Chapter 4

ACTIVITY GENERATOR

CONTENTS

Chapter Outline *1*
4.1. Overview *3*
4.2. Terminology *4*
4.3. Key Concepts *5*
4.3.1 Classification and Regression Trees **5**
4.3.2 Discrete Destination Zone Choice Model **7**
4.4. Major Data Inputs *7*
4.4.1 Household Activity Survey **9**
4.4.2 Survey Household Demographic Data **9**
4.4.3 Synthetic Households **10**
4.4.4 Land-Use Data **10**
4.4.5 TRANSIMS Network Data **11**
4.5. Major Data Outputs *12*
4.6. Module Interfaces *15*
4.6.1 Input received from the Population Synthesizer Module **15**
4.6.2 Input received from the Selector Module **16**
4.6.3 Output sent to the Route Planner Module **17**
4.7 Configuration Files *17*
4.8 Algorithms *21*
4.8.1 Creating skeletal activity patterns from the survey households **21**
4.8.2 Using CART algorithm to build a classification tree based on household demographic data **23**
4.8.2.1 Tree-growing step 23
4.8.2.2 Pruning the tree 39
4.8.3 Matching the given synthetic household with a survey household **49**
4.8.3.1 Assigning the given synthetic household to a unique terminal node of the tree . 49
4.8.3.2 Selecting a survey household within the terminal node to match with the given synthetic household that is assigned to this node 50
4.8.3.3 Matching the assigned synthetic household members with the selected survey household members based on the four demographic variables: relate, work, gender, and age 59
4.8.4 Generating activity times and durations **63**
4.8.5 Generating non-home activity locations **69**
4.8.5.1 Selecting the zones 69
4.8.5.2 Choosing destinations within the selected zones 71

ACTIVITY GENERATOR

Chapter Outline

The population synthesizer module discussed the following issues:

- The generation of synthetic households from census data at the block group level or the census tract level.
- The development of associated demographic characteristics (e.g., age, gender, income, etc.) for each synthetic household.
- The placement of each synthetic household on a link in the transportation network.
- The assignment of vehicles to each household, including information regarding the vehicle emission type and the initial vehicle location on the network.

The Activity Generator develops a list of activities for each member of a synthetic household over a 24-hour period. Figure-4.1 details the inputs and outputs of this module, along with its interactions with the other relevant modules of TRANSIMS.

Each activity generated has the following features:

- An activity type (e.g., home or school, etc.) and its priority,
- Starting, ending, and duration time preferences,
- A preferred travel mode,
- A vehicle preference (if appropriate),
- A list of possible locations for an activity, and
- A list of other household participants (if the activity is shared).

The algorithms used to generate the list of activities for each member of a synthetic household involve the following five steps:

1. Create skeletal activity patterns from the survey households,
2. Use the CART (Classification and Regression Tree) algorithm to build a classification tree based on household demographic data,
3. Match the given synthetic household with a survey household,
4. Generate activity times and durations, and
5. Generate activity locations.

This Chapter describes the above steps, along with the salient concepts including illustrative examples. Prior to algorithm descriptions, the Chapter begins with an overview of the Activity Generator Module, definitions of various terminologies, key concepts, major data inputs and outputs, various module interfaces, and values for the relevant configuration files.

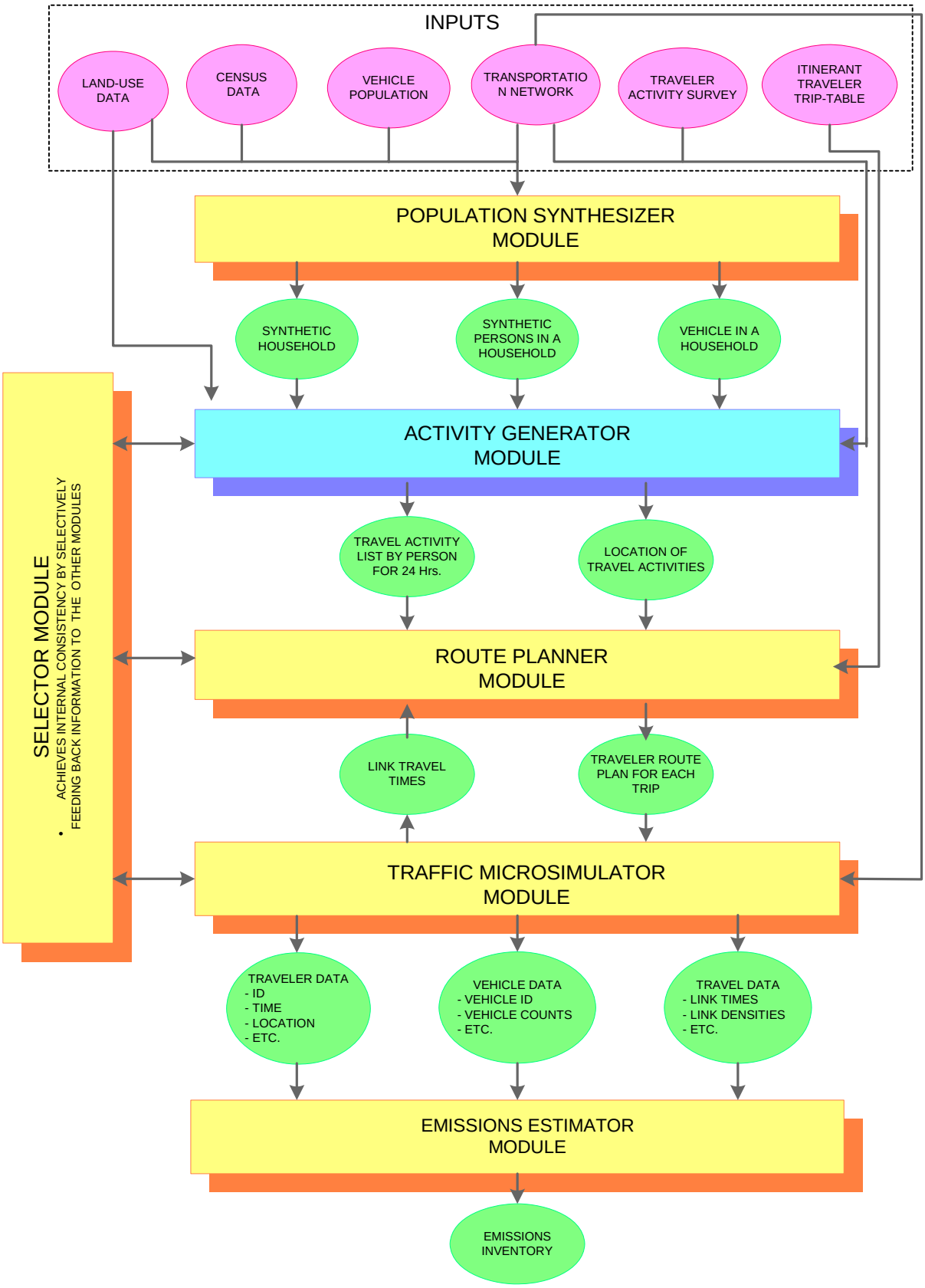


Figure-4.1: TRANSIMS framework

4.1. Overview

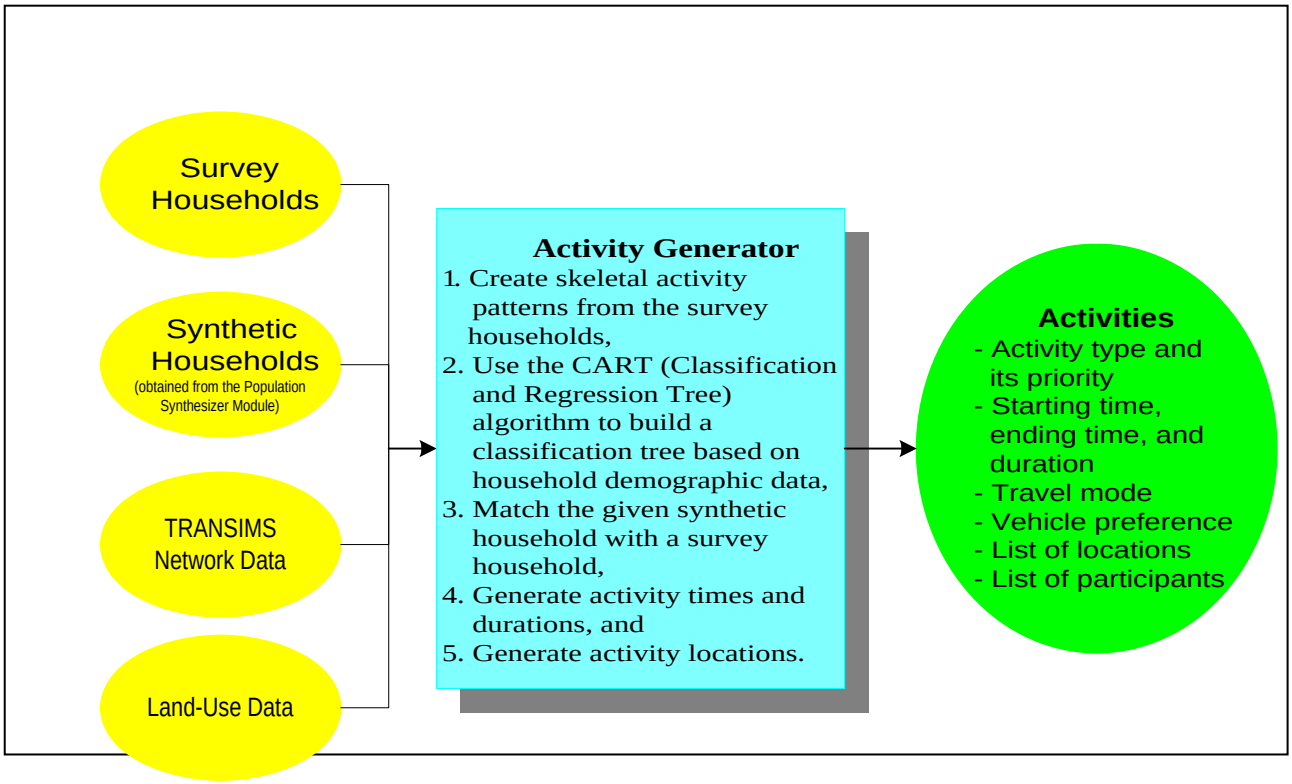


Figure-4.2: Data flow diagram and overall structure of the TRANSIMS Activity Generator Module

The Activity Generator Module, as shown in Figure-4.2, requires principal inputs from the synthetic households (obtained from the Population Synthesizer Module), the survey households, the TRANSIMS network data, and the land-use data. Using these inputs and the CART algorithm the activity generator builds trees, which are used for classifying and matching the households. The lists of these activities are then sent to the Route Planner Module to find the shortest paths for all the transportation-related activities.

The Activity Generator Module has two principal purposes:

- To capture household travel behavior accurately, aside from creating activity/travel patterns for individuals. For example, if one family member takes a child to school, another member need not do so. In other words, household interactions are accommodated within this module.
- To ensure relative simplicity in the modeling of activity location. The location choice models are fairly simple. Feedback from the Traffic Microsimulator and the Route Planner Modules (that are updated and corrected by the Selector Module) are used to refine activity location choices in the feedback iterative process.

4.2. Terminology

- **Survey households:** This is a set of households that have been surveyed to yield the input database for the Activity Generator Module. The data from the survey households include information related to *household demographic data* and *household activity data.*

- **Survey household demographic data:** The survey household demographic data contains information about the characteristics of each household member in the survey sample of households. This data is used for building and matching the demographic characteristics of the synthetic households.

- **Household activity survey:** The set of activities obtained from the survey household is a representative sample of the household's activities, including travel and event-participation information for each household member. Skeletal activity patterns are created by stripping locations from this database. That is, an activity from a survey household assigned to a synthetic household is stripped of its beginning and destination locations. The Activity Generator module determines the destination of this activity through a location-choice model.

- **Zone:** Each zone in the study region contains several activity locations, and is designated by the specification of a geo-coordinate system (*x-y* coordinate). Each zone contains data that captures its characteristics such as its total square footage of retail property, the total employment within the zone, etc. This data is aggregated with **the land use data**, and is used in the Activity Generator Module to select appropriate zones for each of the synthetic household member's activities. The zones could be as large as the usual Traffic Analysis Zones or as small as the individual activity locations.

- **Travel mode:** There are nine different modes available within TRANSIMS (as shown in Table-4.1).

Table-4.1: Currently Recognized Travel Mode Labels

Mode	Mode Label
Walk	*w*
Bike	*i*
Car	*c*
Bus	*b*
Light Rail	*l*
Regional Rail	*g*
Rapid Rail	*p*
Trolley	*y*
Other Transit	*t*

- **Mode string:** The mode string contains a list of admissible travel mode sequences that could be used in the order given along the path from the origin to the destination. For example, given a mode string "*wcw*" for a trip from home to the office, any admissible travel pattern would start via a "walk" mode from home to the "car," use the car as a transportation vehicle, and finally, "walk" to the office.

4.3. Key Concepts

4.3.1 Classification and Regression Trees

The Classification and Regression Tree algorithm (CART) is a tool used to group survey households into a manageable set of categories based on demographic variables, which produce the greatest variation in trip making activity between the categories. Its purpose is to produce an accurate classification of household demographic characteristics, as independent variables (X's, such as number of workers, household size, household income, etc.) based on households travel behaviors, and as dependent variables (Y's, such as times spent at work, at home, at shopping, etc.).

Each survey household is effectively classified by the CART algorithm into one of the tree's terminal nodes/classes representing the end path of selected household demographic characteristics (X's). This tree is chosen to be sensitive to the household behavior characteristics (Y's), but to be parsimonious with respect to household characteristics that do not significantly affect behavior. Tree structured classifiers, or, more correctly, *binary* tree structured classifiers, are constructed by repeated splits of the *active* node into two *subnodes* based on a **split criterion** and a **split value**. The two new subnodes become in turn active nodes. The splits are performed until all leaf nodes can be declared to be **terminal nodes**. The criteria used to mark nodes as *terminal nodes* are **the number of observations in the (children) nodes** ($N_{\min}$) and **the total deviance of the node** ($D(N)$).

The total deviance is computed from the combination of the statistical deviations regarding various household characteristics for the households in the corresponding node, such as the number of workers, the number of cars within the households, etc. The splits are performed until each leaf node is declared as a terminal node. A leaf node becomes a terminal node when the number of observations in at least one child-node is less than N_{min}, or the total deviance of this node is less than a specified value. Otherwise, the node is still marked as an active node and is selected to perform further splits as the algorithm proceeds.

To illustrate a binary-matching tree, consider the simple example given in Figure-4.3. Note that the actual trees used in the Activity Generator Module are more detailed. This tree is constructed based on dependent variables (such as: *time spent at home, time spent at work and time spent in visit*) and three household demographic independent variables (that are used as the split criteria): *number of workers, household size,* and *household income.*

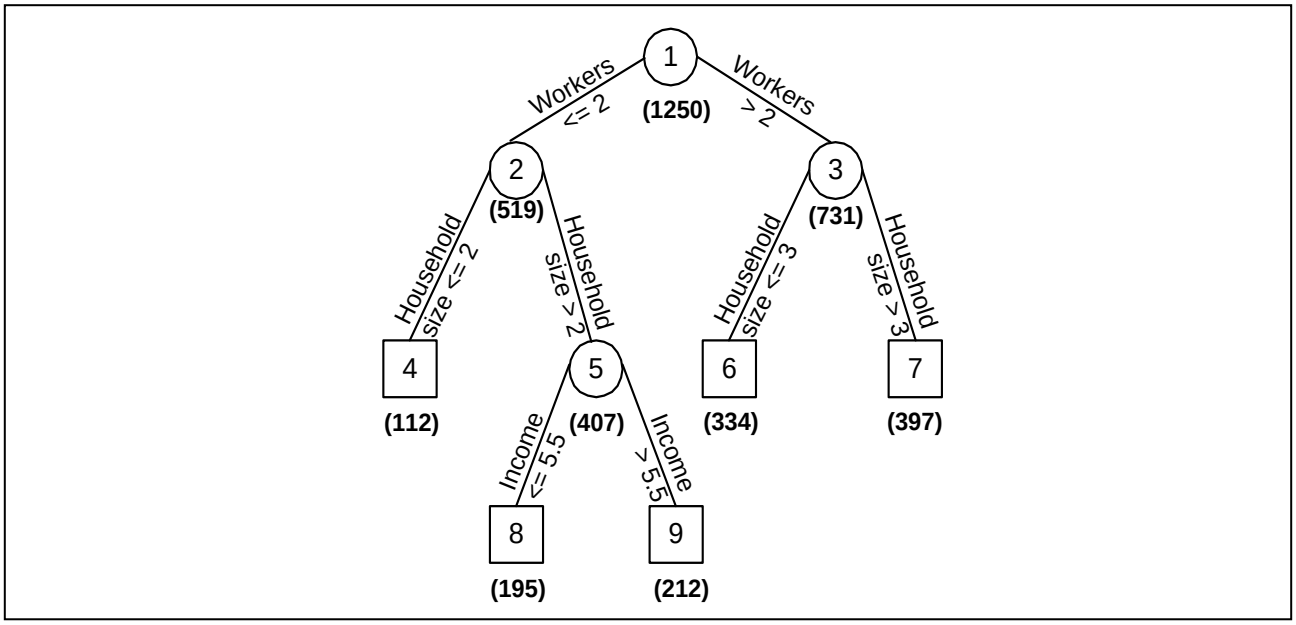


Figure-4.3: An example of a binary tree for a survey of some 1250 households

The number of observation for this survey example is 1250 households. This tree has nine nodes. The number within each node is the node number. Four of the nine nodes are non-terminal nodes indicated by circles and five are terminal nodes (all assumed *inactive or cannot be split further* at this point) indicated by squares. The numbers beneath the nodes in parentheses are number of observations in each node. Starting from node 1, the **split criterion** is the *number of workers in the household* and the **split value** is 2. The CART algorithm finds the best split criterion and split value for each nonterminal node (see Section 8.2 for more details). The algorithm partitions node 1 into two subnodes (nodes 2 and 3) based on the number of workers in the household. For each observation in the survey, if number of workers is less than or equal to 2, then this observation is classified into the left node (node 2); otherwise, it is classified into the right node (node 3). Suppose that $N_{\min}$ is 100 and the total deviance with respect to the dependent variables of nodes 2 and 3 are still much more than a specified value, then nodes 2 and 3 are candidates for further splitting. (Note that if $N_{\min} > 112$, then node 2 would not have been split further against the household size variable as shown in Figure-4.3.) The survey can be classified into five classes, as indicated in the figure, by the five terminal nodes. These are determined by the split criteria and the split values on the branches of the binary tree lying on the chain from the terminal node to the *root* node (node 1). For example, terminal node 4 has households having size ≤ 2, and number of workers ≤ 2.

4.3.2 Discrete Destination Zone Choice Model

Each activity in the household must have a specific location in the system. Initially, the Activity Generator Module selects an appropriate *zone* for the activity using a discrete choice model (this is explained in detail later in section 4.8.5.1). The concept of the discrete choice model **is to develop a probability distribution for selecting each zone based on the activity type, travel mode, and travel time for arriving to the location**, and to use these distributions to assign different types of activity locations to various zones. For example, a zone that contains mostly *industrial plants* would be appropriate for locating a *work* trip. A zone that provides more transportation choices, higher commercial developments, and lesser traffic congestion, would typically have a higher probability for attracting and locating a shopping trip within it.

4.4. Major Data Inputs

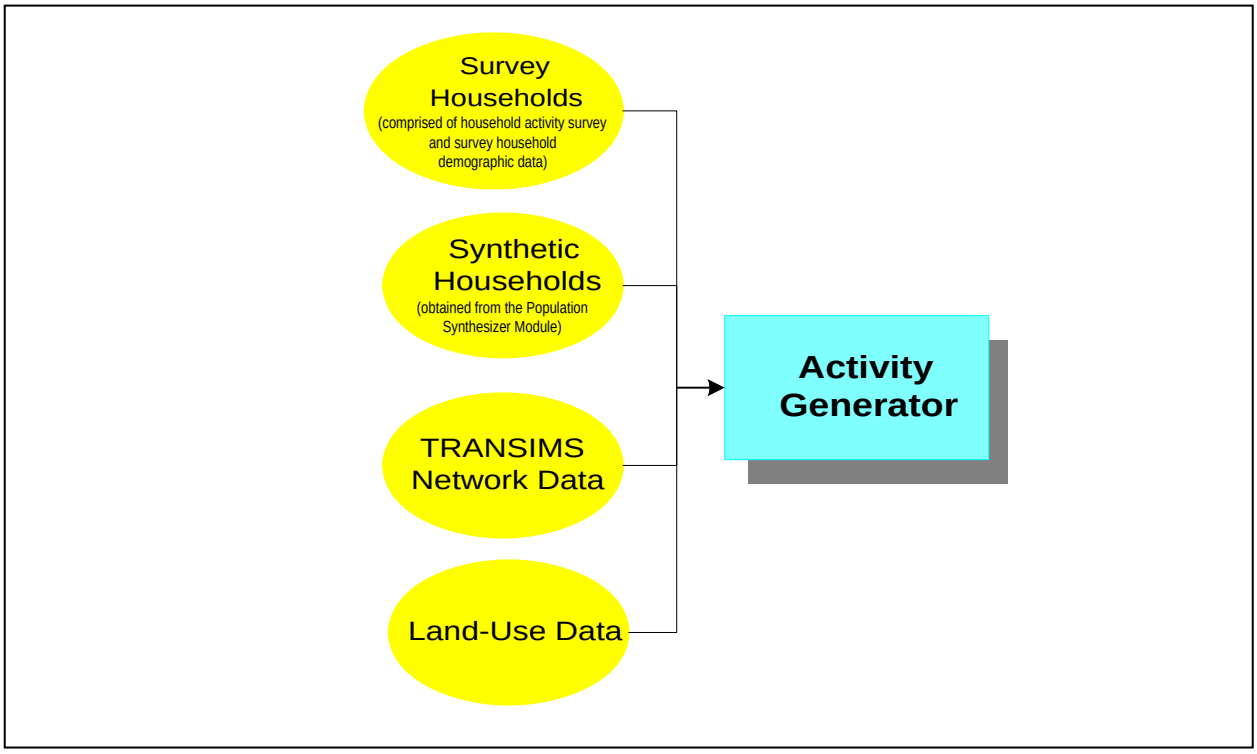


Figure-4.4: The major data inputs to the Activity Generator Module

The major inputs to the Activity Generator Module are comprised of **survey households (including household activity surveys and household demographic survey data)**, **synthetic households**, the **TRANSIMS network data**, and **land-use data**, as exhibited in Figure-4.4. The Activity Generator Module uses these inputs in different parts of the five-step algorithmic process as shown in Figure-4.5.

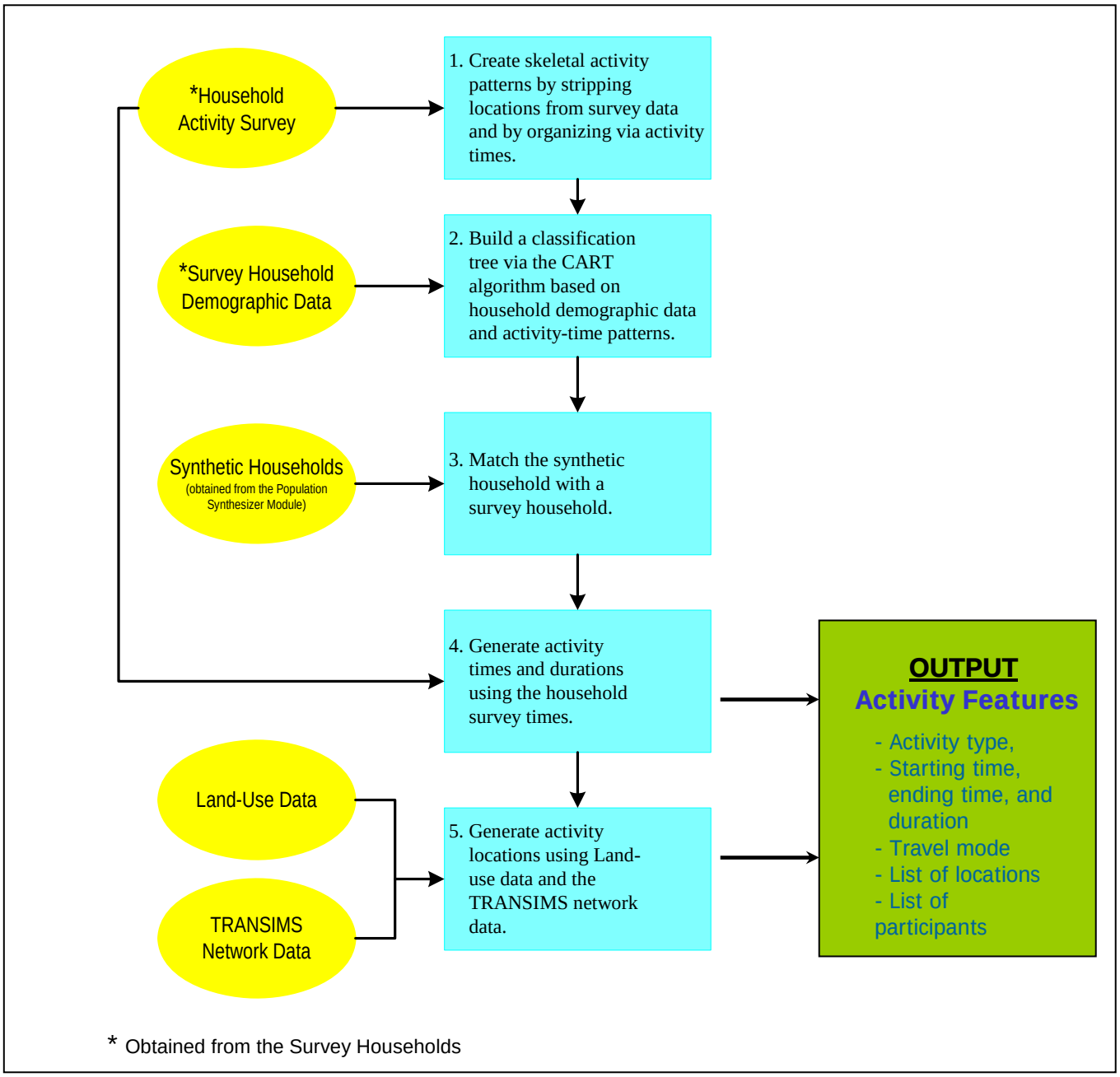


Figure-4.5: The major inputs to the Activity Generator Module

4.4.1 Household Activity Survey

The household activity survey represents a sample of households within the study region; including travel and activity participation information for each household member (see an example of the household activity survey form in *Appendix A*). This survey is used to create skeletal activity patterns by stripping the specific locations from the survey and by organizing the data via activity times. Activities are grouped sequentially for each survey household member (see Section 4.8.1 for the format of the skeletal activity pattern, which is the output created from the activity survey).

