the variables and structures (tours, intermediate stops, etc.) used in ActivitySim, see RSG. *Phase I Activity-Based Model Design Technical Memorandum*, for SEMCOG, March 23, 2020. [↑](#footnote-ref-1)
2. Out of nearly 12,400 households in the SEMCOG 2015 household travel survey, only 165 tours did not start or end at home. These were dropped from the data processing procedures. [↑](#footnote-ref-2)