4.4.2 Survey Household Demographic Data

The CART algorithm uses household demographic survey data (see an example of the household demographic data survey form in *Appendix B*) to build a classification tree. Each survey household is effectively placed by the CART algorithm into one of the tree's terminal nodes. The survey household demographic data associated with each of the terminal nodes of this tree is used to match a given synthetic household with a survey household within a particular terminal node of the tree.

The format of the survey household demographic file must be the same as the format for the synthetic household demographic file. The survey household demographic file contains two header lines, followed by data lines (see Figure-4.6). The first header line contains the overall information about the household demographic characteristics. This information may vary for each household based upon the characteristics of the household. The second headline contains demographic information about each household member, pertaining to the variables: *relate, work, gender,* and *age* (in descending order) as defined by the code numbers in Table-4.2. The subsequent data lines repeat the demographic information for other household members.

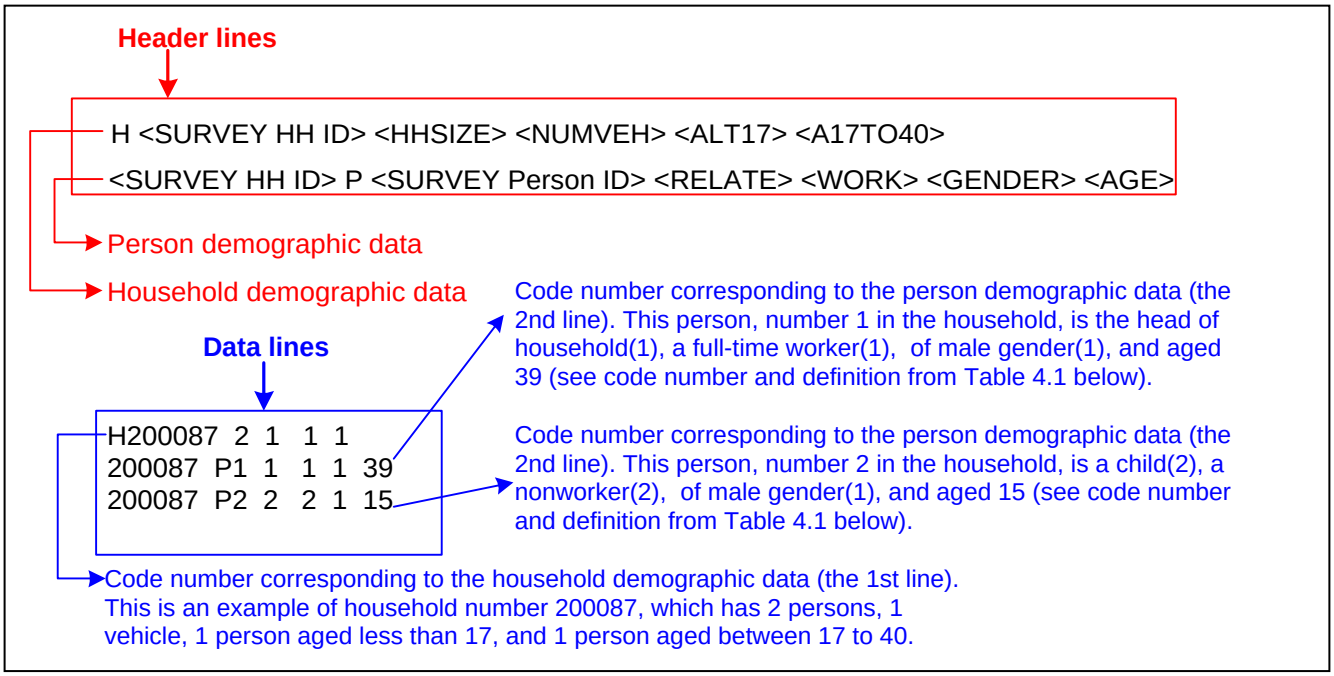


Figure-4.6: An example of the survey household demographic data for a household numbered 200087

Table-4.2: Definition of the code numbers for the four demographic variables: relate, work, gender, and age

Demographic Variable	Code Number and Definition
Relate	This is a variable coded for relationship to the head of household. It must be one of the following: 1 = spouse or partner, head of household 2 = child 3 = adult relative 4 = other.
Work	This is a variable coded for work status of the person. It must be one of the following: 1 = full-time/part-time worker (including self-employed) 2 = nonworker.
Gender	This is a variable coded for the gender of the person. It must be one of the following: 1 = male 2 = female.
Age	This is a variable coded for age of the person (in years).

4.4.3 Synthetic Households

Each synthetic household is matched with a survey household based on its demographic data. This process is explained in Section 4.8.3 in greater detail. The format of the synthetic household demographic file is the same as the survey household demographic file.

4.4.4 Land-Use Data

The Activity Generator uses zonal land use data to establish attraction parameters for selecting appropriate zones for each of the household activities. Each attractor variable represents a specific characteristic of the zone, for example, the total square footage of retail property, the total employment within the zone, etc.

The format of the zone data is comprised of a header line and various data lines. The header line contains the zone number, the easting and northing geocoordinates of the zone, and the activity types contained within the zone. The data lines contain the attractor values for each activity type in the zone in addition to the zone *x*-*y* coordinates and the zone number. The attractor values are obtained from the land-use characteristics of the zones.

Figure-4.7 illustrates and explains an example of the zone data for a test network having ten zones. Six activity types are defined: work, shop, school, visit, other, and serve passenger.

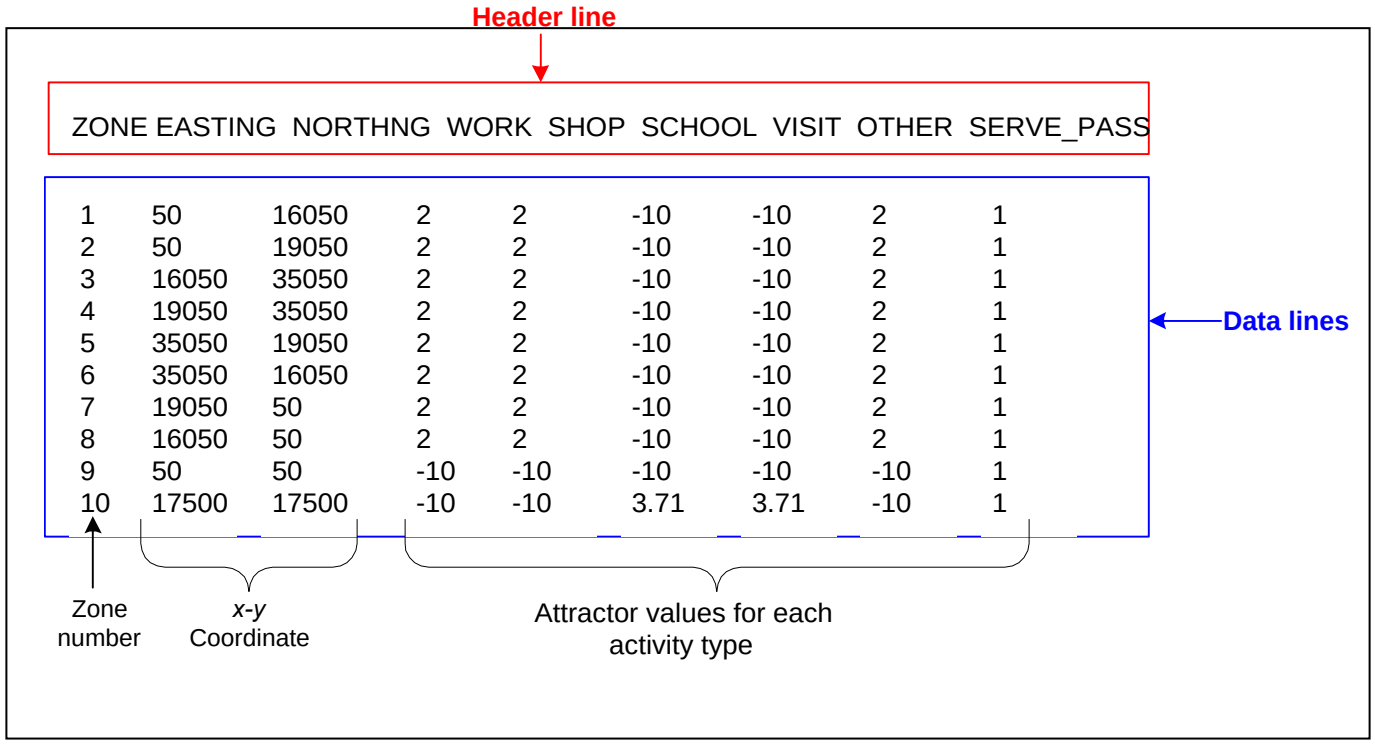


Figure-4.7: Example of the zone data

4.4.5 TRANSIMS Network Data

After obtaining the appropriate zone for each activity, its specific location within the zone must be selected. Activity locations are defined for the TRANSIMS network as offsets from a specified network node along a specified link. The specific location for each activity is stored in the activity location file. The files that the Activity Generator uses from the TRANSIMS network data in order to locate a specific activity are the *TRANSIMS network directory, the activity location table,* and *the node and link tables*. The TRANSIMS network directory shows where the network files reside, for example, **NET_DIRECTORY/home/projects/transims/networks/test** means that the specific files are stored in the home/projects/transims/networks/test directories, respectively. The format and the detailed description of the TRANSIMS network data files are discussed in Chapter 2.

4.5. Major Data Outputs

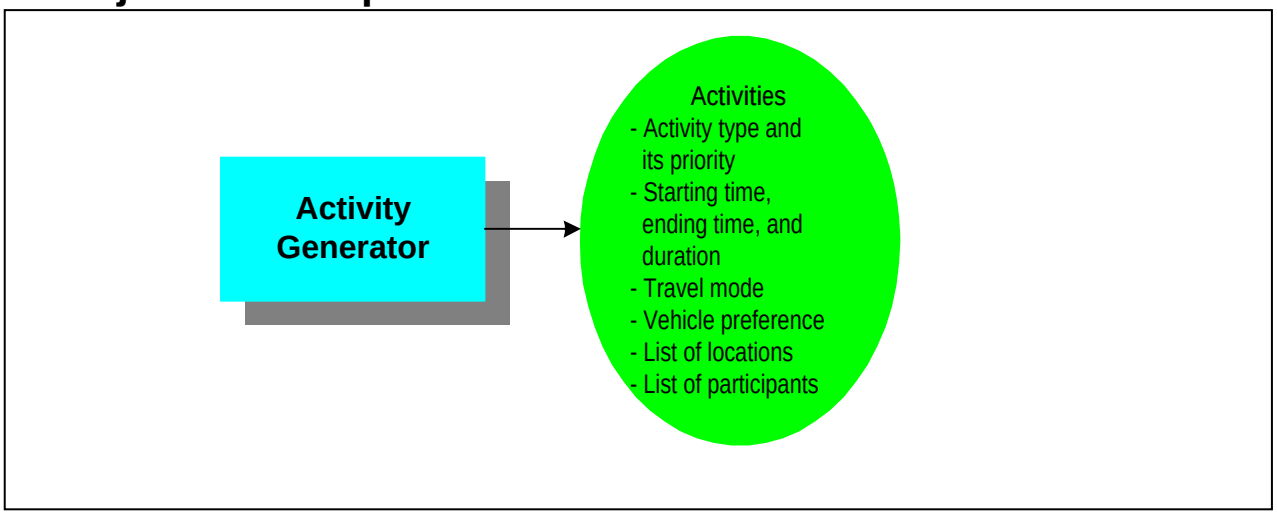


Figure-4.8: The major outputs from the Activity Generator Module

The primary output of the Activity Generator Module is the list of activities for each member of a synthetic household. Each activity list is comprised of the following:

- An activity type (e.g., home or school, etc.) and its priority,
- Starting, ending, and duration time preferences,
- A preferred travel mode,
- A vehicle preference (if appropriate),
- A list of possible locations for an activity, and
- A list of other participating HH members (if the activity is shared).

Table-4.3 provides more detailed outputs and a description of the activities called *TRANSIMS activities* and illustrates an example of the actual outputs from the Activity Generator Module.

Table-4.3: Format of the TRANSIMS activities (output from the Activity Generator Module)

Column Name/Abbreviation	**Description**
Household ID/HHID	Each household has a unique household ID. These numbers are assigned in the population file.
Person ID/PERID	Each person in a household is given a unique ID in the population file.
Activity ID/ACTNO	Each activity in the household has a unique ID.
Activity Type/ACTTYP	Each activity must be specified a type, for example, 0 = Home, 1 = Work, 2 = Shop, 3 = School, 4 = Visit, 5 = Other, 6 = Serve Passenger, etc.
Column Name/Abbreviation	**Description**
Activity Priority/PRIORITY	A 0 (zero) is an activity having a lowest priority; a priority of 9 means the

	activity must be done.
Starting Time Lower Bound /St Time-*a*	This time (hours since midnight) is the earliest time that the activity can start.
Starting Time Upper Bound /St Time-*b*	This time (hours since midnight) is the latest time that the activity can start.
Preferred Starting Time *a* parameter/St-*a*	This is a parameter for a beta distribution of the starting time lower bound. The parameter indicates a shape of the distribution (see Section 8.5 for more details).
Preferred Starting Time *b* parameter/St-*b*	This is a parameter for a beta distribution of the starting time upper bound. The parameter indicates a shape of the distribution (see Section 8.5 for more details).
Ending Time Lower Bound /End T-*a*	This time (hours since midnight) is the earliest time that the activity can end.
Ending Time Upper Bound /End T-*b*	This time (hours since midnight) is the latest time that the activity can end.
Preferred Ending Time *a* parameter/Et-*a*	This is a parameter for a beta distribution of the ending time lower bound. The parameter indicates a shape of the distribution (see Section 8.5 for more details).
Preferred Ending Time *b* parameter/Et-*b*	This is a parameter for a beta distribution of the ending time upper bound. The parameter indicates a shape of the distribution (see Section 8.5 for more details).
Duration Lower Bound /Dur Tim-*a*	This time (hours) is the shortest length of the activity.
Duration Upper Bound /Dur Tim-*b*	This time (hours) is the longest length of the activity.
Duration *a* parameter /D-*a*	This is a parameter for a beta distribution of the duration lower bound. The parameter indicates a shape of the distribution (see Section 8.5 for more details).
Duration *b* parameter /D-*b*	This is a parameter for a beta distribution of the duration upper bound. The parameter indicates a shape of the distribution (see Section 8.5 for more details).
Mode Preference for Arriving at the Activity/Mode	This number represents a string that defines the mode preference to the Route Planner (e.g., *wcw*, *wtw*, …), where the label codes are as follows: 1 = Walk (*w*), 2 = Car (*c*), 3 = Transit (*t*), 4 = Light Rail (*l*), 5 = Park and Ride outgoing, 6 = Park and Ride incoming, 7 = Bicycle (*i*), 8 = Magic Move (*k*), etc.
Vehicle ID/VEHid	This is the vehicle ID for all activities having a mode preference of car, either as a driver or as a passenger. This field should be set to –1 for all other modes.
Number of Possible Locations for Activity/N Loc	This is the number of possible locations in the List of Locations field. The value 0 is not allowed. If this field is –1, then the single value in the List of Locations field is an index representing a group of activities.
List of Activity Locations/Location	If the Number of Possible Locations field is 1 or greater, this field contains a list of activity location IDs where some activity might occur. If the Number of Possible Locations field is –1, this field contains a number that is an index representing a group of activities.
Number of Other Participants in the Activity/N Others	This is the number of other household members who might participate and use the same transportation mode (e.g., the same car) to reach this activity location. The number is 0 if the person is to travel alone to the activity location.
List of Other Participants/OthersID	This contains the person IDs of participants using the same transportation mode. This field should be present only when the value of the Number of Other Participants field is positive. The first entry in this list is always the driver ID.
Activity Group Number /G Number	Every activity for an **individual** will have a number. Sets of activities that must be done together will have the same number.

Table-4.4 displays an example of the actual output from the Activity Generator Module contained in a TRANSIMS Activities file for a household having an ID 13092. In this example, there are two people in the household. The first person has an ID 13300, and the second person has an ID 13301. There are three activities for each person. These two individuals have the same type of activities, which relate to going from "home" to "work," and back to "home" (this is coded in the activity type (**ACTTYP**) column in Table-4.3, where 0 designates home, and 1 designates work). Every activity must be performed because each activity has a 9 priority level. The starting time, ending time, and duration for each activity varies as seen from the table. Although the "home" activities correspond to the same location, which is a house having an ID 845654 (see the **Location** column), the "work" activities are different. The first person goes to work at location ID 833503 via a car (as seen in coded form in the **Mode** column, where 1 designates walk, and 2 designates car), and where the car has an ID 13217 (this is specified in the vehicle ID (**VEHid**) column), while the second person goes to work at location ID 853676 via a car having ID 13218. These two persons travel alone to the "work" activities (this is recorded in the column designating the number of other participants in the activity (**N Others**)).

Table-4.4: Example of a TRANSIMS Activities file (part 1)

HHID	PERID	ACTNO	ACTTYP	PRIORITY	St Time-*a*	St Time-*b*	St-*a*	St-*b*	End T-*a*	End T-*b*	Et-*a*	Et-*b*
13092	13300 1st person	1	0 (home)	9	0	0	-1	-1	7.0833	8.5833	1	1
13092	13300	2	1 (work)	9	7.5833	8.0833	1	1	14.75	15.25	1	1
13092	13300	3	0 (home)	9	14.75	16.25	1	1	24	24	-1	-1
13092	13301 2nd person	4	0 (home)	9	0	0	-1	-1	8.25	9.75	1	1
13092	13301	5	1 (work)	9	9.1667	9.6667	1	1	16.25	16.75	1	1
13092	13301	6	0 (home)	9	16.25	17.75	1	1	24	24	-1	-1

(part 2)

Dur Tim-*a*	Dur Tim-b	D-*a*	D-*b*	Mode	VEHid	N Loc	Location	N Others	OthersID	G Number
7.0833	8.5833	1	1	1	-1	1	845654 (home)	0	-	1
6.9167	7.4167	1	1	2	13217	1	833503 (work1)	0	-	1
7.75	9.25	1	1	2	13217	1	845654 (home)	0	-	1
8.25	9.75	1	1	1	-1	1	845654 (home)	0	-	1
6.8333	7.333	1	1	2	13218	1	853676 (work2)	0	-	1
6.25	7.75	1	1	2	13218	1	845654 (home)	0	-	1

The information given in Table-4.4 is then sent to the Route Planner Module to find shortest paths for the various transportation activities.

4.6. Module Interfaces

An outline of the module interfaces for the Activity Generator Module is shown in Figure-4.9.

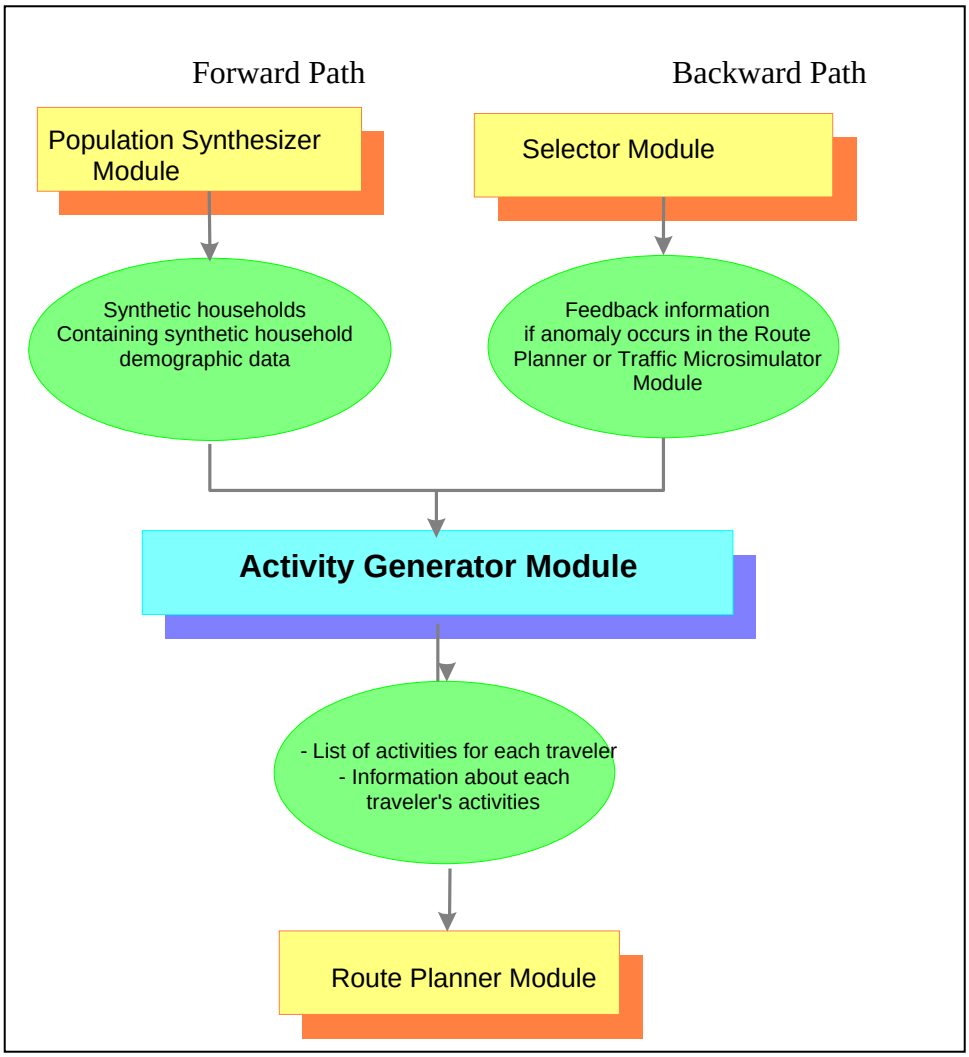


Figure-4.9: Flowchart of the module interfaces for the Activity Generator Module

4.6.1 Input received from the Population Synthesizer Module

The inputs from the Population Synthesizer Module provide information on **the synthetic households**. The format of the synthetic household demographic file is the same as the survey household demographic file as specified before.

4.6.2 Input received from the Selector Module

When the Route Planner or the Traffic Microsimulator Module detects a set of anomalous activities, they will feed them back to the Selector Module. The list of anomalous activities is comprised of the following:

- Activity pattern (comprised of *no path* and *invalid shared ride* anomalies)*.
- Location for an activity.
- Travel mode for arriving at the activity.
- Activity times (comprised of *invalid time* and *invalid shared ride time* anomalies). Each of these anomalies has three possible types of anomalous times: *start, end,* and *duration**.
- Travel time between two activity locations.

For more details of these anomalies, see Section 6.4 in the* ***Route Planner Module**.*

The Selector Module uses this feedback to determine whether to request changes by the Activity Generator Module. If a change is needed by the Activity Generator Module, the Selector Module will create feedback command files that tell the Activity Generator Module which activity needs to be regenerated and what action needs to be taken for each specified activity. The Activity Generator Module may **partially** change the activity or regenerate a **new activity list** for the entire household that contains the activity.

To regenerate a **new activity list** for the entire household, the **synthetic household** is **re-matched with a new survey household** in the regression tree in order to determine an alternative activity pattern for the household. Otherwise, a **partial change** in the location, travel mode, or activity times are used to correct the anomalies.

The format of the feedback command file contains a header line, followed by data lines (see Figure-4.10). The header line contains the household ID, the activity ID that needs to change for the household, and a command. This command is a string of codes that may contain optional command parameters. Currently, there are four types of commands:

- **L** = change the location for the activity;
- **M (followed by a new travel-mode value)** = change the travel mode for the activity to a new specified one;
- **LM (followed by a new travel-mode value)** = change the travel mode for the activity to a new specified one and then change the location for the activity;
- **T (followed by a new *starting time, ending time, lower bound parameter-a, upper bound parameter-b*)** = change the starting time, ending time, lower bound parameter (*a*), upper bound parameter (*b*) for the activity to the new specified values, and
- **R** = regenerate the entire activity list for the household by re-matching it with a new survey household.

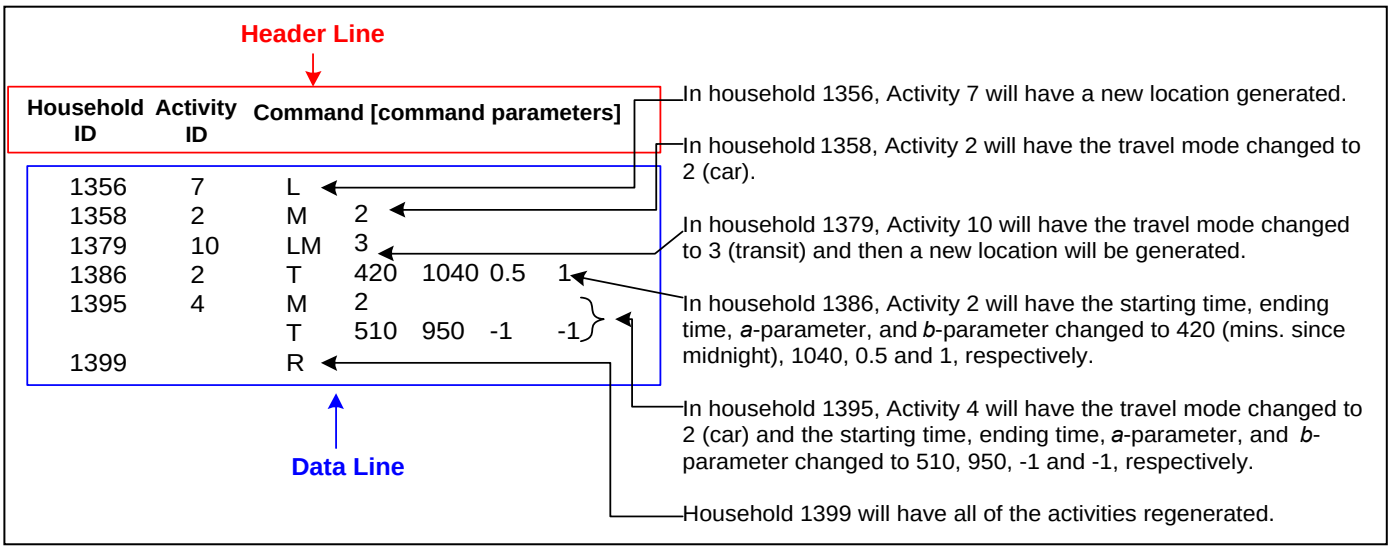


Figure-4.10: An example of the feedback command from the Selector Module

4.6.3 Output sent to the Route Planner Module

The output sent to the Route Planner Module is the list of activities for each member of a matched synthetic household, along with information regarding each traveler's activities. The Route Planner Module uses this information to generate the trip request for each traveler, and then finds shortest paths for the various transportation activities. See Chapter 5 for more details on the output sent to the Route Planner Module.

4.7 Configuration Files

Table-4.5 lists all files used to run the Activity Generator Module and their descriptions, including the default values.

Table-4.5: The list of all files used by the Activity Generator Module

Configuration Key	Description
ACT_ACCESS_HEADER	The user data column header in the network activity location file used to specify access to transit.
ACT_ACTIVITY_TYPE	Used to specify the activity types used by the Activity Generator. The base key must be followed with_*n* where n is an integer to indicate the nth specification of the activity type (nonnegative integer).
ACT_BICYCLE_MODE	The number of the bicycle mode (*wiw*) (integer).
ACT_BLOCKGROUP_HEADER	The user data column header in the network activity location file used to specify a block group.
ACT_CAR_MODE	The number of the car mode (*wcw*) (integer).

Configuration Key	Description
ACT_CARPOOL_ACTIVITY_TYPE	The number of the car pool activity type (nonnegative integer).
ACT_DICISION_TREE_FILE	The name of the file containing the regression tree for the Activity Generator.
ACT_DEFAULT_CAR_SPEED	The default speed for automobiles in meters/second (floating point number) Default = 37.5
ACT_DEFAULT_INTRAZONE_TRAVEL_TIME	The default travel time within a zone in seconds (integer). Default = 60
ACT_DEFAULT_TRANSIT_MODE	The number of the default transit mode (*wtw*) (integer).
ACT_DEFAULT_TRANSIT_SPEED	The default transit speed in meters/second (floating point number). Default = 30.5
ACT_END_OF_DAY_TIME_RANGE	The time range in hours for lower and upper bounds of start and end times for the end-of-day activity (positive floating number). Default = 0.75
ACT_HOME_ACTIVITY_TYPE	The number of the home activity type (nonnegative integer).
ACT_HOME_DURING_DAY_TIME_RANGE	The time range in hours for lower and upper bounds of start and end times for non-work activities originating at home (positive floating point number). Default = 0.75
ACT_HOME_HEADER	The user data column header in the network activity location file used to specify single family home locations.
ACT_INITIAL_HOME_TIME_RANGE	The time range in hours for lower and upper bounds of start and end times for the initial at-home activity (positive floating point number). Default = 0.75
ACT_LOCATION_HEADER	Used to specify the header for the activity type for activity locations in the network activity location table. The headers must correspond to the activity types defined with the ACT_ACTIVITY_TYPE_N keys. The base key must be followed with _*n* where *n* is an integer to indicate the nth specification of the header.
ACT_MAGIC_MOVE_MODE	The number of the magic-move mode (*wkw*) (integer).
ACT_MAX_END_TIME	The maximum end time for an activity in hours past midnight on the starting day (positive floating point number).
ACT_MODE_WEIGHT_FILE	The name of the file containing mode coefficients for the activity types. This must contain a coefficient for every mode and activity type.
ACT_MULTI_FAMILY_HEADER	The user data column header in the network activity location file used to specify multifamily home locations.
ACT_OUT_OF_HOME_TIME_RANGE	The time range in hours for lower and upper bounds of start and end times for non-work activities that do not originate at home (positive floating point number). Default = 0.75
ACT_PERSON_DEMOG__AGE_HEADER	The name of the age demographic header for the persons in the population used by the Activity Generator.

Configuration Key	Description
ACT_PERSON_DEMOG_GENDER_HEADER	The name of the gender demographic header for the persons in the population used by the Activity Generator.
ACT_PERSON_DEMOG_RELATION_HEADER	The name of the relationship demographic header for the persons in the population used by the Activity Generator.
ACT_PERSON_DEMOG_WORKER_HEADER	The name of the worker demographic header for the persons in the population used by the Activity Generator.
ACT_POPULATION_FILE	The name of the file containing a located synthetic population, with household and person demographic data that exactly match the variables in the Activity Generator regression tree. This file is an output from the population converter program.
ACT_RANDOM_SEED	The random number seed used by the Activity Generators.
ACT_REQUIRED_HH_DEMOG	This specifies the required household demographic data in the synthetic population used by the Activity Generator. The base key must be followed with _*n* where *n* is an integer to indicate the nth specification of required demographic data. The demographic data must exactly match and be ordered the same (1-*n*) as the demographic variables in the Activity Generator's regression tree.
ACT_SCHOOL_ACTIVITY_TYPE	The number of the school activity type (nonnegative integer).
ACT_SHARED_RIDE_DISTANCE_RANGE	The distance range in meters for matching activity locations for shared rides. Default = 2000
ACT_SHARED_RIDE_TIME_RANGE_MAX	The maximum time range in minutes for matching activities for shared rides. Default = 60
ACT_SHARED_RIDE_TIME_RANGE_MIN	The minimum time range in minutes for matching activities for shared rides. Default = 15
ACT_SURVEY_ACTIVITY_FILE	The name of the file containing activity patterns for the survey households.
ACT_SURVEY_HOUSEHOLD_FILE	The name of the file containing the survey household population and demographic data.
ACT_SURVEY_WEIGHTS_FILE	The name of the file containing the relative weights of the survey households.
ACT_TAZ_HEADER	The user data column header in the network activity location file used to specify traffic analysis zones.
ACT_TRACT_HEADER	The user data column header in the network activity location file used to specify census tracts.
ACT_TRAVEL_TIMES_FILE	The name of the file containing travel time information between zones.
ACT_TRIP_TABLE_OUTPUT	The name of the file that will be output from the Trip Table Activity Generator.
ACT_TRIP_TABLE_VEHICLE_FILE	The name of the vehicle file that will be output from the Trip Table Activity Generator.
ACT_TRIPTABLE_FILE	The name of the file containing the trip table matrix.
ACT_TRIPTABLE_STARTING_HH_ID	The starting household ID for households generated from the Trip Table Activity Generator.

Configuration Key	Description
ACT_TRIPTABLE_STARTING_PERSON_ID	The staring person ID for travelers generated from the Trip Table Activity Generator.
ACT_TRIPTABLE_STARTING_VEHICLE_ID	The staring vehicle ID for vehicles generated from the Trip Table Activity Generator.
ACT_TRIPTIME_FILE	The name of the file containing the time-of-day trip table data.
ACT_WALKING_MODE	The number of the walking mode (*w*) (integer).
ACT_WORK_ACTIVITY_TYPE	The number of the work activity type (nonnegative integer).
ACT_WORK HEADER	The user data column header in the network activity location file used to specify work locations.
ACT_WORK_TIME_RANGE	The time range in hours for lower and upper bounds of start and end times for work activities (positive floating point number). Default = 0.25
ACT_ZONE_HEADER	Used to specify the header for the zone attractors ,which must match the activity types (ACT_ACTIVITY_TYPE_N). The base key must be followed with _*n* where *n* is an integer to indicate the nth specification of the header.
ACT_ZONE_INFO_FILE	The name of file containing zone attractor data by activity type for the Activity Generator.
ACTIVITY FILE	The output file from the Activity Generator.
NET_ACTIVITY_LOCATION_TABLE	The network activity location table name.
NET_DIRECTORY	The directory where the network files reside.
NET_LINK_TABLE	The network link table name.
NET_NODE_TABLE	The network node table name.
ROUTER_BIKING_SPEED	The approximate speed in meters/sec for bicycles.
ROUTER_WALKING_SPEED	The approximate speed in meters/sec for walking trips.
VEHICLE_FILE	The name of the TRANSIMS vehicle file for the population.
ACT_PARTIAL_OUTPUT	The name of the file that will be output from the partial regeneration of activities.
ACT_FEEDBACK_FILE	The file containing a list of travelers and associated commands for activity regeneration.
ACT_HHDENSITY_HEADER	The column header of the household density values in the network activity location tables.
ACT_POPULATION_FILE	The name of the file containing a located synthetic population with household and person demographic data that exactly match the variables in the Activity Generator regression tree. This file is output from the Population Converter program.
POP_LOCATED_FILE	The name of the file containing a located population.

4.8 Algorithms

The algorithms used comprise the following five steps:

- Create skeletal activity patterns from the survey households,
- Use the CART (Classification and Regression Tree) algorithm to build a classification tree based on household demographic data,
- Match each synthetic household with a survey household,
- Generate activity times and durations, and
- Generate non-home activity locations.

The first two steps use the input data for the Activity Generator module, while the last three steps generate the output data of the module. These steps are described in detail below.

4.8.1 Creating skeletal activity patterns from the survey households

The activity lists for the members of each survey household are organized by trips and are stripped of locations to create a library of skeletal activity/travel patterns as shown in Table-4.6. Activities are grouped sequentially for each member in the household as collected from the household activity survey form (see the examples of the household activity survey form in *Appendix A*).

The format of the skeletal activity pattern is shown in Table-4.6, and an example is given in Figure-4.11.

Table-4.6: Survey activity file format

Column Name/Abbreviation	Description
Survey Household number /SAMPNO	Each survey household has a unique number. These numbers are assigned in the population file.
Person Number /PERSNO	Each person in a household has a unique number, starting with 1. Numbers are only unique within the household.
Activity Number /ACTNO	Each activity in the household has a unique number.
Activity Type /ACTID	Each activity must be specified a type, for example, 0 = at-home activity 1 = work 2 = shop 3 = school 4 = visit 5 = other 6 = serve passenger, etc.
At Home /AT_HOME	This number is 1 if the activity is the at-home activity, and is 2 otherwise.
Were-you-there /WUTHERE	This number is 1 if the person is already at the location, and is 2 otherwise.

Table-4.6 Contd.

Column Name/Abbreviation	Description
Mode for arriving at activity /MODE	This number represents a string that defines the travel mode for arriving at the specified activity (e.g., *wcw*, *wtw*, …), where the label codes are as follows: 1 = Walk (*w*), 2 = Car (*c*), 3 = Transit (*t*), 4 = Light Rail (*l*), 5 = Park and Ride outgoing, 6 = Park and Ride incoming, 7 = Bicycle (*i*), 8 = Magic Move (*k*), etc.
Driver /DRIVER	The number is 1 if the person is a driver, a 2 if the person is a passenger, and is zero otherwise.
Number in Vehicle /NUMVEH	This is the number of people in a vehicle (including a driver).
Activity Start Time /ACTSTART	This time (minutes since midnight) is the time that the activity can start.
Activity End Time /ACTEND	This time (minutes since midnight) is the time that the activity can end.
Geocode *x* /GEOX	This is an easting geocoordinate for the activity location obtained from the TRANSIMS network data.
Geocode *y* GEOY	This is a northing geocoordinate for the activity location obtained from the TRANSIMS network data.

The example given in Figure-4.11 specifies the skeletal activity pattern for a household numbered 200087 that contains two persons. In this example, the survey of household activities has started at 3:00 a.m. and ended at midnight. The **first** person has nine activities including an initial at-**home** activity that started at 3:00 a.m., **service a passenger** activity (being a driver to take the second person to school), **working** activity, **other** activity, going back to **working** activity, **visiting** some place activity, **service a passenger** activity (being the driver to pick up the second person from school), and then followed by two at-**home** activities. The second person in that sample has three activities listed below person 1.

SAMPNO	PERSNO	ACTNO	ACTID	AT_HOME	WUTHERE	MODE	DRIVER	NUMVEH	ACTSTART	ACTEND	GEOX	GEOY
200087	1	0	0 (home)	1 (yes)	2 (no)	1 (walk)	0	0	180	460	7637347.0000	687428.0625
200087	1	1	6 (serve passenger)	2 (no)	2	2 (car)	1 (driver)	2	480	482	7648077.5000	713295.7500
200087	1	2	1 (work)	2	2	2	1	1	510	715	7640385.5000	683294.6875
200087	1	3	5 (other)	1	2	2	1	1	730	760	7637347.0000	687428.0625
200087	1	4	1 (work)	2	2	2	1	1	775	960	7640385.5000	683294.6875
200087	1	5	4 (visit)	2	2	2	1	1	990	1010	7637347.0000	687428.0625
200087	1	6	6 (serve passenger)	2	2	2	1	1	1020	1023	7648077.5000	713295.7500
200087	1	7	0 (home)	1	2	1	0	0	1050	1320	7637347.0000	687428.0625
200087	1	8	0 (home)	1	2	1	0	0	1320	1440	7637347.0000	687428.0625
200087	2	0	0 (home)	1	2	1	0	0	180	460	7637347.0000	687428.0625
200087	2	1	3 (school)	2	2	2	2 (passenger)	2	480	1023	7648077.5000	713295.7500
200087	2	2	0 (home)	1	2	1	0	0	1050	1440	7637347.0000	687428.0625

Figure-4.11: An example of the skeletal activity pattern for the survey household numbered 200087

4.8.2 Using CART algorithm to build a classification tree based on household demographic data

This algorithm is comprised of two steps: the **tree growing step** and **the cross-validation step for reducing or pruning the tree.**

4.8.2.1 Tree-growing step

The CART algorithm is a technique for modeling a regression relationship between one or more dependent variables Y and independent variables $X_1, X_2, \ldots, X_K$. A classification tree is constructed using the total time a household spends in some 15 broadly classified activity types, obtained from the skeletal activity patterns in Section 8.1, as the dependent variables Y_1, Y_2, …, Y_{15}. For example, Y_1 = total time spent by the household in working, etc. An additional variable Y_{16} is used to represent the total number of trips made by the household over the specified 24-hour horizon. The independent variables X_k identify various household demographic attributes that are obtained from the survey household demographic data specified in Section 4.2. For example, X_1 = household size, X_2 = number of vehicles in the household, etc.

Mathematical Terminology and Definition

$\boldsymbol{S}$: survey sample comprised of n data observations indexed $i = 1, \ldots, n$.

$\boldsymbol{X_k, k = 1, \ldots, K}$: given set of K independent demographic attributes.

$\boldsymbol{X_{ik}}$ = value of demographic attribute X_k in the ith survey observation, $i = 1, \ldots, n$, and $k = 1, .., K$.

$\boldsymbol{Y_j, j = 1, \ldots, p}$: given set of p dependent variables used to measure the performance characteristics of the household (in TRANSIMS, $p = 16$).

$\boldsymbol{Y_{ij}}$ = value of measure Y_j for the ith survey observation, $i = 1, \ldots, n$, and $j = 1, .., p$.

$\boldsymbol{T}$ = classification tree, comprised of nodes and branches.

$\boldsymbol{N}$ = a particular current node of the tree T consisting of a subset of households from S (we will generally view N as being a subset of S).

$\boldsymbol{\tilde{T}}$ = set of terminal or leaf or end nodes of T.

$\boldsymbol{\tilde{T}_A}$ = set of active terminal nodes, which are the nodes in $\tilde{T}$ that may be split into two subnodes. (Terminal nodes in $\tilde{T}$ - $\tilde{T}_A$ are considered as being *inactive.*)

$\boldsymbol{N_{\min}}$: this is a parameter used to declare a terminal node as being inactive. Specifically, if the number of households that belong to either of the children nodes of a terminal node when it is split are less than the number defined as $N_{\min}$, then this terminal node is designated to be an inactive terminal node.

$\boldsymbol{D_j(N)}$ = deviance of the node N (where we also consider $N \subseteq S$) with respect to measure j, for j = 1, .., p:

$$D_j(N) = \sum_{i \in N} (Y_{ij} - \overline{Y}_{Nj})^2 \ , \qquad (1)$$

where $\overline{Y}_{Nj} = \dfrac{\sum_{i \in N} Y_{ij}}{|N|}$ = Mean value of dependent variable j for observations in Node N

$\boldsymbol{D(N)}$ = total deviance of node N:

$$D(N) = \sum_{j=1}^{p} s_j D_j(N) \ = \text{weighted sum of the deviances of all dependent variables,} \qquad (2)$$

where s_j = 1/variance (Y_j) = Total variance in the value of dependent variable j over the entire survey population.

$$= \frac{n-1}{\sum_{i=1}^{n} (Y_{ij} - \overline{Y}_j)^2}$$

and where $\overline{Y}_j = \dfrac{\sum_{i=1}^{n} Y_{ij}}{n}$.

$\boldsymbol{t_k(N)}$ = set of possible split values for independent variable k such that there exist observations in N having $X_{ik} \leq t$ and others having $X_{ik} > t$ for each $N \subseteq S$, and k = 1, …, K.

$\boldsymbol{N_{1kt}}$ **and** $\boldsymbol{N_{2kt}}$**:** they are children nodes of node N based on an independent variable X_k, split at a value t, such that $X_{ik} \leq t \ \forall \ i \in N_{1kt}$ and $X_{ik} > t \ \forall \ i \in N_{2kt}$.

$\boldsymbol{\Delta_{kt}(N)}$: decrement in the total deviance upon partitioning the current node N into two subnodes N_{1kt} and N_{2kt}:

$$\Delta_{kt}(N) = D(N) - [D(N_{1kt}) + D(N_{2kt})].$$

DOWN(·): if the current node N is split into two subnodes N_1 and N_2, we set DOWN (N_1) = N and DOWN (N_2) = N as well.

INITIALIZATION STEP

1. From the given survey population comprised of n data observations (households) indexed i = 1, …, n, let the root node be designated as S, representing the total survey. Initialize the tree $T = \{S\}$ to have the node S, with the set of terminal nodes $\tilde{T} = \{S\}$, and the set of active terminal nodes $\tilde{T}_A = \{S\}$ as well.
2. Specify the measures X_{ik} (value of the household demographic attribute X_k for the ith survey observation) for k = 1, …, K, obtained from the survey household demographic data in Section 4.2, and Y_{ij} (value of measure Y_j for the ith survey observation) for j = 1, …, 16 , which represent the set of dependent variables as described above. The information on the Y_j-variables is obtained from the skeletal activity patterns as shown in Section 8.1.
3. Let $N_{\min} = 10$ (this is one of the criteria for declaring terminal nodes as mentioned above).
4. Compute $D(S)$, and let $\beta \equiv 0.01 D(S)$ be the threshold for another criterion for designating terminal nodes as being active and inactive.
5. Let DOWN (S) $\equiv$ 0.

MAIN STEP

1. If the set of **active terminal nodes** $\tilde{T}_A$ is empty, stop the tree-growing procedure with the **output** tree T and its **terminal** nodes $\tilde{T}$. Otherwise, proceed to the next step.
2. Pick an **active terminal** node N that has the maximum deviance.
3. Compute $\Delta_{kt}(N)$ for each k = 1, .., K, and $t \in t_k$ (N) and find the largest such value: $\Delta_{k^*t^*}(N) = \max_{k,t}\{\Delta_{kt}(N)\}$. Let the corresponding partition of the current node N be $\{N_1, N_2\} \equiv \{N_{1k^*t^*}, N_{2k^*t^*}\}$.
4. Check if number of observations in each of the subnodes N_1 and N_2 is at least greater than some minimal value $N_{\min}$ (taken as 10). If so, then conduct this partitioning by proceeding to Step 5. Otherwise, do not perform the partitioning, declare the terminal node N as inactive, and remove N from the set of active terminal nodes $\tilde{T}_A$. (Note that this terminal node N will end up being a terminal node of the **final** tree that is constructed.) Repeat the Main Step.
5. Partition N into N_1 and N_2, and add N_1 and N_2 to the tree T with DOWN (N_1) = DOWN (N_2) = N.
6. Compute and store $D(N_1)$ and $D(N_2)$. Remove N from the set of terminal nodes $\tilde{T}$ and the set of active terminal nodes $\tilde{T}_A$, and add N_1 and N_2 to $\tilde{T}$. If $D(N_i) > \beta$, then also add N_i to $\tilde{T}_A$, for each i = 1, 2. Repeat the Main Step.

Figure-4.12 displays a flow-chart for this Tree-Growing procedure.

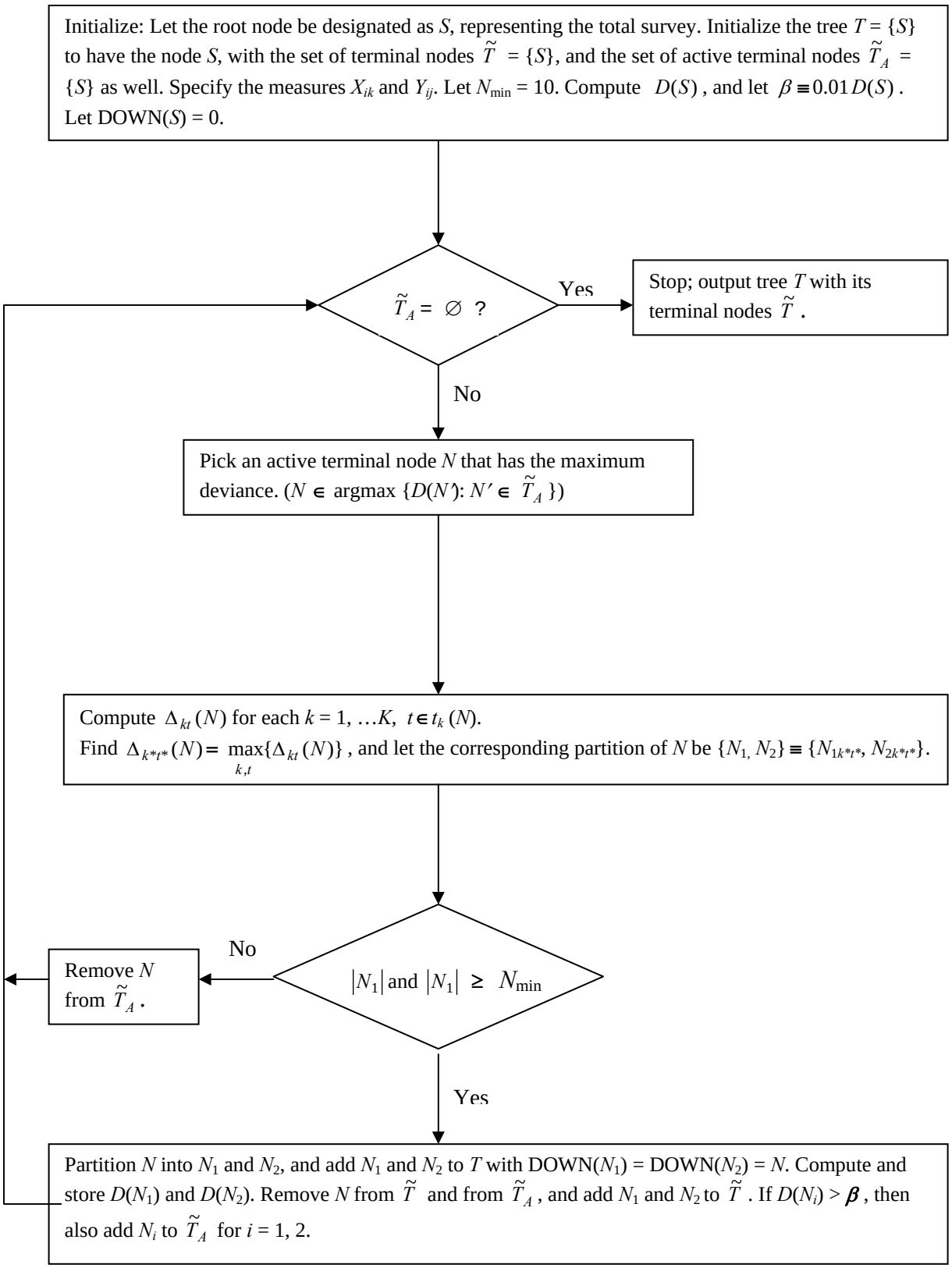


Figure-4.12: Flow-Chart for the Tree-Growing Procedure

Simple example for the tree-growing step

The following example illustrates the above method for growing a tree using a small database. Suppose that we are given a survey of ten households. In addition, we are given the corresponding variables X_k (independent variables) and Y_j (dependent variables) for each household as shown in Table-4.7. (A full description of how to obtain these variables will be provided in the next practical example.)

Table-4.7: Values of the X_k (independent variables) and Y_j (dependent variables) for the ten households

Household Number (i)	HHSIZE	HHAGE	WORKERS	T_Home	T_Work	N_Nonhome
	X1	X2	X3	Y1	Y2	Y3
1	1	29	1	918	417	3
2	1	82	0	1395	0	2
3	2	25	1	2275	385	11
4	1	82	0	1620	0	0
5	1	77	0	1310	0	2
6	3	66	3	2970	770	14
7	1	54	1	1560	0	1
8	4	40	2	4260	475	16
9	3	25	2	3060	1180	6
10	2	52	2	1765	1145	10
				$\overline{Y}_1$ = 2113.3	$\overline{Y}_2$ = 437.2	$\overline{Y}_3$ = 6.5

For the sake of illustration, we use only three X_k (independent variables) and three Y_j (dependent variables). (Generally, in a practical implementation, there would be more variables of each type.) The three dependent variables (Y_j, j = 1, 2, 3) were selected to represent the activity travel pattern of a household as the following:

Y_1 = total household duration (in minutes) for in-home activity (THOME),
Y_2 = total household duration (in minutes) for work (TWORK), and
Y_3 = number of non-home based trips (NONHOME).

The three independent demographic variables (X_k, k = 1, 2, 3) are defined as follows:

X_1 = number of persons in the household (HHSIZE),
X_2 = age of the head of the household (HHAGE), and
X_3 = number of employed household members (WORKERS).

INITIALIZATION STEP

1. From the given survey population (S) comprised of 10 data households; let the root node be designated as *S,* representing the total survey. Initialize the tree $T = \{S\}$ to have the node *S*, with the set of terminal nodes $\tilde{T} = \{S\}$, and the set of active terminal nodes $\tilde{T}_A = \{S\}$ as well (see Figure-4.13 for illustration).

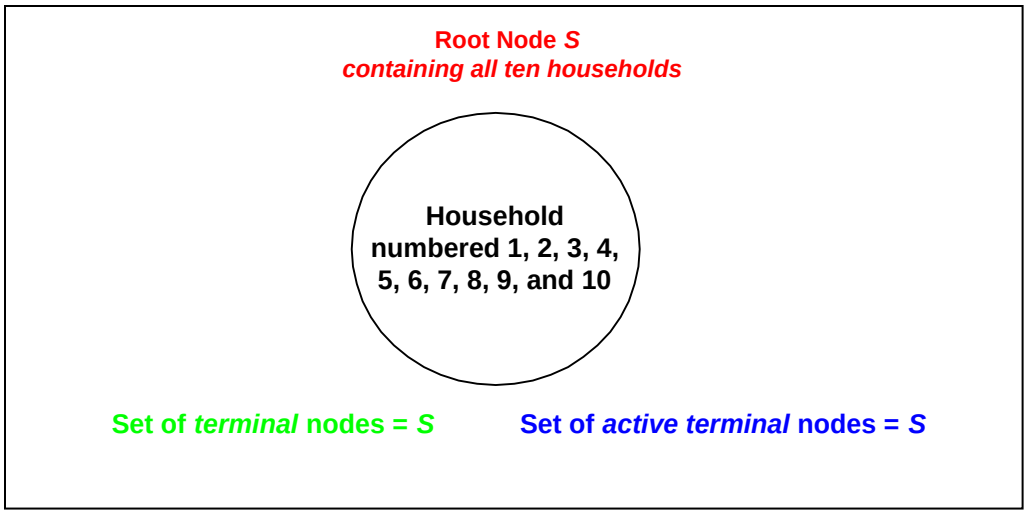


Figure-4.13: An illustration of the tree-growing at the first iteration

2. Specify the measures X_{ik} (value of the household demographic attribute X_k for the ith survey observation) for $k = 1, 2, 3$ and Y_{ij} (value of measure Y_j for the ith survey observation) for $j = 1, 2, 3$ obtained from Table-4.7. For example, the household numbered 1 is indexed as $i = 1$. Hence, $X_{11} = 1$, $X_{12} = 29$, $X_{13} = 1$, $Y_{11} = 918$, $Y_{12} = 417$, and $Y_{13} = 3$.

3. Let $N_{\min} = 2$.

4. Compute $D(S)$ as,

$$D(S) = \sum_{j=1}^{3} s_j D_j(S)$$

where $s_j = 1/\text{variance}\,(Y_j)$

$$= \frac{10-1}{\sum_{i=1}^{10}(Y_{ij} - \bar{Y}_j)^2}.$$

Hence, $D(S) = \sum_{j=1}^{3} \frac{10-1}{\sum_{i=1}^{10}(Y_{ij}-\bar{Y}_j)^2} \times \sum_{i=1}^{10}(Y_{ij}-\bar{Y}_j)^2$

$= \sum_{i=1}^{3}(10-1)$

$= 3(10\text{-}1)$

$= \mathbf{27}$.

Let, $\beta \equiv 0.01 D(S) \equiv 0.01(27) \equiv \mathbf{0.27}$.

5. Let DOWN $(S) \equiv 0$.

We then proceed to the Main Step.

MAIN STEP

1. From Figure-4.13, the set of **active** terminal nodes is **not** empty. Hence, proceed to the next step.
2. At the beginning, there is only **one active** terminal node containing all ten observations (see Figure-4.13). Hence, proceed to the next step.
3. Compute $\Delta_{kt}(N)$ for each k = 1 (HHSIZE), 2 (HHAGE), and 3 (WORKERS) and $t \in t_k(N)$ and find the largest such value: $\Delta_{k^*t^*}(N) = \max_{k,t}\{\Delta_{kt}(N)\}$.

In order to compute the decrement $\Delta_{kt}(N)$, we have to limit the possible values of t_k for each k. We assume that the possible values of HHSIZE ($t_1(N)$) is 1, 2, 3, and 4, the possible values of HHAGE ($t_2(N)$) is 18, 19, …, 90 years (this set of values could be further reduced by considering a more coarse discretization), and the possible values of WORKERS ($t_3(N)$) is 0, 1, 2, 3, and 4. For example, we select the split variable that yields the maximum decrement as **HHSIZE=1** $\in t_1(N)$ (i.e., $k^* = 1$ and $t^* = 1$). This is done here by intuition. However, it is usually selected after all the independent variables and their corresponding split values are tried out, and the one chosen is the variable with the split value that produces the maximum decrement in the total deviance. It would be confusing here to try all these possible combinations manually. However, let us assume that the HHSIZE variable provides the best decrement in the total deviance.

In particular, for the HHSIZE with a split value of 1, the children nodes $N_1 = N_{1kt}$ and $N_2 = N_{2kt}$ (where k = 1 and t = 1) contain households that have HHSIZE$\leq$1 and HHSIZE>1 respectively. The child node N_1 contains 5 households numbered 1, 2, 4, 5, and 7. The child node N_2 contains 5 households numbered 3, 6, 8, 9, and 10. We then compute the total deviance of these nodes by the following equations:

For the node N_{1kt}, $D_j(N_1) = \sum_{i=1,2,4,5,7}(Y_{ij}-\bar{Y}_{N_1 j})^2$ (1)

, where $\bar{Y}_{N_1 j} = \frac{\sum_{i=1,2,4,5,7} Y_{ij}}{5}$ (2)

$$\left.\begin{array}{lll} \text{Hence, for } j\text{=1 (T_Home)}, & \bar{Y}_{N_1\,1} = \dfrac{(918+1395+1620+1310+1560)}{5} & = 1360.6, \\ \text{for } j\text{=2 (T_Work)}, & \bar{Y}_{N_1\,2} = \dfrac{(417+0+0+0+0)}{5} & = 83.4, \text{ and} \\ \text{for } j\text{=3 (N_Nonhome)}, & \bar{Y}_{N_1\,3} = \dfrac{(3+2+0+2+1)}{5} & = 1.6. \end{array}\right\} \quad (3)$$

Substituting (3) into (1), we get

for j=1, $D_1(N_1) = \sum_{i=1,2,4,5,7} (Y_{i1} - 1360.6)^2$

$= (918\text{-}1360.6)^2 + (1395\text{-}1360.6)^2 + (1620\text{-}1360.6)^2 + (1310\text{-}1360.6)^2 + (1560\text{-}1360.6)^2$

$= 306687,$

for j=2, $D_2(N_1) = \sum_{i=1,2,4,5,7} (Y_{i2} - 83.4)^2$

$= (417\text{-}83.4)^2 + (0\text{-}83.4)^2 + (0\text{-}83.4)^2 + (0\text{-}83.4)^2 + (0\text{-}83.4)^2$

$= 139111,$ and

for j=3, $D_3(N_1) = \sum_{i=1,2,4,5,7} (Y_{i3} - 1.6)^2$.

$= (3\text{-}1.6)^2 + (2\text{-}1.6)^2 + (0\text{-}1.6)^2 + (2\text{-}1.6)^2 + (1\text{-}1.6)^2$

$= 5.2.$

Next, we compute

$$\begin{aligned} \boldsymbol{D(N_1)} &= \text{total deviance of node } N_1 \\ &= \sum_{j=1}^{3} s_j D_j(N_1) \qquad (4) \end{aligned}$$

where s_j = 1/variance (Y_j)

$$= \frac{n-1}{\sum_{i=1}^{n} (Y_{ij} - \bar{Y}_j)^2}.$$

Hence, for j=1, $s_1 = \dfrac{10-1}{\sum_{i=1}^{10} (Y_{ij} - 2113.3)^2}$

$$= \frac{9}{(918-2113.3)^2 + ... + (1765-2113.3)^2}$$

$= 1/1058381.$

Using the same procedure, we obtain s_2 = 1/214936, and s_3 = 1/34.

Substituting the values of s_j and $D_j(N_1)$ into Equation (4), we obtain

$$D(N_1) = \sum_{j=1}^{3} s_j D_j(N_1)$$

$$= \frac{306687}{1058381} + \frac{139111}{214936} + \frac{5.2}{34}$$

= **1.09**.

Using the same procedure, we obtain $D(N_2)$ = **7.62**.
The decrement $\Delta_{kt}(N)$ upon partitioning the current node N (containing all the ten households) into the above two subnodes $N_1 = N_{1kt}$ and $N_2 = N_{2kt}$ based on the split value $t = 1$ for $k = 1$ (HHSIZE) is computed according to:

$$\Delta_{kt}(N) = D(S) - [D(N_1) + D(N_2)]$$
$$= 27 - (1.09 + 7.62) = \mathbf{18.29}.$$

The HHSIZE variable at split level $t = 1$ provides the maximum decrement in the total deviance (18.29) as compared with the decrements obtained by considering all the other variables with their corresponding split values.

4. Because the number of observations in each of the subnodes N_1 and N_2 is 5, which is greater than the minimal value $N_{\min}$= 2, then continue this partitioning by proceeding to Step 5.

5. Partition the current node $\{S\}$ containing all the ten households into N_1 (having HHSIZE=1) and N_2 (having HHSIZE>1), and add N_1 and N_2 to the tree T with DOWN(N_1) = DOWN(N_2) = $\{S\}$.

6. Compute and store $D(N_1)$ and $D(N_2)$ using the same procedure described above in Equations (1) through (4). Remove the node $N = \{S\}$ from the set of **terminal** nodes and the set of **active terminal** nodes, and add N_1 and N_2 to the set of **terminal** nodes. Because $D(N_1) = 1.09$ and $D(N_2) = 7.62$, which are both more than $\beta = 0.27$, we also add both N_1 and N_2 to the set of **active terminal** nodes. Repeat the Main Step. Figure-4.14 shows the updated tree at this step.

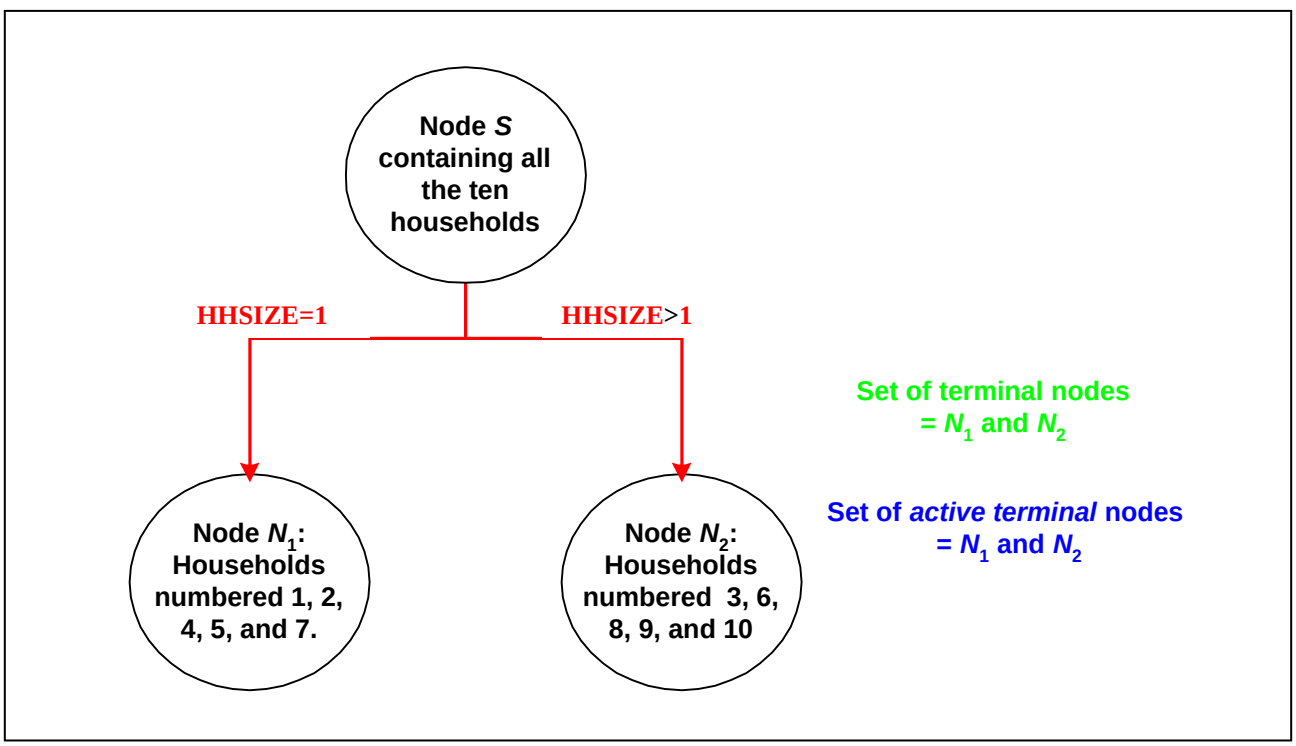


Figure-4.14: The updated tree after finishing the first iteration

Figure-4.15 shows the final result of the classification and regression tree corresponding to the given ten households.

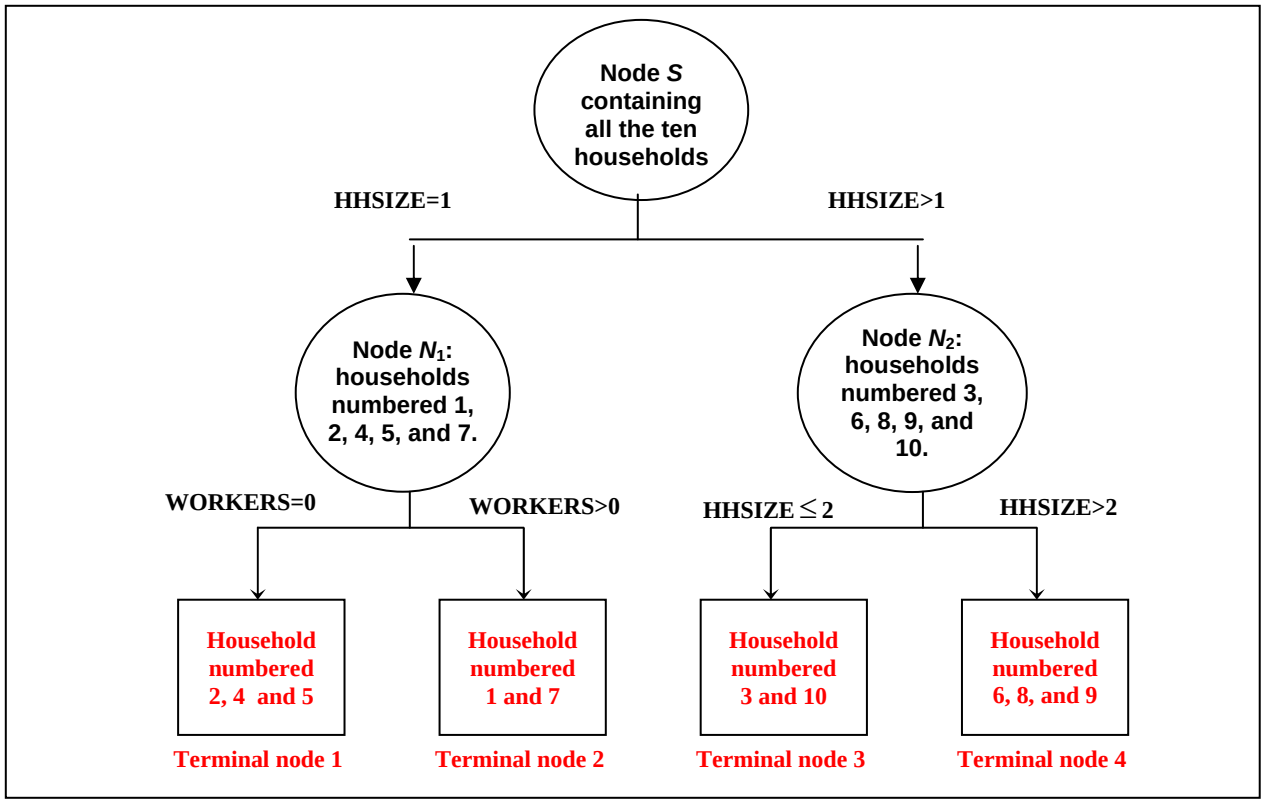


Figure-4.15: A final result of the classification and regression tree corresponding to the given 10 households

After performing the first split using the split variable HHSIZE=1, the next splits occur for WORKER=0 and HHSIZE=2 leading to a final tree having 4 terminal nodes as shown in Figure-4.15.

The first terminal node contains 3 households; numbered 2, 4 and 5, having HHSIZE=1 and WORKERS=0. The second terminal node contains 2 households; numbered 1 and 7, having HHSIZE=1 and WORKERS>0. The third terminal node contains 2 households; numbered 3 and 10, having HHSIZE=2. The fourth terminal node contains 3 households; numbered 6, 8, and 9, having HHSIZE>2. No further splitting is possible because the end nodes contain the minimum number of observations. Any further splitting would violate the $N_{\min}$ requirement. The total deviance of the terminal nodes D(TN1), D(TN2), D(TN3) and D(TN4) are calculated as shown below, and the decrement $\Delta_{kt}(N)$ upon partitioning the nodes *N1 and N2* into the above four terminal nodes is also shown. Further splitting is possible on terminal nodes 2, 3, and 4 based on the total deviance in the node since it is greater than β which is equal to 0.27. However the $N_{\min}$ requirement would be violated.

Terminal Node 1 (TN1)

		j=1	j=2	j=3
	Household Number	T-Home	T-Work	N-Nonhome
		Y_1	Y_2	Y_3
	2	1395	0	2
	4	1620	0	0
	5	1310	0	2
		$\overline{Y}_1 = 1441.666$	$\overline{Y}_2 = 0$	$\overline{Y}_3 = 1.333$
$D_j\ TN_1$	$\sum_{i=1}^{3}\left(Y_i - \overline{Y}\right)^2$	51315.08	0	2.665
	S_j	$\frac{1}{1058381}$	$\frac{1}{214936}$	$\frac{1}{34}$

Terminal Node 2 (TN2)

		j=1	j=2	j=3
	Household Number	T-Home	T-Work	N-Nonhome
		Y_1	Y_2	Y_3
	1	918	417	3
	7	1560	0	1
		$\overline{Y}_1 = 1239$	$\overline{Y}_2 = 208.5$	$\overline{Y}_3 = 2$
$D_j\ TN_2$	$\sum_{i=1}^{2}\left(Y_i - \overline{Y}\right)^2$	206082	86945	2.0
	S_j	$\frac{1}{1058381}$	$\frac{1}{214936}$	$\frac{1}{34}$

$$D(TN_1) = \frac{51315.08}{1058381} + \frac{0}{214936} + \frac{2.665}{34}$$

$$= 0.048 + 0 + 0.078$$

$$= 0.126$$

$$D(TN_2) = \frac{206082}{1058381} + \frac{86945}{214936} + \frac{2.0}{34}$$

$$= 0.195 + 0.404 + 0.059$$

$$= 0.658$$

$$\Delta_{kt}(N_1) = D(N_1) - [(D(TN_1) + D(TN_2)] = 1.09 - (0.126 + 0.658) = 0.306$$

Terminal Node 3 (TN3)

		j=1	j=2	j=3
	Household Number	T-Home	T-Work	N-Nonhome
		Y_1	Y_2	Y_3
	3	2275	385	11
	10	1765	1145	10
		$\overline{Y}_1 = 2020$	$\overline{Y}_2 = 765$	$\overline{Y}_3 = 10.5$
$D_j\ TN_3$	$\sum_{i=1}^{2}\left(Y_i - \overline{Y}\right)^2$	13005.0	288800	0.5
	S_j	$\frac{1}{1058381}$	$\frac{1}{214936}$	$\frac{1}{34}$

Terminal Node 4 (TN4)

		j=1	j=2	j=3
	Household Number	T-Home	T-Work	N-Nonhome
		Y_1	Y_2	Y_3
	6	2970	770	14
	8	4260	475	16
	9	3060	1180	6
		$\overline{Y}_1 = 3430$	$\overline{Y}_3 = 808.33$	$\overline{Y}_3 = 12$
$D_j\ TN_2$	$\sum_{i=1}^{3}\left(Y - \overline{Y}\right)^2$	1,037,400	250717	56
	S_j	$\frac{1}{1058381}$	$\frac{1}{214936}$	$\frac{1}{34}$

$$D(TN_3) = \frac{130050}{1058381} + \frac{288800}{214936} + \frac{.05}{34}$$

$$= 0.123 + 1.343 + 0.0147$$

$$= 1.43$$

$$D(TN_4) = \frac{1037460}{1058381} + \frac{250717}{214936} + \frac{56}{34}$$

$$= 0.980 + 1.17 + 1.647$$

$$= 3.797$$

$$\Delta_{kt}(N_2) = D(N_2) - \left[\left(D(TN_3) + D(TN_4)\right)\right] = 7.62 - (1.43 + 3.797) = 2.393$$

The next example illustrates the tree-growing step using a much larger sample size (3,470 households), implemented by using the CART-VT program.

Example for the tree-growing step using the CART-VT program

The data used in this example came from the Portland, Oregon Activity and Travel Survey of 1994/95 (see survey forms in *Appendices A* and *B*). In the Portland survey, all respondents in the sample households were asked to record in a diary their activity and travel behavior for two consecutive days, that is, for a 48-hour period, with the pairs of days being staggered across the 7 days of the week. The diary information was retrieved via a subsequent telephone interview and approximately 92% of the activity locations were successfully geocoded.

The Portland Activity and Travel Survey was conducted in the spring of 1994 and the fall/winter of 1994/95. The survey resulted in 4,451 completed household surveys for a total of 10,048 individuals. The two-day sampling resulted in 20,096 daily activity-travel patterns distributed across the seven days of the week. The survey data showed that 48.3% of the individuals are male with the average age of the individual being 38.21 years. Employment is at 54% (either full-time, part-time, full-time self employed, or part-time self employed) with 8% working at home and 5% working two or more jobs. 24% of the respondents are either full or part-time students.

Seven dependent variables (Y_j, j = 1, …, 7) were selected to represent the activity travel pattern of each household as given below:

Y_1 = total household duration in minutes for in-home activity (THOME),
Y_2 = total household duration in minutes for work (TWORK),
Y_3 = total household duration in minutes for shop (TSHOP),
Y_4 = total household duration in minutes for school (TSCHOOL),
Y_5 = total household duration in minutes for visit (TVISIT),
Y_6 = total household duration in minutes for other activities (TOTHER), and
Y_7 = number of non-home based trips (NONHOME).

To model the activity travel pattern of an average or typical **week day** and to maintain the independence of the observations, only the **first day's** activity information was used from households whose first diary-day fell on a week day. This limited the sample to 3,505 households. The final sample, excluding those with missing data, includes 3,470 households (i = 1, …, 3470).

Appended to the dependent variables were a set of nine independent demographic variables (X_k, k = 1,…, 9) defined as follows:

X_1 = number of persons in household (HHSIZE),
X_2 = number of vehicles in household (VEHICLES),
X_3 = total income in $K (INCOME),
X_4 = number in household with age less than 5 years (ALT5),
X_5 = number in household with age 5 to 17 (A5TO17),
X_6 = number in household with age 26 to 45 (A26TO45),
X_7 = age of the head of the household (HHAGE),
X_8 = number of employed household members (WORKERS), and
X_9 = residential density of block group in housing units/acre (HDENSTY).

INITIALIZATION STEP

1. From the given survey population (S) comprised of 3,470 data households, let the root node be designated as S representing the total survey. Initialize the tree $T = \{S\}$ to have the node S, with the set of terminal nodes $\tilde{T} = \{S\}$, and the set of active terminal nodes $\tilde{T}_A = \{S\}$ as well.

2. Specify the measures X_{ik} (value of the household demographic attribute X_k for the ith survey observation) for k = 1, …, 9, obtained from the **survey household demographic data** (see an example as shown in Figure-4.16).

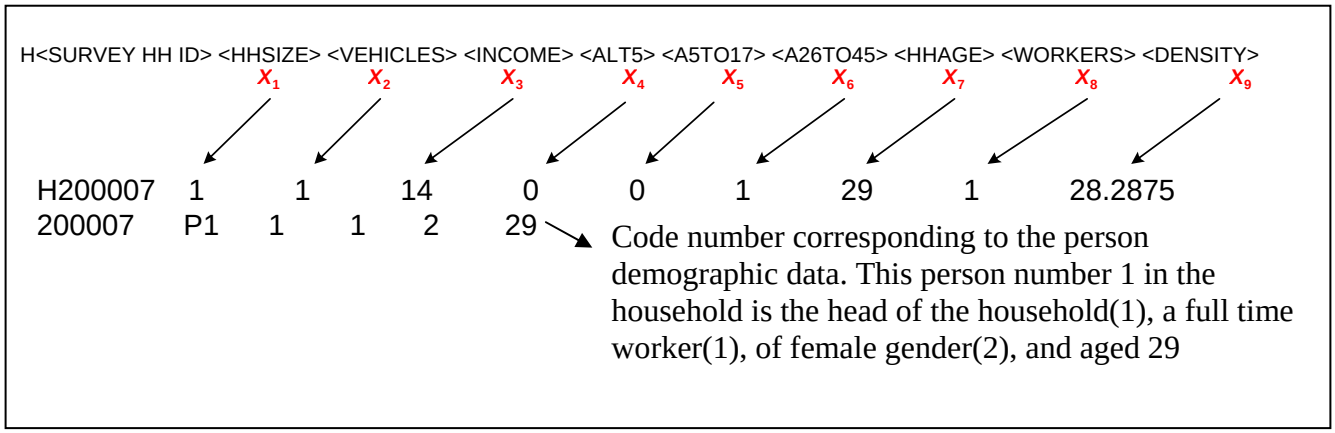


Figure-4.16: An example of a survey household demographic data for the household numbered 200007

From Figure-4.16, the household numbered 200007 is the **first** household in the survey. Hence, X_{11} = 1, X_{12} = 1, X_{13} = 14, X_{14} = 0, X_{15} = 0, X_{16} = 1, X_{17} = 29, X_{18} = 1, and X_{19} = 28.2875.

Then we specify the dependent variables Y_{ij} (value of measure Y_j for the ith survey observation) for j = 1, …, 7. The information on the Y_j-variables are obtained from the skeletal activity patterns (see an example as shown in Figure-4.17).

Used for calculating Y_j

SAMPNO	PERSNO	ACTNO	ACTID	AT_HOME	WUTHERE	MODE	DRIVER	NUMVEH	ACTSTART	ACTEND	GEOX	GEOY
200007	1	0	0 (home)	1	2	1	0	0	0	708	7639843	686973.5625
		1	1 (work)	2	2	2	1	1	708	870	7642798	682822.375
		2	1 (work)	2	2	2	1	1	885	1140	1	12
		3	5 (other)	2	2	2	1	1	1185	1365	7655483.5	681557.625
		4	0 (home)	1	2	2	1	1	1410	1620	7639843	686973.5625

Figure-4.17: An example of the skeletal activity pattern for the survey household numbered 200007

From the example shown in Figure-4.17, the household numbered 200007 has only three types of activities (obtained from the **ACTID** column), which are *in-home activity, work,* and *other*. Hence, Y_3 (duration for *shop*), Y_4 (duration for *school*), and Y_5 (duration for *visit*) are equal to zero. Y_1, Y_2, and Y_6 are calculated from the summary of the actual **ending time** minus the actual **starting time** for each type of activity. Hence,

$$Y_1 \text{ (home)} = (708 - 0) + (1620 - 1410) = 918 \text{ minutes,}$$
$$Y_2 \text{ (work)} = (870 - 708) + (1140 - 885) = 417 \text{ minutes, and}$$
$$Y_6 \text{ (other)} = (1365 - 1185) = 180 \text{ minutes.}$$

Y_7 (number of non-home based trips) here is = 3 (obtained from the **ACTID** column).

After specifying the variables X and Y for every survey household, we then proceed to Step 3.

3. Let $N_{\min} = 50$.

4. Compute $D(S)$ by,

$$D(S) = \sum_{j=1}^{7} s_j D_j(S)$$

where $s_j = 1/\text{variance}\ (Y_j)$

$$= \frac{3470-1}{\sum_{i=1}^{3470}(Y_{ij} - \bar{Y}_j)^2}.$$

Hence, $D(S) = \sum_{j=1}^{7} \frac{3470-1}{\sum_{i=1}^{3470}(Y_{ij} - \bar{Y}_j)^2} \times \sum_{i=1}^{3470}(Y_{ij} - \bar{Y}_j)^2$

$$= \sum_{i=1}^{7}(3470-1)$$

$= 7(3470\text{-}1)$

$=$ **24283**.

Let, $\beta \equiv 0.01 D(S) \equiv 0.01(24283)$

$\equiv$ **242.83**.

5. Let DOWN $(S) \equiv 0$.

We then proceed to the Main Step using the CART-VT program, where the above values of $N_{\min}$= 50 and Minimum Deviance = 0.01 are used as inputs to the program.

MAIN STEP

We have implemented the tree-growing step in the CART-VT program, which is coded according to the Main Step as mentioned earlier. The output produced the classification and regression tree that is shown in Figure-4.18.

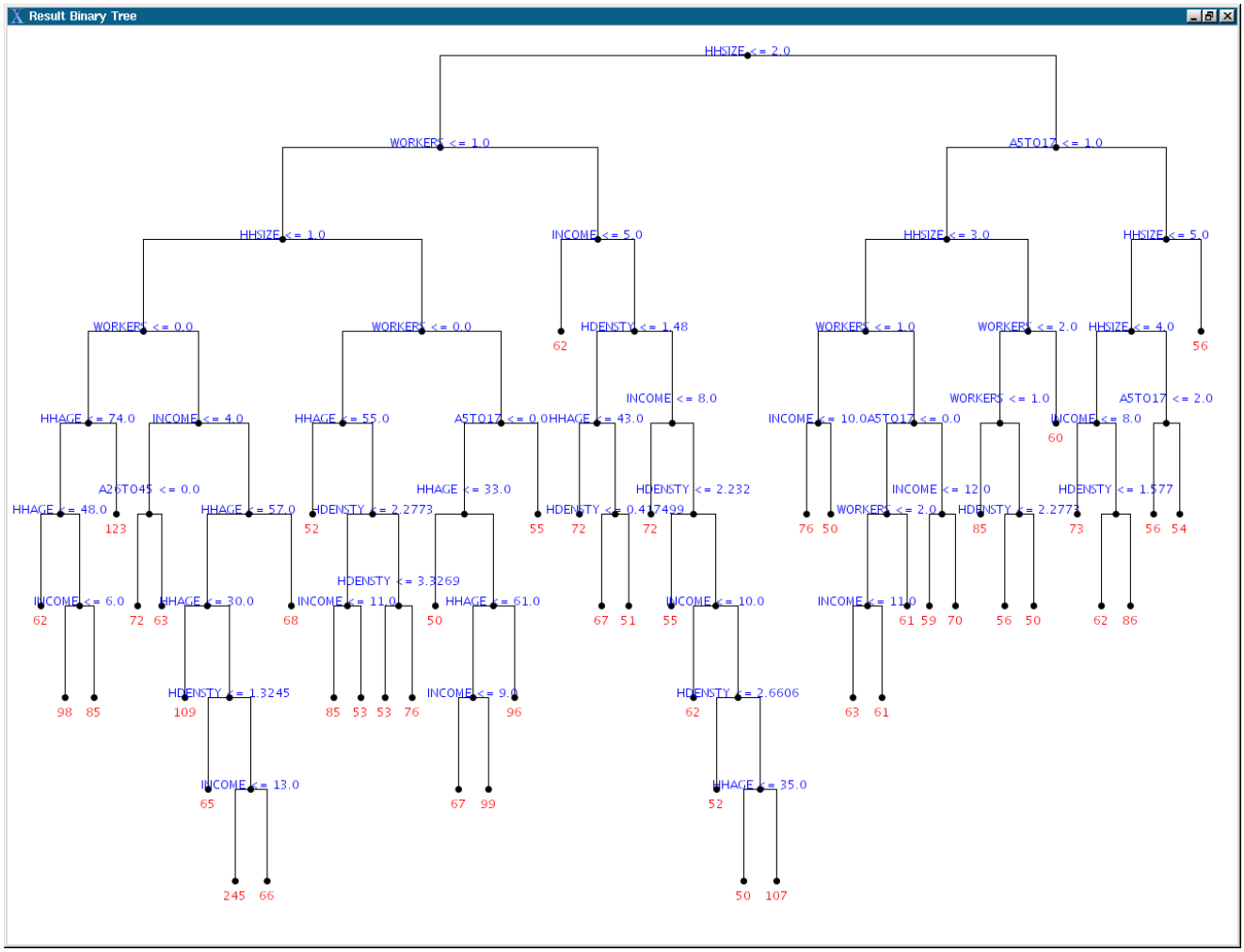


Figure-4.18: A classification and regression tree corresponding to the survey data from Portland, Oregon

Note that each number beneath the terminal node indicates number of observations classified within the node. There are 48 terminal nodes. The demographic variables (X_k) used in the tree are:

X_1 = number of persons in household (HHSIZE),
X_3 = total income in $K (INCOME),
X_5 = number in household with age 5 to 17 (A5TO17),
X_6 = number in household with age 26 to 45 (A26TO45),
X_7 = age of the head of the household (HHAGE),
X_8 = number of employed household members (WORKERS), and
X_9 = residential density of block group in housing units/acre (HDENSTY).

Note that X_2 (number of vehicles in household, VEHICLES), and X_4 (number in household with age less than 5 years, ALT5) are not used here because they **do not significantly** affect behavior of the survey households. Initially, the demographic variable used for the **first** split is $\boldsymbol{X_1}$, **household size (HHSIZE)** with the split value of 2. Hence, X_1 is the most important variable that affects the household behavior. The second most important demographic variables are $\boldsymbol{X_8}$ **(WORKERS)** and $\boldsymbol{X_5}$ **(A5TO17)**. The other variables, $\boldsymbol{X_3}$ **(INCOME),** $\boldsymbol{X_9}$ **(HDENSTY),** $\boldsymbol{X_7}$ **(HHAGE)**, and $\boldsymbol{X_6}$ **(A26TO45)** are relatively less important.

Generally, the tree may create several terminal nodes from the data. Some demographic variables (X_k) should not be used. This can be partly observed by the **number of observations** or **the deviance** of the terminal nodes corresponding to the split variable (X_k). Some terminal nodes contain only 50 observations, which is the least acceptable value ($N_{\min}$), or have a small deviance, compared to the other terminal nodes, which have more than hundred observations, or have higher deviances. The terminal nodes that contain a small amount of observations, or have relatively small deviances should be considered for pruning as described in the following section.

4.8.2.2 Pruning the tree

The binary tree produced by the tree-growing algorithm generally overfits the data in that there are more independent variables (X_k) than needed. Several proposals have been made to determine a more manageable reduced tree. One common way to assess how well a (reduced) tree might represent the data is to use it to predict *a new set of data* or *a partitioned set of the existing data*. In this case, deviance is replaced by a sum-squared prediction error. The best subtree is the one that has a minimum prediction error.

The general algorithm CART recommends two alternative schemes for reducing or *pruning* (or *pruning upward*) the tree *T*, namely, the *cross-validation* approach and the *independent test sample* approach. TRANSIMS implements a variant of the former scheme, which is the one recommended by Breiman et al. (1984). The two approaches are described below.

4.8.2.2.1 Cross-validation step for reducing or pruning the tree

To describe the cross-validation method, suppose that the survey sample *S* is randomly partitioned into 10 roughly equal subsets $S_1,...,S_{10}$. (This number 10 is recommended by Breiman et al., and is used by TRANSIMS as well.) Having done this, nine out of these ten subsets are grouped together in turn, and are used to build a corresponding classification tree and this tree is then evaluated with respect to how well it fits the last subset as well. Of these ten trees tested in this manner, the one that provides the best overall fit (yields the minimum prediction error) is prescribed for implementation. Presumably, because of the smaller sample size, each of these ten trees will tend to be smaller than the tree *T* obtained using the full sample *S*.

Mathematical Terminology and Definition

$\boldsymbol{v = 1, \ldots, 10}$: index for the partitioned data sets.
S : the set of all observations
$S_v \subseteq S$: a subset of S

$S^{(v)} = S - S_v$ (In other words, S is split into S_v and $S^{(v)}$ (built set)).

$T^{(v)}$ = the tree constructed using CART algorithm based on the (tree growing) observations in $S^{(v)}$ alone.

$\tilde{T}^{(v)}(i)$ = the terminal node of $T^{(v)}$ where the (test) observation *i* would fall into.

$\overline{Y}_{\tilde{T}^{(v)}(i),j}$ = the average of value of the measure *j* (Y_{kj}) for the (tree growing) observations *k* in $S^{(v)}$ that reside in the terminal node $\tilde{T}^{(v)}(i)$:

$$\overline{Y}_{\tilde{T}^{(v)}(i),j} = \frac{\sum_{k \in \tilde{T}^{(v)}(i)} Y_{kj}}{\left|\tilde{T}^{(v)}(i)\right|}$$

$e_i(T^{(v)})$ = a *prediction error* for the (test) observation *i* with respect to the tree $T^{(v)}$:

$$e_i(T^{(v)}) = \sum_{j=1}^{p} s_j (Y_{ij} - \overline{Y}_{\tilde{T}^{(v)}(i),j})$$

For cross validation method, the prediction error of a tree is defined by the sum of prediction errors of all observations:

$$e(T^{(v)}) = \sum_{i \in S} e_i(T^{(v)})$$

INITIALIZATION STEP

1. From the given survey sample *S* comprised of the *n* household data observations, randomly partition *S* into ten roughly equal subsets $S_1, ..., S_{10}$.
2. Let the set $S^{(v)}$ be comprised of the nine subsets excluding S_v:

   $S^{(v)} = S - S_v$ for each v = 1, …, 10.

   Hence, each $S^{(v)}$ is comprised of some 90% of the overall sample size, excluding the particular 10% of surveys designated by S_v.

MAIN STEP

1. Build the tree $T^{(v)}$ based on the observations in $S^{(v)}$ alone, for each v = 1, …, 10, via the algorithm described in Section 8.2.1: Tree-growing step.

2. For each tree $\tau \in \{T^{(v)}, v=1,...,10\}$, compute the *prediction error* $e(\tau)$ with respect to the tree τ.

3. Then, the *cross-validation method* in TRANSIMS selects a tree $T^{(v*)}$ (out of the ten resulting trees) that minimizes the *total* prediction error:

   $$T^{(v*)} \in \text{argmin } \{e(T^{(v)}), v = 1, \ldots, 10\}.$$

   This tree is prescribed as the final choice for a tree that classifies the given survey data.

Example for the cross-validation step for reducing or pruning the tree

Suppose that we randomly partition the Portland, Oregon survey sample *S,* comprised of the **3470** household data observations, into ten roughly equal subsets $S_1,...,S_{10}$. Let set $S^{(v)}$ be comprised of the 9 subsets excluding S_v:

$$S^{(v)} = S - S_v \text{ for each } v = 1, \ldots, 10.$$

Hence, each $S^{(v)}$ is comprised of some 90% of the overall sample size (3123 households), excluding the particular 10% (347 households) of surveys designated by S_v.

Build the tree $T^{(v)}$ based on the observations in $S^{(v)}$ alone, for each v = 1, …, 10, via the algorithm described in Section 8.2.1: Tree-growing step. For each tree $\tau \in \{T^{(v)}, v=1,...,10\}$, compute the *prediction error* $e(\tau)$ with respect to the tree τ. Table-4.8 shows values of the total prediction errors for the ten trees.

Table-4.8: Total prediction errors for the ten trees

Tree $T^{(v)}$	Total Prediction Error
$T^{(1)}$	15490.51
$T^{(2)}$	15438.31
$T^{(3)}$	15394.38
$T^{(4)}$	15415.22
$T^{(5)}$	15391.04*
$T^{(6)}$	15403.63
$T^{(7)}$	15455.61
$T^{(8)}$	15469.89
$T^{(9)}$	15423.20
$T^{(10)}$	15507.35

Then, the *cross-validation method* in TRANSIMS selects a tree $T^{(v^*)}$ (out of the ten resulting trees) that minimizes the *total* prediction error, which is the **fifth tree** ($T^{(5)}$). This tree is prescribed as the final choice for a tree that classifies the given survey data. Figure-4.19 shows the pruned tree.

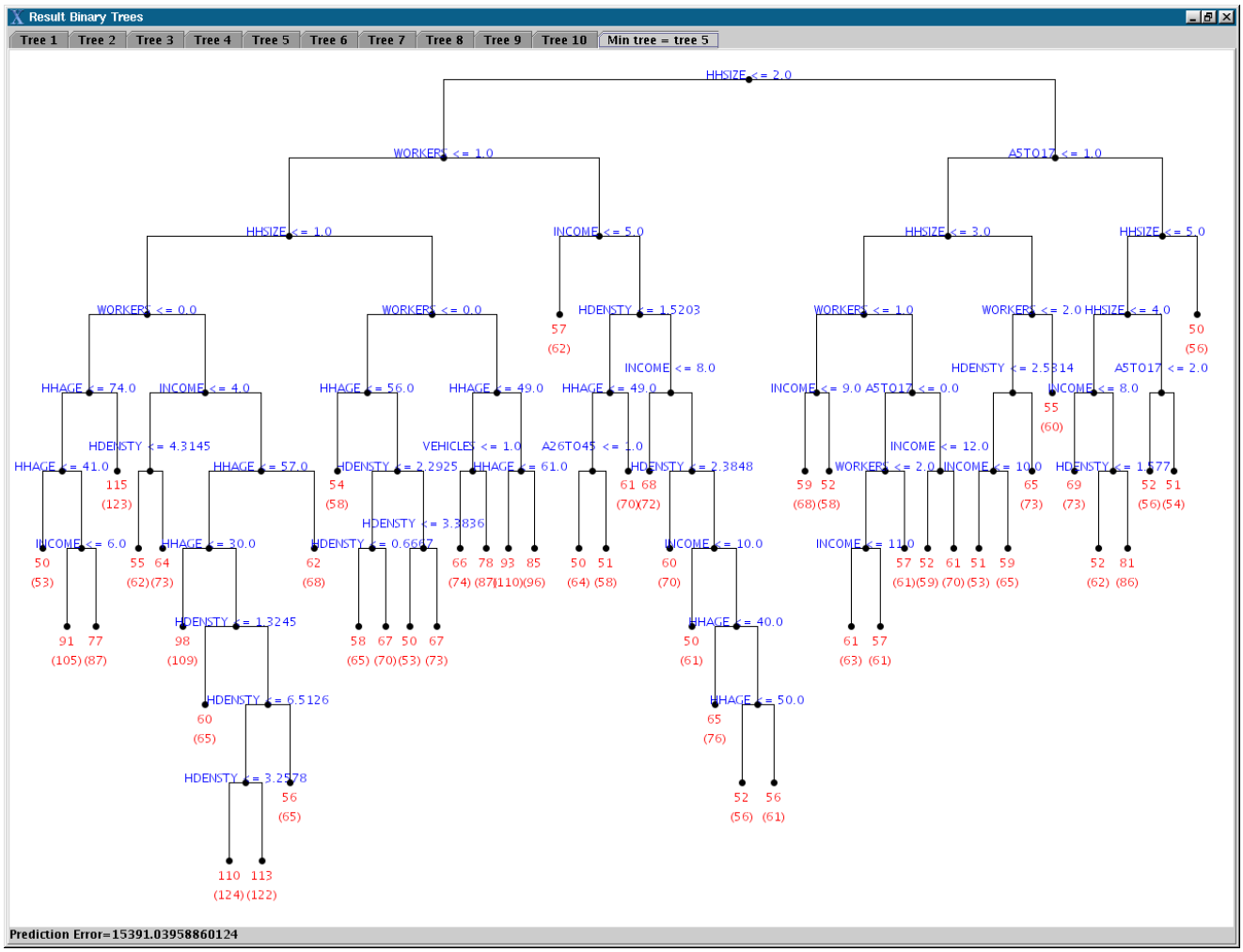


Figure-4.19: The pruned tree corresponding to the survey data from Portland, Oregon using cross-validation approach

Note that the pruned tree has 48 terminal nodes, which is the same size as the original tree.

The cross validation approach is quite computationally intensive because it has to create 10 different trees from the data directly. If the data size is very large, this approach will take a very long time to process. Besides, the tree will not be pruned very much as we can see in the example above. Therefore, another approach is recommended when a large sample size is used.

4.8.2.2.2 Independent Test Sample step for reducing or pruning the tree

This alternative approach to the cross-validation method is more suited for large sample sizes, but is not currently a part of TRANSIMS. To describe this method, we first have to define some mathematical terminology and definition.

Mathematical Terminology and Definition

S : the set of all observations
$S_v \subseteq S$: a subset of S
$S^{(v)} = S - S_v$ (In other words, S is split into S_v (test set) and $S^{(v)}$ (built set)).
$T^{(v)}$ = the tree constructed using CART algorithm based on the (tree growing) observations in $S^{(v)}$ alone.
$\tilde{T}^{(v)}(i)$ = the terminal node of $T^{(v)}$ where the (test) observation *i* would fall.
$\overline{Y}_{\tilde{T}^{(v)}(i),j}$ = the average of value of the measure *j* (Y_{kj}) for the (tree growing) observations *k* in $S^{(v)}$ that reside in the terminal node $\tilde{T}^{(v)}(i)$:

$$\overline{Y}_{\tilde{T}^{(v)}(i),j} = \frac{\sum_{k \in \tilde{T}^{(v)}(i)} Y_{kj}}{\left|\tilde{T}^{(v)}(i)\right|}$$

$e_i(T^{(v)})$ = a *prediction error* for the (test) observation *i* with respect to the tree $T^{(v)}$:

$$e_i(T^{(v)}) = \sum_{j=1}^{p} s_j (Y_{ij} - \overline{Y}_{\tilde{T}^{(v)}(i),j})$$

For the independent test sample method, the prediction error of a tree is defined by the sum of prediction errors of the observations in the test set:

$$e_{S_v}(T^{(v)}) = \sum_{i \in S_v} e_i(T^{(v)})$$

$D(\tau)$: deviance measure defined via (1)-(2) for any tree τ as the total deviance over its terminal nodes (denoted $\tilde{\tau}$) given by

$$D(\tau) = \sum_{N \in \tilde{\tau}} D(N). \qquad (3)$$

$D_\alpha(\tau)$: *penalty function* that augments the deviance $D(\tau)$ with a penalty term given by some penalty parameter $\alpha > 0$ multiplied by the number of terminal nodes $|\tilde{\tau}|$ in τ. Hence,

$$D_\alpha(\tau) = D(\tau) + \alpha\,|\tilde{\tau}|. \qquad (4)$$

$\tau_N \subseteq \tau$: *subtree rooted at node N* of τ given by *N* and its *descendents* (or *successors*) *within* τ.

Given any tree τ (where the initial τ equals the original tree T) and a node N of τ, where $N \notin \tilde{\tau}$, and given any parameter α, we can compute the penalty function value similar to (4) for the *node* N and for the *subtree* τ_N that is rooted at N as follows

$$D_\alpha(N) = D(N) + \alpha \tag{5a}$$

$$D_\alpha(\tau_N) = D(\tau_N) + \alpha\,|\tilde{t}_N|. \tag{5b}$$

Note that from (5 a, b), we will have

$$D_\alpha(\tau_N) < D_\alpha(N) \text{ so long as } \alpha < \alpha(N) \equiv \frac{D(N) - D(\tau_N)}{|\tilde{t}_N| - 1}. \tag{6}$$

In other words, for any $\alpha \geq \alpha(N)$, using the penalty function (4), we would prefer collapsing the subtree τ_N rooted at N into this node N. Suppose that we compute this value $\alpha(N)$ given by (6) for all the nonterminal nodes N of τ, and find

$$N^* \in \arg\min \{\alpha(N) : N \text{ is a node of } \tau,\ N \notin \tilde{\tau}\}. \tag{7}$$

Given any tree τ (where the initial τ equals the original tree T) a node N^* is determined that provides the *weakest junction* of τ in the sense that among the nodes $N \notin \tilde{\tau}, N^*$ yields the smallest value of the penalty parameter α beyond which it becomes preferable to collapse the subtree rooted τ_{N^*} at the node N^* with respect to the merit function (4).

This establishes an inductive process whereby after finding N^* and collapsing τ_{N^*} into N^*, the process is repeated with the *pruned* tree given by $\tau - \tau_{N^*}$. This produces a sequence of *nested trees*, leading ultimately to the single root node r, say, of the original tree.

The independent test sample can be described as follows. First partition the survey sample set S into the sets S_1 and S_2, where S_1 is used for growing a tree T_1, and S_2 is the *independent test sample* that can subsequently be used to select a tree among the nested trees derived from T_1 based on a suitable error prediction function. Figure-4.20a displays the flowchart for this approach.

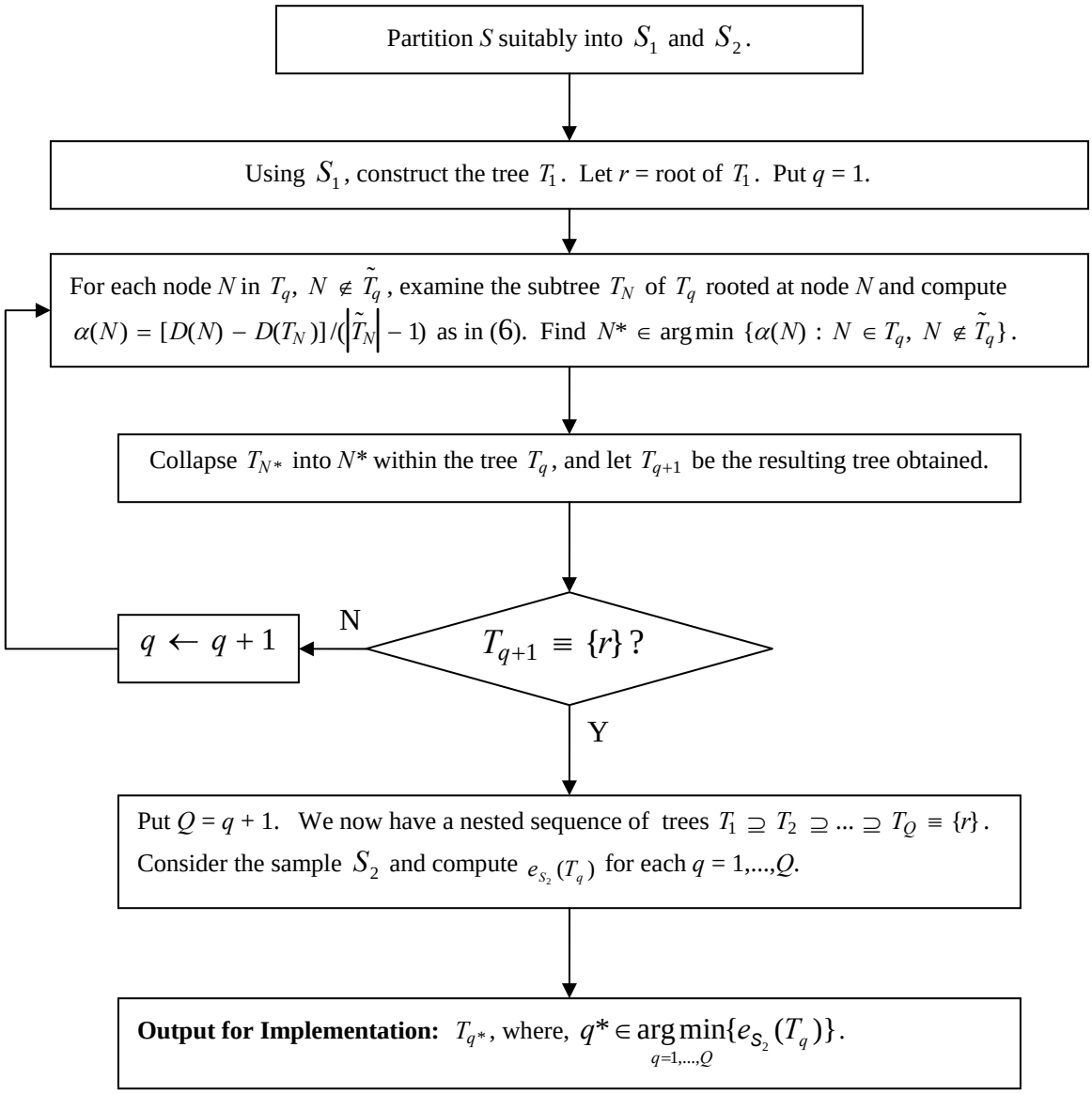


Figure-4.20a: Flowchart for the Pruning using the Independent Test Sample Approach.

Moreover, in any case, note that if S_1 is a sufficiently representative sample, and if T_1 is an elaborately expanded classification tree, then it is quite likely that $q^* = 1$ by the procedure of Figure 4-20a, leading to no pruning action.

Example for the independent test sample step for reducing or pruning the tree

Suppose that we randomly partition the Portland, Oregon survey sample *S,* comprised of the **3470** household data observations, into two parts S_1 and S_2, where S_1 is comprised of some 90% of the overall sample size (3123 households), and S_2 is the remaining 10% of the surveys (347 households).

Build the nested sequence of trees $T_1 \supseteq T_2 \supseteq \ldots \supseteq T_Q \equiv \{r\}$ based on the algorithm given in Figure 4-20a. Compute the *prediction error* $e_{S_2}(T_q)$ for each tree in the nested sequence. Figure-4.20b shows plot of the prediction errors of the nested trees. The algorithm select T_{72} as the output, which is shown in Figure-4.20c.

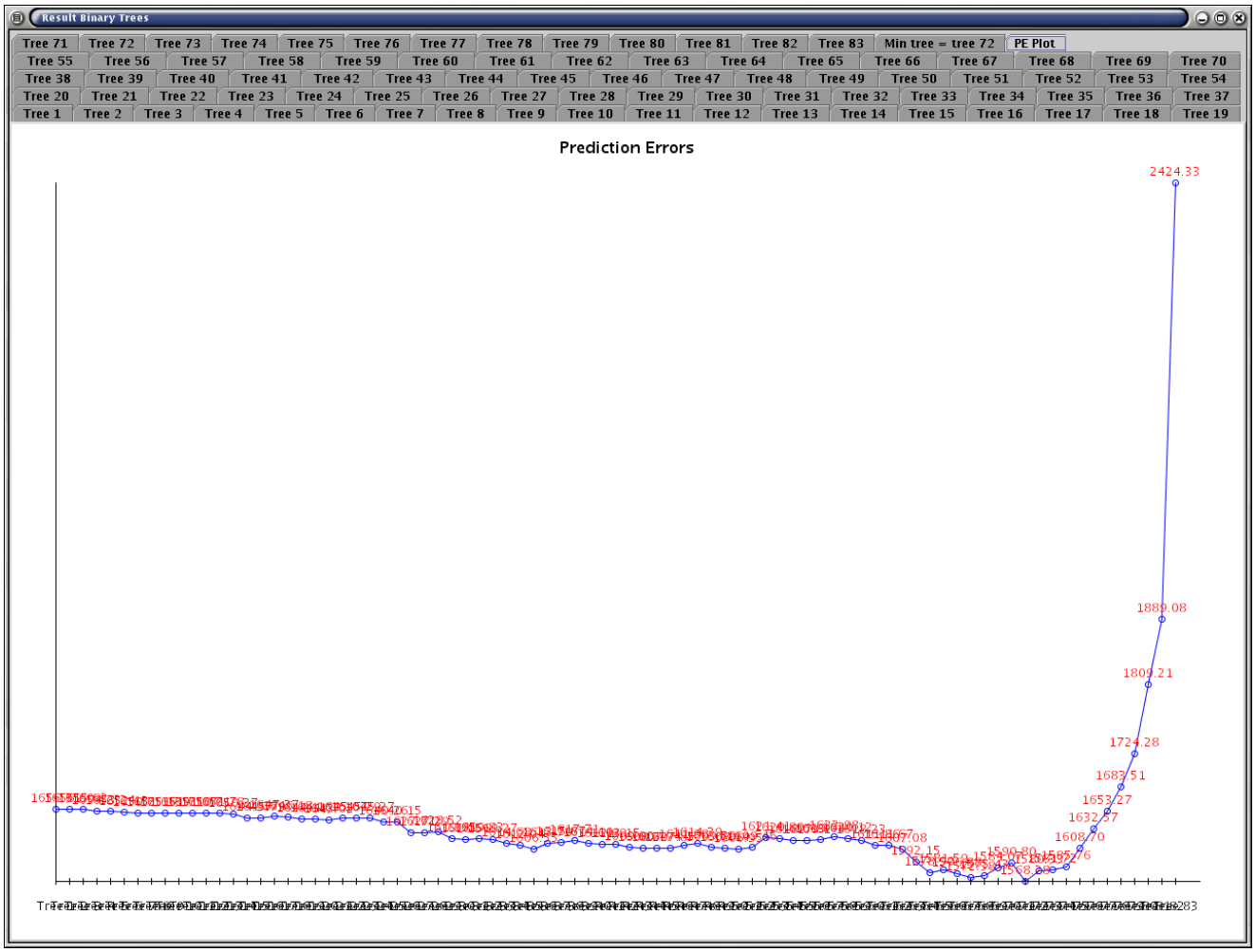


Figure-4.20b: Prediction error plot

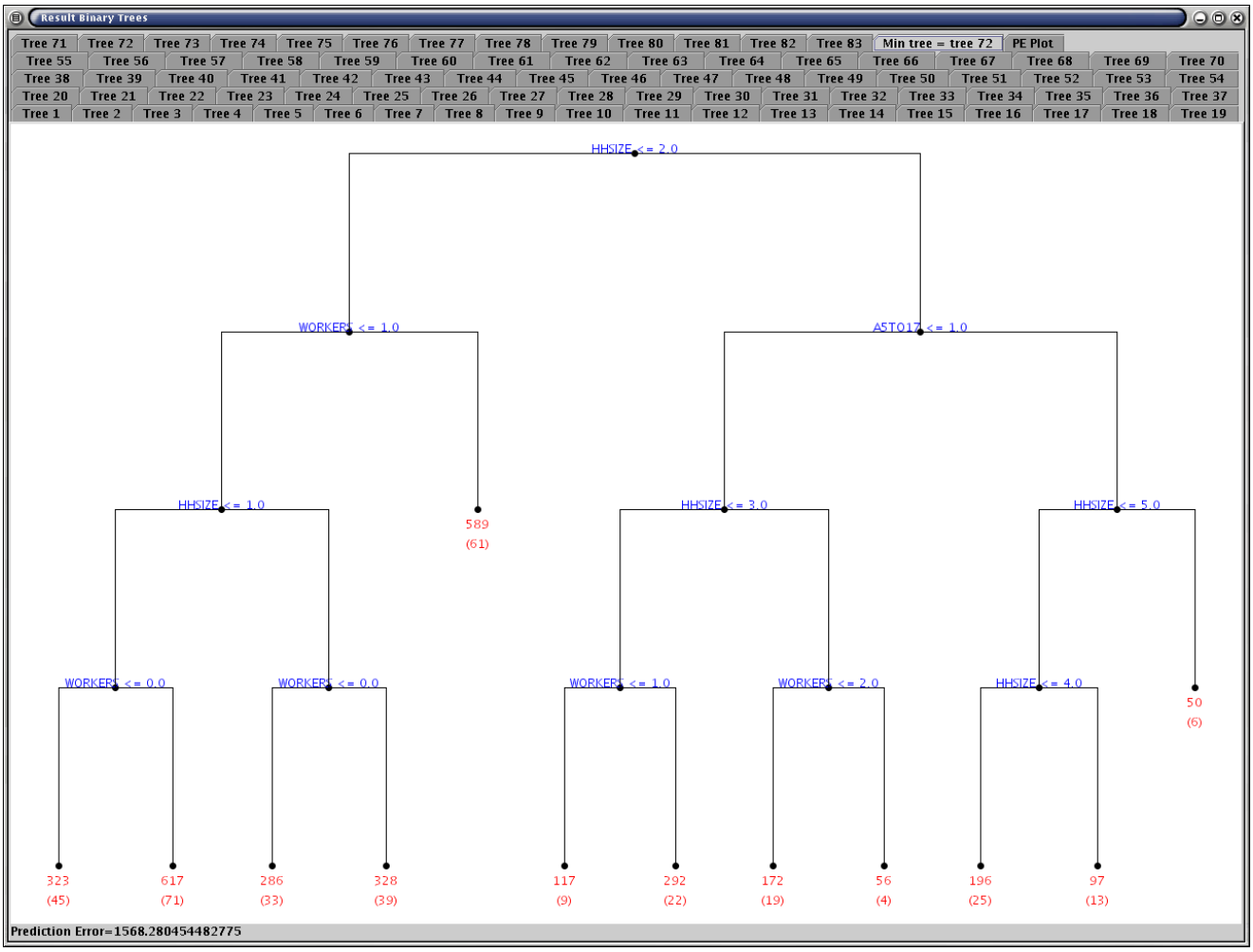


Figure-4.20c: The pruned tree corresponding to the survey data from Portland, Oregon using independent test approach

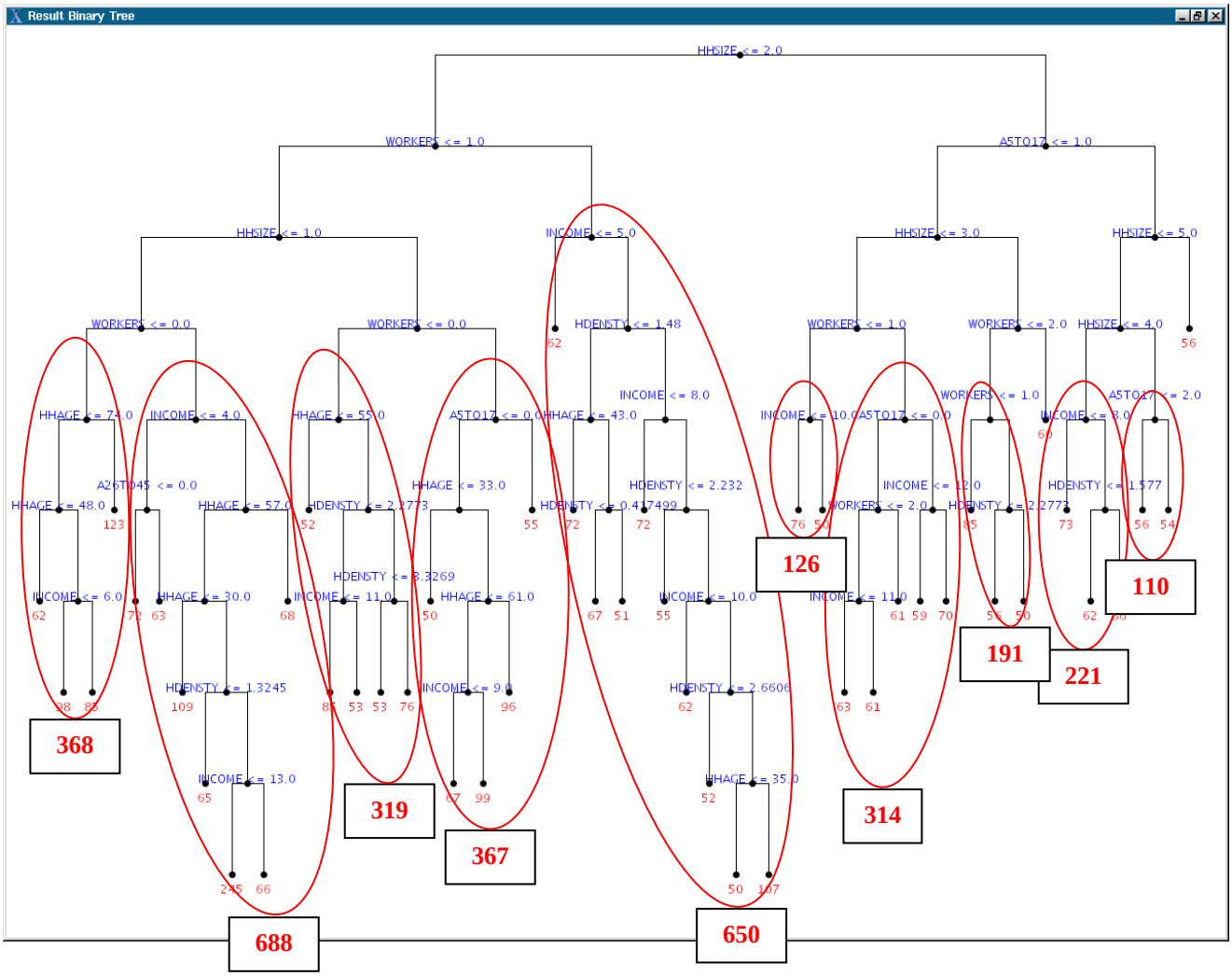


Figure-4.20d: An illustration of how the original tree is pruned

Note that there are 10 parts that are pruned from the main tree as indicated by the 10 oval regions shown in Figure-4.20d. After pruning, the nodes within each oval will be collapsed into **one** terminal node and all of the split variables (X_k) used within the oval will be deleted. The number beneath each oval is the number of **total** observations within the oval, which represents the number of the observations in the collapsed **new** terminal node. For example, considering the first oval on the left side, the four terminal nodes containing 62, 98, 85 and 123 observations, respectively, are collapsed into one terminal containing 368 observations (the sum of 62, 98, 85 and 123 observations). The splits (HHAGE$\leq$74, HHAGE$\leq$48, and INCOME$\leq$6) are deleted. Hence, the new terminal node is placed onto a **left** side of the WORKERS$\leq$0-split.

4.8.3 Matching the given synthetic household with a survey household

This procedure involves three steps:

4.8.3.1 Assign each synthetic household (obtained from the Population Synthesizer Module) to a unique terminal node of the tree built from the survey households.

4.8.3.2 Select a survey household within the terminal node to match with the given synthetic household that is assigned to this node.

4.8.3.3 Match the assigned synthetic household members with the selected survey household members based on the following four demographic variables: *relate, work, gender,* and *age*.

4.8.3.1 Assigning the given synthetic household to a unique terminal node of the tree

The given synthetic household has certain specified demographic variable values. These values are used to classify this household into the appropriate terminal node of the classification tree.

Example for assigning the synthetic household to a unique terminal node of the tree

Suppose that we are given a synthetic household with an ID 26931, which has the demographic data as shown in Table-4.9.

Table-4.9: Synthetic household demographic data for a household ID 26931

Synthetic Household-Demographic Data										
HHID	**HHSIZE**	VEHICLES	Location	INCOME	ALT5	A5TO17	A26TO45	HHAGE	**WORKERS**	DENSITY
26931	**3**	1	841405	5	0	2	1	44	**1**	7.29399

Person ID	RELATE	WORK	GENDER	AGE
55728	1 (head of HH)	1 (worker)	2 (female)	44
55729	2 (child)	2 (nonworker)	2 (female)	12
55730	2 (child)	2 (nonworker)	2 (female)	10

These demographic values are used to classify this household into the appropriate terminal node of the pruned tree as displayed in Figure-4.21.

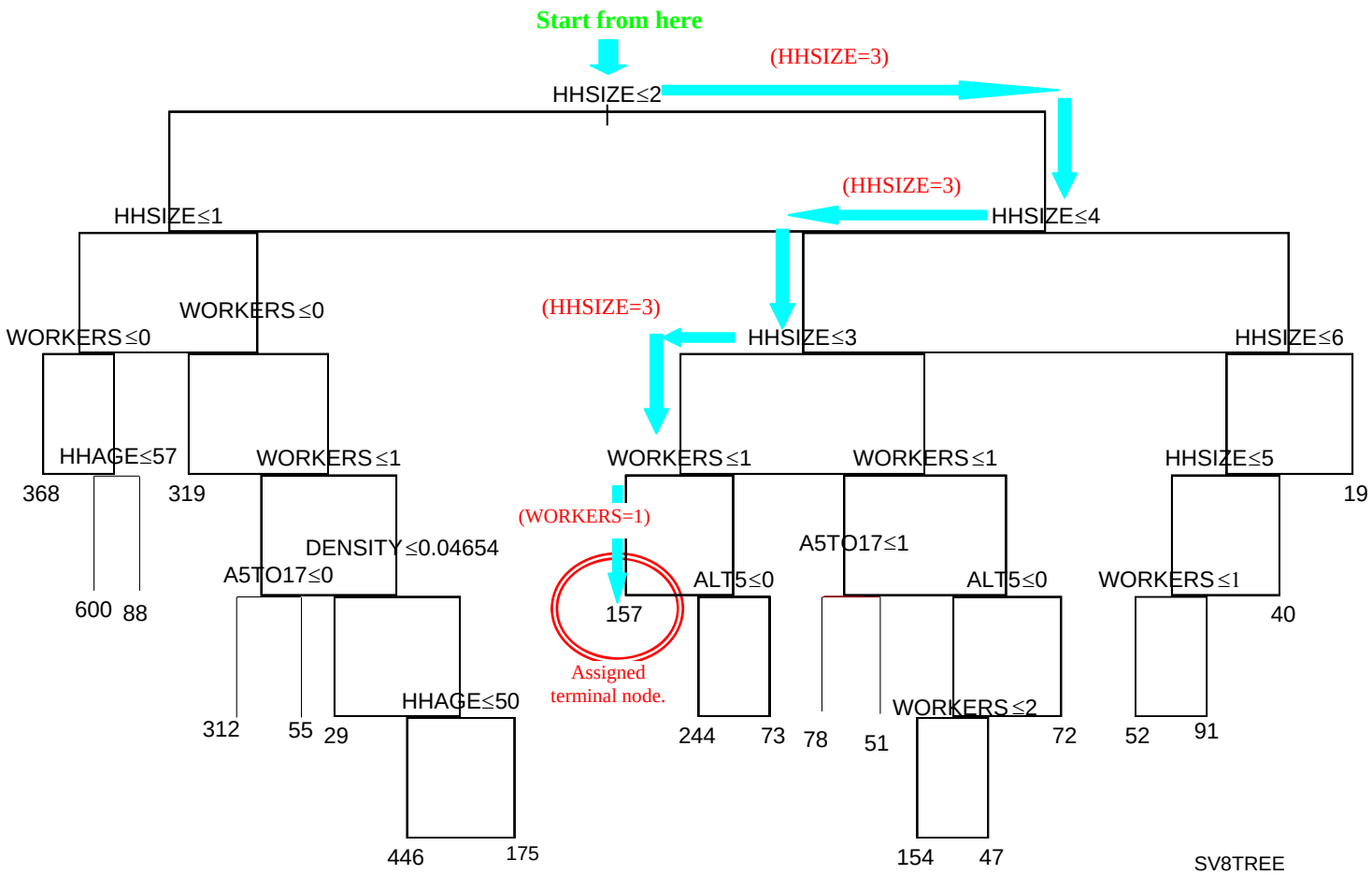


Figure-4.21: Assigning the synthetic household ID 26931 to a unique node of the pruned tree

From Figure-4.21, the assigned terminal node contains 157 survey households. The key demographic data for assigning survey households to this terminal node are **household size =3**, and **number of workers =0** or **1**.

4.8.3.2 Selecting a survey household within the terminal node to match with the given synthetic household that is assigned to this node

After the synthetic household is assigned to a specific terminal node, a survey household is chosen at random within this terminal node to obtain a matching household. For flexibility, weights are used in choosing the survey household. Each survey household has a weight w_i assigned to it in the Survey Weight File. If N is the assigned terminal node for the synthetic household, the survey household i in node N is chosen with probability $p(i) = \frac{w_i}{\sum_{j \in N} w_j}$.

The weighting scheme used in the Portland Study is described below.

Because of underreporting of trips in the survey households or sampling households with less active lifestyles, the total number of trips generated for the synthetic population using this procedure is usually too low. To remedy this, the survey households with more trips are weighted higher than those with fewer trips.

Suppose there is *p*% too few trips. (In the Portland Study, *p*=25% for combined drive and transit trips). Let $X = (1 + p/100)$. At a node in the tree, let the number of drive and transit trips for the m survey households at the node be

$$n_1, n_2 \ldots n_m$$

The average number of drive plus transit trips generated for the synthetic households at this node is

$$\bar{n} = \sum n_i / m$$

Here we use

$$w_i = (1 + \beta n_i^k)$$

as the general form of the weighting function, where β and k are calibration constants.
The form of this function for various values of k is shown in Figure-4.22.

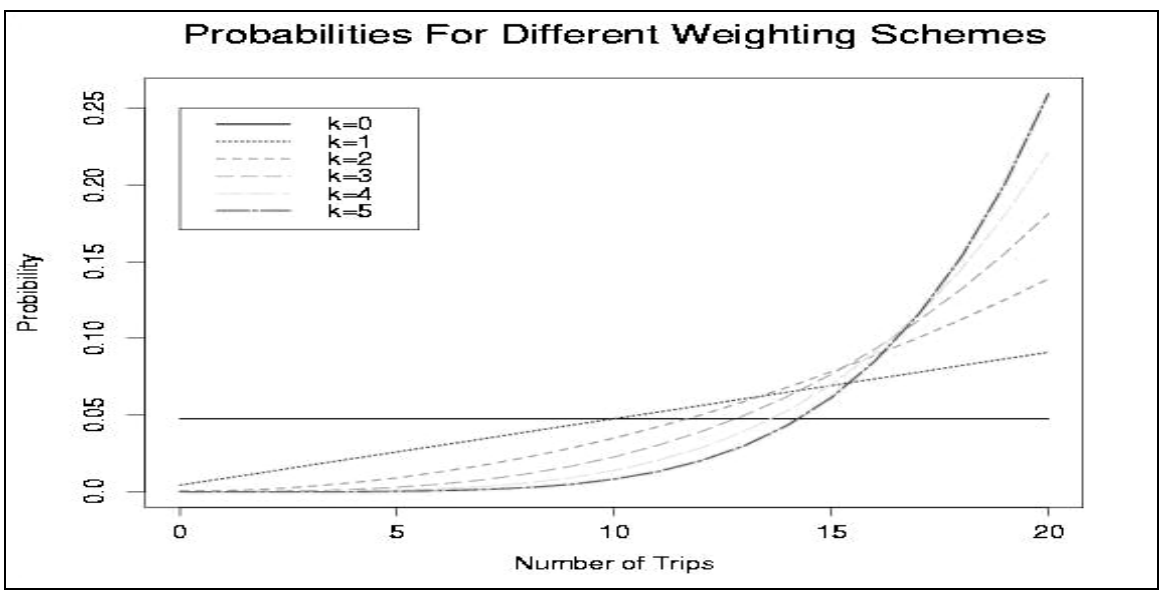


Figure-4.22: Probabilities for different weighting schemes

Those lines with the smallest k values give more similar weighting to all of the households at the node of the tree. When k equals to 0, all of the survey households get the same weight. It may not be desirable to assign one household significantly more weight than all of the others. Therefore, the smallest integer value of k (at least 1, if not expecting the same weight for every household) was determined for each node so that β is positive. As a result, all of the weights remained positive.

β is obtained by solving

$$\sum w_i n_i / \sum w_i = X\bar{n}$$

$$W_i = (1 + \beta \times n_i^k)$$

$$\sum_{i=1}^{m} W_i = m + \beta \sum_{i=1}^{m} n_i^k$$

$$\sum_{i=1}^{m} W_i \times n_i = \sum n_i + \beta \sum_{i=1}^{m} n_i^{k+1}$$

$$\frac{\sum_{i=1}^{m} W_i \times n_i}{\sum_{i=1}^{m} W_i} = \frac{\sum n_i + \beta \sum_{i=1}^{m} n_i^{k+1}}{m + \beta \sum_{i=1}^{m} n_i^k}$$

$$X \times \bar{n} = \frac{\sum n_i + \beta \sum_{i=1}^{m} n_i^{k+1}}{m + \beta \sum_{i=1}^{m} n_i^k}$$

$$X \times \bar{n} \times (m + \beta \sum_{i=1}^{m} n_i^k) = \sum n_i + \beta \sum_{i=1}^{m} n_i^{k+1} = m \times \bar{n} + \beta \sum_{i=1}^{m} n_i^{k+1}$$

$$X \times m \times \bar{n} - m \times \bar{n} = \beta \sum_{i=1}^{m} n_i^{k+1} - X \times \bar{n} \times \beta \sum_{i=1}^{m} n_i^k$$

$$(X - 1) \times m \times \bar{n} = \beta \left[\sum_{i=1}^{m} n_i^{k+1} - X \times \bar{n} \sum_{i=1}^{m} n_i^k \right]$$

$$\beta = \frac{(X - 1) \times m \times \bar{n}}{\left[\sum_{i=1}^{m} n_i^{k+1} - X \times \bar{n} \sum_{i=1}^{m} n_i^k \right]}$$

The solution for β in this equation is

$$\beta = (X - 1) m\bar{n} / [\sum n_i^{k+1} - X\bar{n} \sum n_i^k]$$

For a particular value of k, the solution for β in the above equation may be less than 0. For these cases, a larger value of k must be selected. Solutions of the equation with negative values of β lead to negative weights for some of the households.

Calculation Steps:

1. Let $k = 1$.

2. Get the value of β from the following formula:

$$\beta = (X-1)m\bar{n} / [\sum n_i^{k+1} - X\bar{n}\sum n_i^k]$$

3. If $\beta > 0$, go to step 4;
   Otherwise, $k = k + 1$; go to step 2.

4.Calculate the weight of each household:

$$w_i = (1 + \beta n_i^k)$$

5. Calculate the chosen probability for each household:

$$p(i) = \frac{w_i}{\sum_{j \in N} w_j}$$

Example for selecting a survey household within the terminal node

From Figure-4.22, the assigned terminal node contains 157 survey households having household size =3 and number of workers =0 or 1. The list of the 157 survey households is shown in Table-4.10.

Table-4.10: The list of the 157 survey households having HHSIZE = 3, and WORKERS = 0 or 1

ORDER	HHID	HHSIZE	VEHICLES	INCOME	ALT5	A5TO17	A26TO45	HHAGE	WORKERS	HDENSITY	TRIPS	WEIGHT	PROB
1.	200124	**3**	1	8	0	1	2	38	**1**	5.2314	12	3.4032	0.0092
2.	200210	**3**	1	14	0	2	0	46	**1**	1.8818	1	1.2003	0.0033
3.	200291	**3**	2	3	0	1	0	62	**0**	0.0964	5	2.0013	0.0054
4.	200354	**3**	2	13	1	0	2	40	**1**	2.9716	11	3.2029	0.0087
5.	200417	**3**	2	11	0	0	0	65	**1**	0.0348	2	1.4005	0.0038
6.	200454	**3**	1	3	0	2	1	45	**0**	0.0731	4	1.8011	0.0049
7.	200760	**3**	1	6	0	1	1	65	**0**	6.3328	4	1.8011	0.0049
8.	201002	**3**	2	14	0	1	1	64	**1**	3.3008	13	3.6034	0.0098
9.	201003	**3**	2	11	1	0	2	31	**1**	3.3008	7	2.4018	0.0065
10.	201417	**3**	3	8	0	1	0	50	**1**	4.4075	16	4.2042	0.0114
11.	201642	**3**	3	11	0	0	0	62	**1**	0.2819	5	2.0013	0.0054
12.	201671	**3**	2	6	1	0	2	39	**1**	0.0172	0	1.0000	0.0027
13.	201673	**3**	2	5	0	1	0	49	**1**	0.7661	14	3.8037	0.0103
14.	201869	**3**	2	7	1	0	2	28	**1**	3.8775	6	2.2016	0.0060
15.	202153	**3**	2	3	0	0	2	27	**1**	2.8272	7	2.4018	0.0065
16.	202265	**3**	2	13	0	1	2	33	**1**	0.5886	9	2.8024	0.0076
17.	202689	**3**	2	7	0	0	0	58	**1**	5.5296	12	3.4032	0.0092
18.	202773	**3**	3	13	0	0	1	65	**0**	1.4800	2	1.4005	0.0038
19.	202963	**3**	3	12	0	0	0	47	**1**	2.3848	7	2.4018	0.0065
20.	203130	**3**	3	14	0	0	0	48	**1**	7.0896	20	5.0053	0.0136
21.	203156	**3**	2	4	0	1	2	34	**1**	2.8272	11	3.2029	0.0087

ORDER	HHID	HHSIZE	VEHICLES	INCOME	ALT5	A5TO17	A26TO45	HHAGE	WORKERS	HDENSITY	TRIPS	WEIGHT	PROB
22.	203401	**3**	1	4	0	1	0	66	**0**	3.3269	2	1.4005	0.0038
23.	203681	**3**	4	9	0	0	1	44	**0**	0.0234	3	1.6008	0.0043
24.	203704	**3**	1	2	1	0	2	31	**0**	0.2549	9	2.8024	0.0076
25.	203799	**3**	3	14	0	1	2	30	**1**	0.3893	9	2.8024	0.0076
26.	203829	**3**	2	2	2	0	1	28	**1**	1.5560	0	1.0000	0.0027
27.	203868	**3**	1	9	0	0	3	36	**0**	2.4933	15	4.0039	0.0108
28.	203914	**3**	1	14	0	2	1	37	**0**	3.4679	4	1.8011	0.0049
29.	203999	**3**	4	4	0	0	1	64	**1**	5.2382	2	1.4005	0.0038
30.	204122	**3**	2	14	1	0	2	29	**1**	1.0599	5	2.0013	0.0054
31.	204147	**3**	2	10	0	0	2	36	**1**	0.0792	3	1.6008	0.0043
32.	204271	**3**	1	9	2	0	1	41	**1**	0.0497	10	3.0026	0.0081
33.	204563	**3**	3	11	0	0	0	52	**1**	5.8694	6	2.2016	0.0060
34.	204838	**3**	1	7	0	2	0	48	**1**	5.3647	4	1.8011	0.0049
35.	204854	**3**	2	7	0	0	2	41	**0**	4.1182	8	2.6021	0.0071
36.	205054	**3**	1	14	1	0	2	30	**1**	2.4801	2	1.4005	0.0038
37.	205159	**3**	2	13	1	0	2	32	**1**	0.1215	7	2.4018	0.0065
38.	205258	**3**	2	3	0	0	1	67	**1**	7.8221	2	1.4005	0.0038
39.	205432	**3**	2	10	1	0	2	32	**1**	0.4064	5	2.0013	0.0054
40.	205457	**3**	2	6	1	0	1	40	**1**	1.9141	1	1.2003	0.0033
41.	205558	**3**	0	4	2	0	1	33	**1**	2.2976	6	2.2016	0.0060
42.	205588	**3**	3	1	0	0	0	63	**0**	1.5560	6	2.2016	0.0060
43.	205597	**3**	6	6	0	0	0	46	**1**	0.0272	6	2.2016	0.0060
44.	206078	**3**	2	5	2	0	1	29	**1**	3.8458	4	1.8011	0.0049
45.	206132	**3**	1	3	0	2	1	26	**1**	3.0264	8	2.6021	0.0071
46.	206421	**3**	3	6	0	0	1	77	**1**	0.0341	3	1.6008	0.0043
47.	206606	**3**	2	8	1	0	1	22	**1**	2.1634	1	1.2003	0.0033
48.	206691	**3**	2	11	1	0	2	30	**1**	3.2001	7	2.4018	0.0065
49.	206854	**3**	2	1	0	2	1	45	**1**	1.3245	8	2.6021	0.0071
50.	206881	**3**	2	10	0	0	0	56	**1**	1.2087	12	3.4032	0.0092
51.	206987	**3**	2	14	0	0	1	60	**1**	6.5126	3	1.6008	0.0043
52.	207180	**3**	2	5	0	0	1	76	**1**	0.2398	0	1.0000	0.0027
53.	207361	**3**	4	14	0	0	0	62	**1**	3.4075	4	1.8011	0.0049
54.	207710	**3**	1	9	1	0	0	20	**1**	1.5560	5	2.0013	0.0054
55.	207786	**3**	1	13	0	2	1	41	**1**	0.7979	13	3.6034	0.0098
56.	207923	**3**	1	11	0	1	1	50	**1**	2.6143	2	1.4005	0.0038
57.	208283	**3**	3	14	0	0	0	61	**1**	6.5946	8	2.6021	0.0071
58.	208542	**3**	2	13	0	0	0	58	**1**	0.7979	10	3.0026	0.0081
59.	208698	**3**	2	13	0	2	1	43	**1**	5.4203	6	2.2016	0.0060
60.	208765	**3**	3	10	0	1	2	29	**1**	2.4991	3	1.6008	0.0043
61.	208947	**3**	2	11	0	0	0	62	**1**	2.2398	6	2.2016	0.0060
62.	209069	**3**	2	8	1	0	2	28	**1**	2.8951	11	3.2029	0.0087
63.	209281	**3**	2	8	0	1	2	35	**1**	1.5770	12	3.4032	0.0092
64.	209345	**3**	3	8	0	0	0	50	**1**	2.3947	2	1.4005	0.0038
65.	209466	**3**	2	5	0	0	0	70	**0**	1.2645	5	2.0013	0.0054
66.	209728	**3**	1	9	1	1	1	40	**1**	2.4719	15	4.0039	0.0108
67.	209838	**3**	1	14	0	0	0	21	**1**	7.8259	14	3.8037	0.0103
68.	209903	**3**	2	8	1	0	2	36	**1**	3.8963	1	1.2003	0.0033
69.	209998	**3**	2	6	0	0	1	69	**1**	4.1182	1	1.2003	0.0033
70.	210060	**3**	1	7	0	1	1	41	**1**	1.7639	2	1.4005	0.0038
71.	210238	**3**	2	10	0	0	0	51	**1**	2.1795	0	1.0000	0.0027
72.	210249	**3**	2	3	1	0	2	28	**1**	0.0543	3	1.6008	0.0043
73.	210293	**3**	2	13	0	2	0	50	**0**	0.1215	8	2.6021	0.0071
74.	210359	**3**	3	13	0	1	0	48	**1**	1.2771	17	4.4045	0.0119
75.	210509	**3**	3	4	0	0	1	67	**1**	3.3008	15	4.0039	0.0108
76.	210616	**3**	3	13	0	1	1	61	**1**	0.2549	6	2.2016	0.0060
77.	210824	**3**	0	3	1	0	2	34	**0**	2.6606	5	2.0013	0.0054
78.	210857	**3**	2	5	0	1	2	30	**1**	5.8622	1	1.2003	0.0033
79.	210971	**3**	3	14	0	1	2	45	**1**	2.3124	9	2.8024	0.0076
80.	211199	**3**	2	8	0	1	1	39	**1**	2.8641	3	1.6008	0.0043
81.	211238	**3**	2	7	0	0	1	64	**1**	4.1098	1	1.2003	0.0033

ORDER	HHID	HHSIZE	VEHICLES	INCOME	ALT5	A5TO17	A26TO45	HHAGE	WORKERS	HDENSITY	TRIPS	WEIGHT	PROB
82.	211326	**3**	2	13	2	0	1	40	**1**	0.0346	2	1.4005	0.0038
83.	211411	**3**	3	6	1	0	2	33	**1**	2.0662	3	1.6008	0.0043
84.	211668	**3**	2	14	0	1	2	44	**1**	2.4755	8	2.6021	0.0071
85.	211750	**3**	2	8	0	2	0	50	**1**	4.4075	13	3.6034	0.0098
86.	211772	**3**	2	5	0	1	0	72	**0**	3.5586	2	1.4005	0.0038
87.	211841	**3**	3	10	0	0	1	68	**1**	2.8915	2	1.4005	0.0038
88.	211903	**3**	2	14	1	0	2	32	**1**	2.4909	11	3.2029	0.0087
89.	212010	**3**	2	11	1	0	2	27	**1**	0.5020	9	2.8024	0.0076
90.	212025	**3**	0	3	0	2	1	42	**1**	1.5560	4	1.8011	0.0049
91.	212162	**3**	1	5	1	0	2	26	**1**	6.3328	5	2.0013	0.0054
92.	212273	**3**	1	4	0	2	1	38	**1**	2.8951	9	2.8024	0.0076
93.	212346	**3**	2	14	0	1	2	26	**1**	0.0869	9	2.8024	0.0076
94.	212437	**3**	1	4	0	2	1	33	**1**	7.8259	6	2.2016	0.0060
95.	212599	**3**	0	6	1	1	0	24	**1**	4.1182	1	1.2003	0.0033
96.	212667	**3**	1	6	0	2	1	39	**1**	1.7639	3	1.6008	0.0043
97.	212690	**3**	1	8	1	0	1	22	**1**	2.2925	17	4.4045	0.0119
98.	212747	**3**	2	9	0	2	1	42	**1**	1.9141	13	3.6034	0.0098
99.	213202	**3**	3	13	0	1	0	56	**1**	0.0792	2	1.4005	0.0038
100.	213497	**3**	1	10	1	0	1	47	**1**	13.1715	12	3.4032	0.0092
101.	213498	**3**	3	13	1	0	2	35	**1**	1.5431	6	2.2016	0.0060
102.	213569	**3**	3	4	2	0	0	25	**1**	2.7252	8	2.6021	0.0071
103.	213812	**3**	0	5	0	1	2	33	**1**	3.2606	1	1.2003	0.0033
104.	214026	**3**	4	13	0	1	1	62	**1**	2.2320	4	1.8011	0.0049
105.	214230	**3**	3	8	1	0	2	31	**1**	1.2771	3	1.6008	0.0043
106.	214351	**3**	1	7	1	0	1	23	**1**	2.1634	6	2.2016	0.0060
107.	214454	**3**	3	9	0	1	0	53	**1**	0.4064	5	2.0013	0.0054
108.	214617	**3**	1	14	1	0	0	72	**0**	1.2087	8	2.6021	0.0071
109.	214731	**3**	1	9	0	0	1	74	**1**	2.8272	3	1.6008	0.0043
110.	214939	**3**	1	13	0	2	1	44	**1**	2.8272	7	2.4018	0.0065
111.	214957	**3**	1	5	0	2	1	38	**0**	3.2578	11	3.2029	0.0087
112.	215271	**3**	4	14	0	1	0	67	**1**	3.8458	10	3.0026	0.0081
113.	215307	**3**	1	6	0	2	1	31	**1**	1.9141	18	4.6047	0.0125
114.	215387	**3**	1	7	0	2	1	37	**1**	0.0869	17	4.4045	0.0119
115.	215749	**3**	1	14	0	2	1	43	**1**	0.0188	7	2.4018	0.0065
116.	215781	**3**	3	14	1	0	2	43	**1**	1.8817	6	2.2016	0.0060
117.	216445	**3**	4	10	0	2	1	30	**1**	5.8694	5	2.0013	0.0054
118.	216461	**3**	1	1	0	2	1	33	**1**	1.4800	2	1.4005	0.0038
119.	216575	**3**	1	14	0	2	1	41	**1**	0.4064	5	2.0013	0.0054
120.	216599	**3**	2	6	0	1	0	48	**1**	3.4418	1	1.2003	0.0033
121.	216613	**3**	3	13	1	0	1	42	**1**	0.2323	6	2.2016	0.0060
122.	216635	**3**	1	5	1	0	1	30	**1**	1.7013	15	4.0039	0.0108
123.	217070	**3**	3	6	0	0	1	69	**1**	2.4909	7	2.4018	0.0065
124.	217090	**3**	3	8	0	0	1	60	**1**	0.0188	3	1.6008	0.0043
125.	217123	**3**	2	14	0	2	1	44	**1**	0.0497	10	3.0026	0.0081
126.	217435	**3**	2	14	2	0	0	20	**0**	2.6606	3	1.6008	0.0043
127.	217576	**3**	2	13	0	0	1	50	**1**	0.7979	14	3.8037	0.0103
128.	217818	**3**	2	13	1	0	2	38	**1**	0.4743	14	3.8037	0.0103
129.	217828	**3**	2	5	0	2	1	39	**1**	2.0662	7	2.4018	0.0065
130.	218641	**3**	0	4	0	2	1	34	**1**	8.4041	4	1.8011	0.0049
131.	218789	**3**	2	6	0	1	1	43	**1**	1.6464	5	2.0013	0.0054
132.	218914	**3**	1	5	0	1	2	37	**1**	3.8775	11	3.2029	0.0087
133.	219223	**3**	1	8	1	1	1	44	**1**	2.2452	2	1.4005	0.0038
134.	219921	**3**	1	9	0	0	0	57	**1**	4.4075	5	2.0013	0.0054
135.	220013	**3**	1	8	0	2	1	41	**1**	2.8951	12	3.4032	0.0092
136.	220185	**3**	2	6	2	0	0	25	**1**	4.8826	3	1.6008	0.0043
137.	220436	**3**	3	13	0	0	0	55	**1**	0.4491	11	3.2029	0.0087
138.	220469	**3**	1	14	0	2	1	32	**1**	2.4735	12	3.4032	0.0092
139.	220629	**3**	1	3	0	2	1	34	**1**	0.0752	14	3.8037	0.0103
140.	221197	**3**	2	5	1	0	1	23	**1**	7.0896	9	2.8024	0.0076
141.	221198	**3**	1	4	0	1	2	33	**0**	4.6723	6	2.2016	0.0060

ORDER	HHID	HHSIZE	VEHICLES	INCOME	ALT5	A5TO17	A26TO45	HHAGE	WORKERS	HDENSITY	TRIPS	WEIGHT	PROB
142.	221209	**3**	1	14	0	0	1	33	**1**	6.3328	9	2.8024	0.0076
143.	221278	**3**	1	2	0	2	1	26	**0**	3.7102	11	3.2029	0.0087
144.	221282	**3**	2	6	1	0	2	27	**1**	2.3505	6	2.2016	0.0060
145.	222383	**3**	2	14	0	1	0	47	**1**	0.7979	5	2.0013	0.0054
146.	223322	**3**	1	1	0	2	1	38	**1**	11.4504	9	2.8024	0.0076
147.	223350	**3**	2	8	0	0	0	62	**0**	2.6143	4	1.8011	0.0049
148.	223754	**3**	1	4	1	1	1	40	**1**	2.8272	0	1.0000	0.0027
149.	224495	**3**	2	13	0	1	0	50	**1**	1.5431	4	1.8011	0.0049
150.	225025	**3**	1	13	1	0	2	27	**1**	6.5126	7	2.4018	0.0065
151.	225966	**3**	1	10	0	0	0	47	**1**	2.5314	6	2.2016	0.0060
152.	226028	**3**	2	11	1	0	2	39	**1**	1.8049	11	3.2029	0.0087
153.	226162	**3**	2	13	0	1	0	48	**1**	2.9255	8	2.6021	0.0071
154.	226185	**3**	2	14	0	0	0	99	**1**	0.2049	8	2.6021	0.0071
155.	226246	**3**	3	13	0	1	0	48	**1**	2.1896	9	2.8024	0.0076
156.	226344	**3**	2	6	0	1	0	49	**1**	0.3657	4	1.8011	0.0049
157.	227374	**3**	3	11	0	1	0	65	**1**	21.6827	4	1.8011	0.0049

After the synthetic household ID 26931 is assigned to the terminal node containing these 157 survey households, a survey household is chosen at random within this terminal node to obtain a matching household.

First, we need to get the weight of each household using the scheme we described before.

1. Starting from $k = 1$.
2. Get $\beta = 0.2003$.
3. Since $\beta > 0$, go to step 4.
4. Calculate the weight of each household (as Shown in Table-9).
5. Calculate the chosen probability for each household (as shown in Table-4.10).

We randomly select a survey household according to this probability by using the associated cumulative distribution (as shown in Figure-4.23), which is divided into four parts to represent the total survey household IDs of 157 households. This is done by generating a random number between 0 and 1, and then selecting that survey household for which the random number falls within its defined interval.

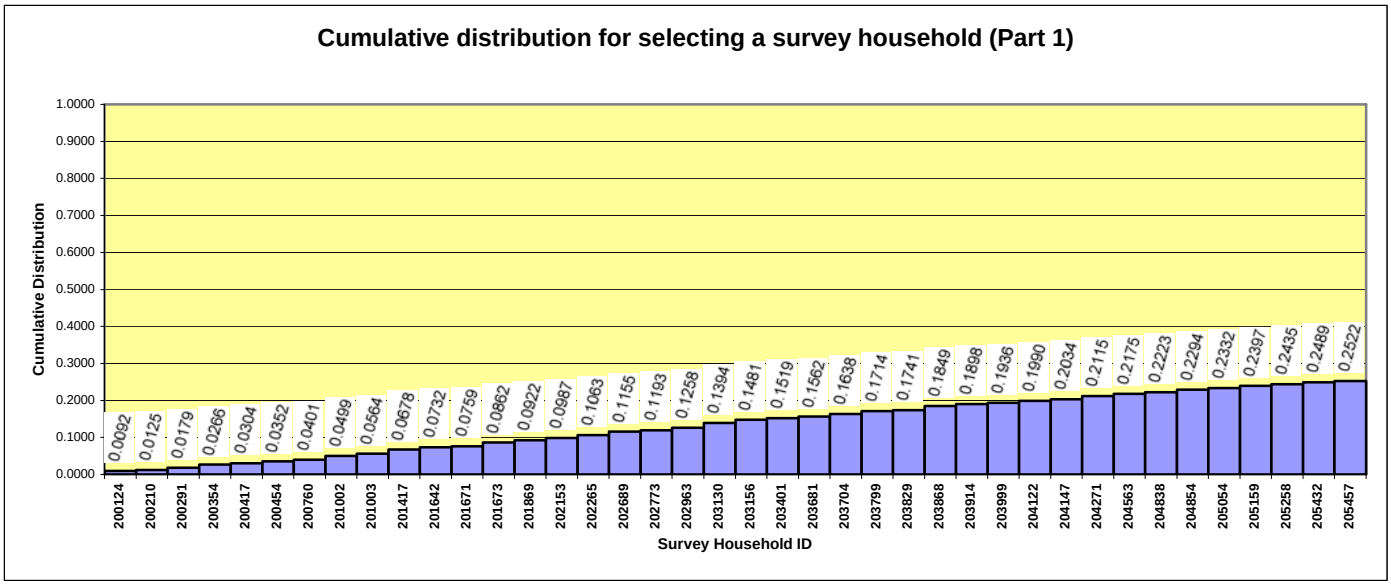


Figure-4.23: Cumulative distribution (part 1) for selecting a survey household within the terminal node

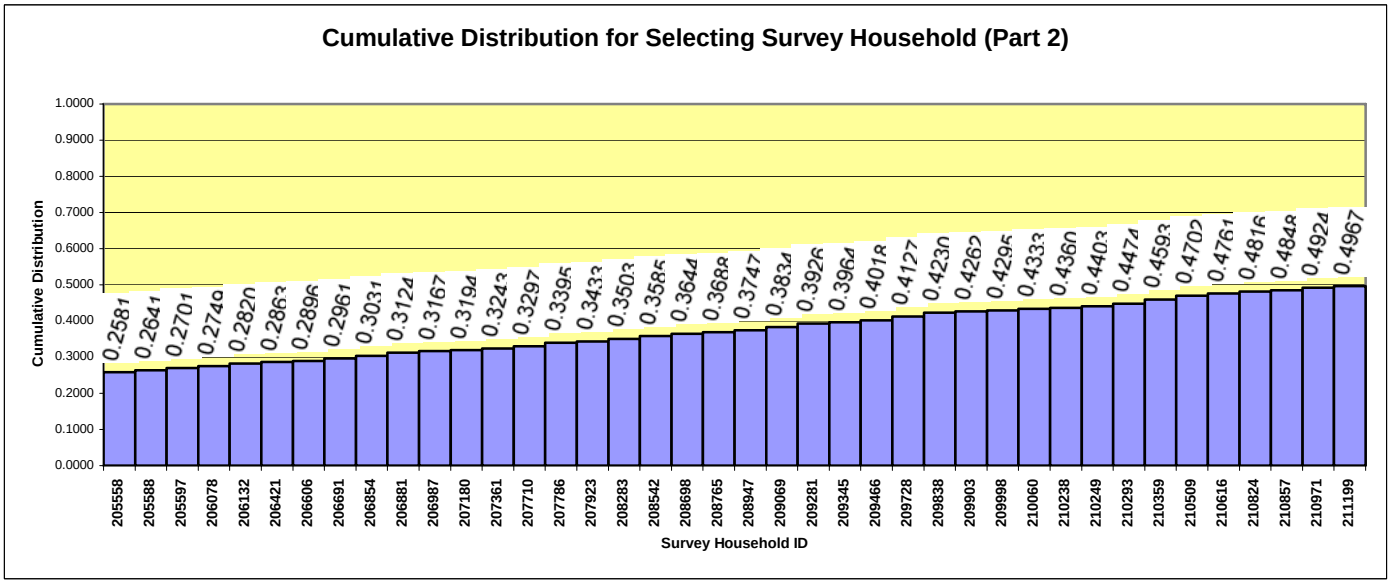


Figure-4.24: Cumulative distribution (part 2) for selecting a survey household within the terminal node

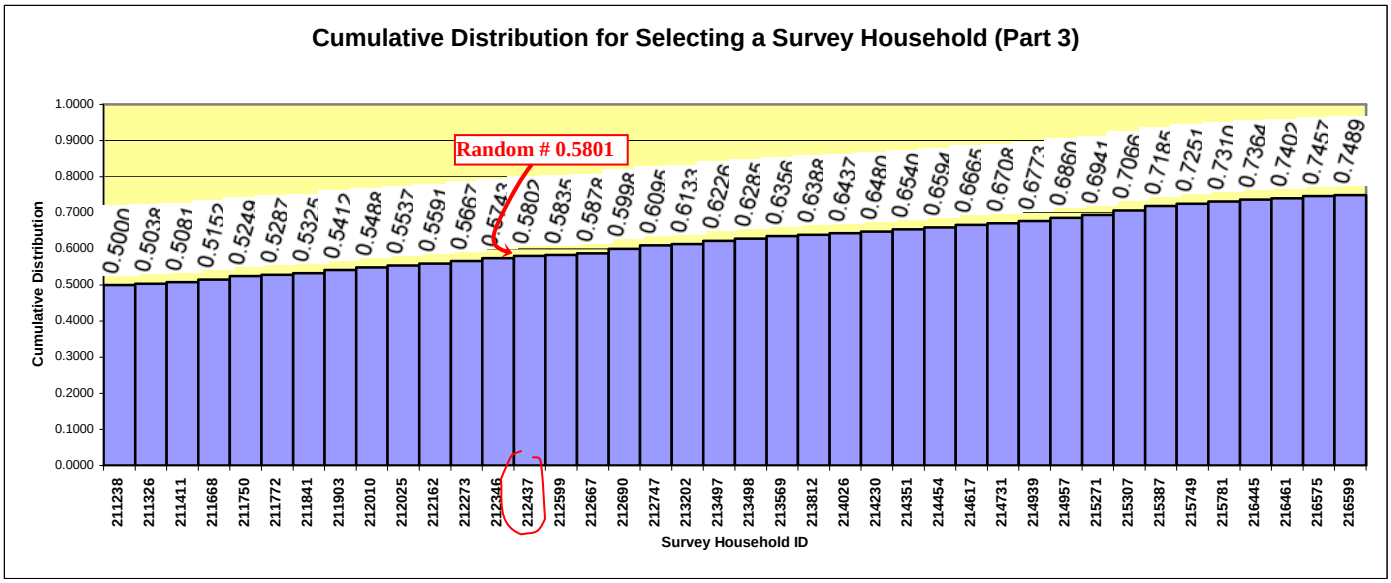


Figure-4.25: Cumulative distribution (part 3) for selecting a survey household within the terminal node

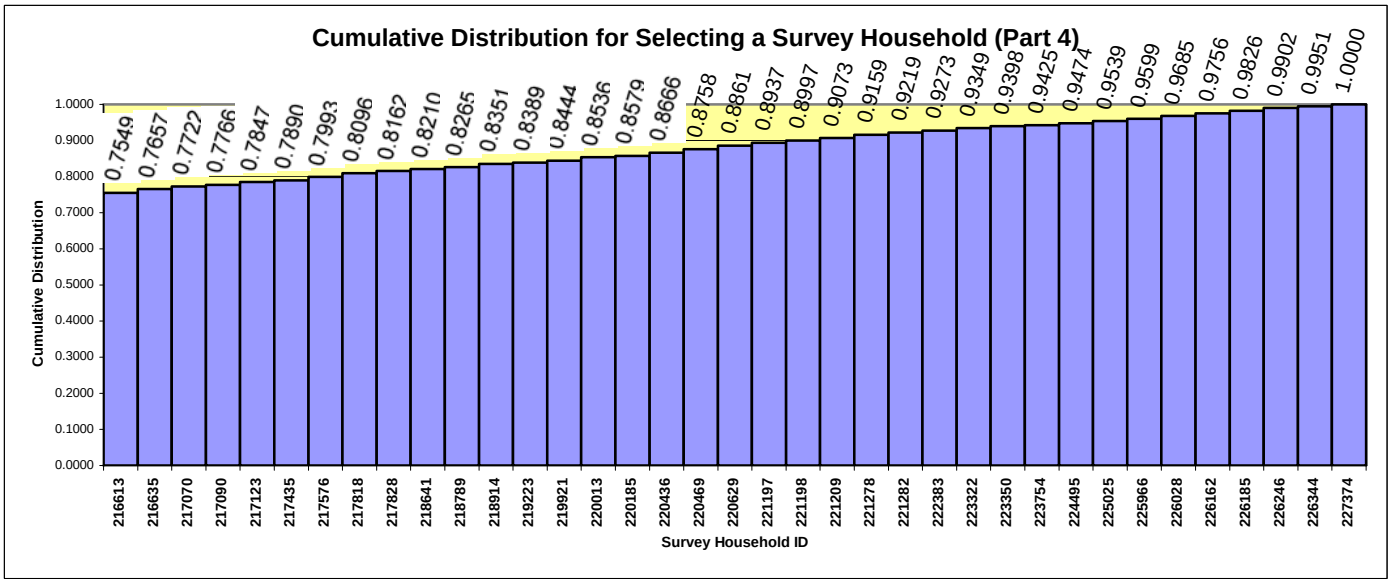


Figure-4.26: Cumulative distribution (part 4) for selecting a survey household within the terminal node

For example, suppose that we randomly generate the number **0.5801**. Then, the **survey household ID** selected would be **212437** (see Figure-4.25).

4.8.3.3 Matching the assigned synthetic household members with the selected survey household members based on the four demographic variables: relate, work, gender, and age

The four demographic variables relate, work, gender, and age are contained in the household member demographic information in both the *survey household demographic file* (see Section 4.4.2) and the *synthetic households file* (see Section 4.4.3). The household members from the synthetic household must be matched (in a one-to-one fashion) with the household members from the survey household based on these four variables. Following this, the information in the skeletal activity patterns of the survey household members (see Section 4.8.1) is designated as the activity patterns for the corresponding synthetic household member.

Although perfect matches are not always possible, the Activity Generator Module does attempt to find the best matches possible via the following process.

STEP 1: Matching Children

1. Sort the children in the synthetic household by *gender* and *age* (in descending order). The children in the survey household are sorted in an identical manner.
2. A one-to-one match is made between the two sorted lists. If the survey household has more children than the synthetic household, the extra children in the survey household are ignored. If the synthetic household contains more children than the survey household, the activities of the last child in the survey household are replicated as often as necessary to create activities for the remaining children in the synthetic household.

STEP 2: Matching Adults

1. The adults in the synthetic and the survey households are sorted by the four demographic variables: *relate, work, gender,* and *age* (in descending order).
2. Again, a one-to-one match is made using the same rules as for the children.

STEP 3: Assigning the information in the skeletal activity patterns for each person in the matched survey household to the corresponding person in the synthesized household

The skeletal activity patterns for each person in the survey household are obtained from the data described in Section 4.8.1. The major information used relate to *the list of activities, their priorities* (TRANSIMS currently sets the value of all priorities as nine; this is the maximum priority)*, travel modes,* and *vehicle preferences* (obtained from the Vehicle Population)*.* If any transportation activity is shared, the appropriate participating household members who might use the same mode of transportation are grouped (e.g., those that use the same car).

Example for matching the assigned synthetic household members with the selected survey household members

From the previous example, the given synthetic household is one with an ID 26931, and the selected survey household that is matched with it is numbered 212273. Tables 4.11 and 4.12 display the demographic data of these two households. Only the four demographic variables *relate, work, gender,* and *age* are used for matching purposes.

Table-4.11: Synthetic household demographic data for a household ID 26931

Synthetic Household-Demographic Data										
HHID	HHSIZE	VEHICLES	Location	INCOME	ALT5	A5TO17	A26TO45	HHAGE	WORKERS	DENSITY
26931	3	1	**841405**	5	0	2	1	44	1	7.29399

Person ID	RELATE	WORK	GENDER	AGE
55728	1 (head of HH)	1 (worker)	2 (female)	44
55729	2 (child)	2 (nonworker)	2 (female)	12
55730	2 (child)	2 (nonworker)	2 (female)	10

(The four variables used for matching purposes. These rows are matched with the corresponding rows in Table 8.7.)

Table-4.12: Survey household demographic data for a household numbered 212273

Survey Household-Demographic Data										
HH#	HHSIZE	VEHICLES	Location	INCOME	ALT5	A5TO17	A26TO45	HHAGE	WORKERS	DENSITY
212273	3	1	-	4	0	2	1	38	1	2.8951

Person #	RELATE	WORK	GENDER	AGE
4426	1 (head of HH)	1 (worker)	2 (female)	38
4427	2 (child)	2 (nonworker)	1 (male)	9
4428	2 (child)	2 (nonworker)	2 (female)	7

(The four variables used for matching purposes. These rows are matched with the corresponding rows in Table 8.6.)

The matching is as follows:

- The **person ID 55728 in the synthetic household** who is the head of the household, an only worker, of female gender, and aged 44, is matched with the **person numbered 4426 in the survey household** who is the head of the household, an only worker, of female gender, and aged 38.
- The **person ID 55729 in the synthetic household** who is a child, a nonworker, of female gender, and aged 12, is matched with the **person numbered 4427 in the survey household** who is a child, a nonworker, of male gender, and aged 9. Although they have different gender, they provide the best match.
- The **person ID 55730 in the synthetic household** who is a child, a nonworker, of female gender, and aged 10, is matched with the **person numbered 4428 in the survey household** who is a child, a nonworker, of female gender, and aged 7.

Fortunately, the number of adults and children in both households are exactly the same. Hence, there is no need to ignore any extra adults/children in the survey household or replicate activities of the last adult/child in the survey household for any remaining adults/children in the synthetic household as described in Section 4.8.3.3.

Next, we assign the information in the **skeletal activity patterns** for each person in the matched survey household numbered 212273 (as shown in Table-4.12) to the corresponding person in the synthetic household ID 26931. The information assigned **at this step** is shown in its output format in Table-4.13.

Table-4.13: The skeletal activity pattern for the survey household numbered 212273

HH#	PERSNO	ACTNO	ACTID	AT_HOME	WUTHERE	MODE	DRIVER	NUMVEH	ACTSTAR	ACTEND	GEOX	GEOY	Note
212273	4426	0	0 (home)	1 (yes)	2 (no)	1 (walk)	0 (no)	0	0	570	7639240	662963.5	Home
		1	1 (work)	2 (no)	2	2 (car)	1 (driver)	1	570	810	7625769	650143.5	Work
		2	5 (other)	1	2	2	1	1	825	1050	7639240	662963.5	Home
		3	5	2	2	2	1	3	1180	1210	7639300	2	Other1
		4	2 (shop)	2	2	2	1	3	1215	1235	7639938	665034.3	Shop 1
		5	0	1	2	2	1	3	1240	1620	7639240	662963.5	Home
	4427	0	0	1	2	1	0	0	0	398	7639240	662963.5	Home
		1	4 (visit)	2	2	1	0	0	398	840	7639355	662843.5	Visit 1
		2	0	1	2	1	0	0	960	1050	7639240	662963.5	Home
		3	5	2	2	2	2 (passenger)	3	1180	1210	7639300	2	Other1
		4	2	2	2	2	2	3	11215	1235	7639938	665034.3	Shop 1
		5	0	1	2	2	2	3	1240	1620	7639240	662963.5	Home
	4428	0	0	1	2	1	0	0	0	398	7639240	662963.5	Home
		1	4 (visit)	2	2	1	0	0	398	740	7639355	662843.5	Visit 1
		2	0	1	2	1	0	0	960	1050	7639240	662963.5	Home
		3	5	2	2	2	2 (passenger)	3	1180	1210	7639300	2	Other1
		4	2	2	2	2	2	3	11215	1235	7639938	665034.3	Shop 1
		5	0	1	2	2	2	3	1240	1620	7639240	662963.5	Home

Table-4.14: TRANSIMS Activities file (part 1) corresponding to the matched synthetic household ID 26931 (assigned at Section 4.8.3)

HHID	PERID	ACTNO	ACTTYP	PRIORITY	St Time-*a*	St Time-*b*	St-*a*	St-*b*	End T-*a*	End T-*b*	Et-*a*	Et-*b*

26931	55728	1	0 (home)	9								
		2	1 (work)	9								
		3	5 (other)	9								
		4	5	9								
		5	2 (shop)	9								
		6	0	9								
	55729	1	0	9								
		2	4 (visit)	9								
		3	0	9								
		4	5	9								
		5	2	9								
		6	0	9								
	55730	1	0	9								
		2	4	9								
		3	0	9								
		4	5	9								
		5	2	9								
		6	0	9								

Table-4.14: TRANSIMS Activities file (part 2) corresponding to the matched synthetic household ID 26931 (assigned at Section 8.3)

Dur Tim-*a*	Dur Tim-b	D-*a*	D-*b*	Mode	VEHid	N Loc	Location	N Others	OthersID	G Number
				1 (walk)	-1 (no vehicle)	1	841405 (home, obtained	0	-	1

							directly from the synthetic household)			
				2 (car)	45554 (from vehicle file)	1		0	-	1
				2	45554	1	841405	0	-	1
				2	45554	1		3	55728, 55729, and 55730	1
				2	45554	1		3	55728, 55729, and 55730	1
				2	45554	1	841405	3	55728, 55729, and 55730	1
				1	-1	1	841405 (home)	0	-	1
				1	-1	1		0	-	1
				1	-1	1	841405	0	-	1
				2	45554	1		3	55728, 55729, and 55730	1
				2	45554	1		3	55728, 55729, and 55730	1
				2	45554	1	841405	3	55728, 55729, and 55730	1
				1	-1	1	841405	0	-	1
				1	-1	1		0	-	1
				1	-1	1	841405	0	-	1
				2	45554	1		3	55728, 55729, and 55730	1
				2	45554	1		3	55728, 55729, and 55730	1
				2	45554	1	841405	3	55728, 55729, and 55730	1

Note that we directly obtain the particular **home ID** from the synthetic household demographic data (see Table-4.10). The IDs for the other locations will be generated later in Section 4.8.5.

4.8.4 Generating activity times and durations

Each activity in the matched synthetic household has a preferred *start-time, end-time,* and *duration*. These times are taken from the skeletal patterns of the household survey (see Section 4.8.1). The range of each of these times is specified by a beta distribution, which requires four parameters:

- Lower bound (L)
- Upper bound (U), and
- *Shape* parameters (a and b).

When a and b equal 1, the lower and upper bounds have different values, and therefore, there exists a range about the observed time, (see the specific bounds of these ranges in Table-5.16). When a and b equal –1, the lower and upper bounds have the same values, and so, there exists no range around the observed times. The observed times are taken directly from the skeletal patterns. Table-4.16 gives the values of the parameters and their ranges for each type of activity.

Table-4.16: The settings of time parameters for each type of activity. (Obs. Means observed time taken from the skeletal pattern. The times are in hours)

Type of Activity		Lower Bound (*L*) (hours)	Upper Bound (*U*) (hours)	*a*	*b*
Work	Start	Obs. – 0.25	Obs. + 0.25	1	1
	End	Obs. – 0.25	Obs. + 0.25	1	1
	Duration	*Obs.** – 0.25	*Obs.** + 0.25	1	1
Out-of-home	Start	Obs. – 0.50	Obs. + 0.50	1	1
	End	Obs. – 0.50	Obs. + 0.50	1	1
	Duration	*Obs.** – 0.3(*Obs.**)	*Obs.** + 0.3(*Obs.**)	1	1
Activity at home that **starts** at 12:00 AM.	Start	0	0	-1	-1
	End	Obs. – 0.75	Obs. + 0.75	1	1
	Duration	(Obs. of the ending time)–0.75	(Obs. of the ending time)+0.75	1	1
Activity at home during the day (activity that does not start or end at 12:00 AM.)	Start	Obs. – 0.75	Obs. + 0.75	1	1
	End	Obs. – 0.75	Obs. + 0.75	1	1
	Duration	*Obs.** – 1.00	*Obs.** + 1.00	1	1
Activity at home that **ends** at 12:00 AM.	Start	Obs. – 0.75	Obs. + 0.75	1	1
	End	24.00	24.00	-1	-1
	Duration	24.00 – (Obs. of the starting time) – 0.75	24.00 – (Obs. of the starting time) + 0.75	1	1

Note that the observed duration is calculated as the* ***observed ending time minus* *the* ***observed starting time****.

Example for generating activity times and durations

To complete Table-4.13, we need to generate information regarding times for the activities in this table via the observed times taken from the skeletal pattern (Table-4.12), and using the time parameters for each activity type displayed in Table-4.16. For the sake of illustration, Table-4.17 is created to show the computation of the **observed duration** for each activity. This is calculated as the **observed ending time minus** the **observed starting time.** Note that we have to convert the units of the starting and ending times from **minutes since midnight** to **hours since midnight**.

Table-4.17: Calculation of the observed durations for the various activities

HH#	PERSNO	ACTNO	ACTTYPE	ACTSTAR (hours since midnight)	ACTEND (hours since midnight)	DURATION (hours since midnight)
212273	4426	0	0 (home)	0	570/60 = **9.5**	**9.5**

		1	1 (work)	570/60 = **9.5**	810/60 = **13.5**	**4**
		2	5 (other but taking place at home)	825/60 =**13.75**	1050/60 = **17.5**	**3.75**
		3	5 (other)	1180/60 = **19.6667**	1210/60 = **20.1667**	**0.5**
		4	2 (shop)	1215/60 = **20.25**	1235 = **20.5833**	**0.3333**
		5	0 (home)	1240/60 = **20.6667**	**24** (default value)	**3.3333**
	4427	0	0 (home)	0	398/60 = **6.6333**	**6.6333**
		1	4 (visit)	398/60 = **6.6333**	840/60 = **14**	**7.3667**
		2	0 (home)	960/60 = **16**	1050/60 = **17.5**	**1.5**
		3	5 (other)	1180/60 = **19.6667**	1210/60 = **20.1667**	**0.5**
		4	2 (shop)	1215/60 = **20.25**	1235 = **20.5833**	**0.3333**
		5	0 (home)	1240/60 = **20.6667**	**24** (default value)	**3.3333**
	4428	0	0 (home)	0	398/60 = **6.6333**	**6.6333**
		1	4 (visit)	398/60 = **6.6333**	740/60 = **12.3333**	**5.7**
		2	0 (home)	960/60 = **16**	1050/60 = **17.5**	**1.5**
		3	5 (other)	1180/60 = **19.6667**	1210/60 = **20.1667**	**0.5**
		4	2 (shop)	1215/60 = **20.25**	1235 = **20.5833**	**0.3333**
		5	0 (home)	1240/60 = **20.6667**	**24** (default value)	**3.3333**

Using the information from Table-4.16, the calculations of the **lower** and **upper** bounds for each time, and the shape parameters (*a* and *b*) are exhibited in Table-4.18.

Table-4.18: Calculation of the lower and upper bounds for each time, and the shape parameters (a & b)

HH#	PERSNO	ACTNO	ACTTYPE	ACTSTAR (hours since midnight)	ACTEND (hours since midnight)	DURATION (hours since midnight)
212273	4426	0	0 (home)	**0** a = -1, b =-1	**9.5** ± 0.75 a = 1, b = 1	**9.5** ± 0.75 a = 1, b = 1
		1	1 (work)	**9.5** ± 0.25 a = 1, b = 1	**13.5** ± 0.25 a = 1, b = 1	**4** ± 0.25 a = 1, b = 1
		2	5 (other but taking place at home)	**13.75** ± 0.75 a = 1, b = 1	**17.5** ± 0.75 a = 1, b = 1	**3.75** ± 1 a = 1, b = 1
		3	5 (other)	**19.6667** ± 0.5 a = 1, b = 1	**20.1667** ± 0.5 a = 1, b = 1	**0.5** ± 0.3(**0.5**) a = 1, b = 1
		4	2 (shop)	**20.25** ± 0.5 a = 1, b = 1	**20.5833** ± 0.5 a = 1, b = 1	**0.3333** ± 0.3(**0.3333**) a = 1, b = 1
		5	0 (home)	**20.6667** ± 0.75 a = 1, b = 1	**24** (default value) a = -1, b = -1	24 – **20.6667** ± 0.75 a = 1, b = 1
	4427	0	0 (home)	**0** a = -1, b =-1	**6.6333** ± 0.75 a = 1, b = 1	**6.6333** ± 0.75 a = 1, b = 1
		1	4 (visit)	**6.6333** ± 0.5 a = 1, b = 1	**14** ± 0.5 a = 1, b = 1	**7.3667** ± 0.3(**7.3667**) a = 1, b = 1
		2	0 (home)	**16** ± 0.75 a = 1, b = 1	**17.5** ± 0.75 a = 1, b = 1	**1.5** ± 1 a = 1, b = 1
		3	5 (other)	**19.6667** ± 0.5 a = 1, b = 1	**20.1667** ± 0.5 a = 1, b = 1	**0.5** ± 0.3(**0.5**) a = 1, b = 1
		4	2 (shop)	**20.25** ± 0.5 a = 1, b = 1	**20.5833** ± 0.5 a = 1, b = 1	**0.3333** ± 0.3(**0.3333**) a = 1, b = 1
		5	0 (home)	**20.6667** ± 0.75 a = 1, b = 1	**24** (default value) a = -1, b = -1	24 – **20.6667** ± 0.75 a = 1, b = 1
	4428	0	0 (home)	**0** a = -1, b =-1	**6.6333** ± 0.75 a = 1, b = 1	**6.6333** ± 0.75 a = 1, b = 1
		1	4 (visit)	**6.6333** ± 0.5 a = 1, b = 1	**12.3333** ± 0.5 a = 1, b = 1	**5.7** ± 0.3(**5.7**) a = 1, b = 1
		2	0 (home)	**16** ± 0.75 a = 1, b = 1	**17.5** ± 0.75 a = 1, b = 1	**1.5** ± 1 a = 1, b = 1
		3	5 (other)	**19.6667** ± 0.5 a = 1, b = 1	**20.1667** ± 0.5 a = 1, b = 1	**0.5** ± 0.3(**0.5**) a = 1, b = 1
		4	2 (shop)	**20.25** ± 0.5 a = 1, b = 1	**20.5833** ± 0.5 a = 1, b = 1	**0.3333** ± 0.3(**0.3333**) a = 1, b = 1
		5	0 (home)	**20.6667** ± 0.75 a = 1, b = 1	**24** (default value) a = -1, b = -1	24 – **20.6667** ± 0.75 a = 1, b = 1

The final calculations for the activity times and the shape parameters are displayed in Table-4.19 in **bold** numbers.

Table-4.19: TRANSIMS Activities file (part 1) corresponding to the matched synthetic household ID 26931 (assigned in Sections 4.8.3 and 4.8.4)

HHID	PERID	ACTNO	ACTTYP	PRIORITY	St Time-*a*	St Time-*b*	St-*a*	St-*b*	End T-*a*	End T-*b*	Et-*a*	Et-*b*
26931	55728	1	0 (home)	9	**0**	**0**	**-1**	**-1**	**8.75**	**10.25**	**1**	**1**
		2	1 (work)	9	**9.25**	**9.75**	**1**	**1**	**13.25**	**13.75**	**1**	**1**
		3	5 (other)	9	**13**	**14.5**	**1**	**1**	**16.75**	**18.25**	**1**	**1**
		4	5	9	**19.1667**	**20.1667**	**1**	**1**	**19.6667**	**20.6667**	**1**	**1**
		5	2 (shop)	9	**19.75**	**20.75**	**1**	**1**	**20.0833**	**21.0833**	**1**	**1**
		6	0	9	**19.9167**	**21.4167**	**1**	**1**	**24**	**24**	**-1**	**-1**
	55729	1	0	9	**0**	**0**	**-1**	**-1**	**5.8833**	**7.3833**	**1**	**1**
		2	4 (visit)	9	**6.1333**	**7.1333**	**1**	**1**	**13.5**	**14.5**	**1**	**1**
		3	0	9	**15.25**	**16.75**	**1**	**1**	**16.75**	**18.25**	**1**	**1**
		4	5	9	**19.1667**	**20.1667**	**1**	**1**	**19.6667**	**20.6667**	**1**	**1**
		5	2	9	**19.75**	**20.75**	**1**	**1**	**20.0833**	**21.0833**	**1**	**1**
		6	0	9	**19.9167**	**21.4167**	**1**	**1**	**24**	**24**	**-1**	**-1**
	55730	1	0	9	**0**	**0**	**-1**	**-1**	**5.8833**	**7.3833**	**1**	**1**
		2	4	9	**6.1333**	**7.1333**	**1**	**1**	**11.8333**	**12.8333**	**1**	**1**
		3	0	9	**15.25**	**16.75**	**1**	**1**	**16.75**	**18.25**	**1**	**1**
		4	5	9	**19.1667**	**20.1667**	**1**	**1**	**19.6667**	**20.6667**	**1**	**1**
		5	2	9	**19.75**	**20.75**	**1**	**1**	**20.0833**	**21.0833**	**1**	**1**
		6	0	9	**19.9167**	**21.4167**	**1**	**1**	**24**	**24**	**-1**	**-1**

Table-4.20: TRANSIMS Activities file (part 2) corresponding to the matched synthetic household ID 26931 (assigned in Sections 4.8.3 and 4.8.4)

Dur Tim-*a*	Dur Tim-b	D-*a*	D-*b*	Mode	VEHid	N Loc	Location	N Others	OthersID	G Number
8.75	**10.25**	**1**	**1**	1 (walk)	-1 (no vehicle)	1	841405 (home, obtained directly from the synthetic household)	0	-	1
3.75	**4.25**	**1**	**1**	2 (car)	45554 (from vehicle file)	1		0	-	1
2.75	**4.75**	**1**	**1**	2	45554	1	841405	0	-	1
0.35	**0.65**	**1**	**1**	2	45554	1		3	55728, 55729, and 55730	1
0.2333	**0.4333**	**1**	**1**	2	45554	1		3	55728, 55729, and 55730	1
2.5833	**4.0833**	**1**	**1**	2	45554	1	841405	3	55728, 55729, and 55730	1
5.8833	**7.3833**	**1**	**1**	1	-1	1	841405 (home)	0	-	1
5.1567	**9.5767**	**1**	**1**	1	-1	1		0	-	1
0.5	**2.5**	**1**	**1**	1	-1	1	841405	0	-	1
0.35	**0.65**	**1**	**1**	2	45554	1		3	55728, 55729, and 55730	1
0.2333	**0.4333**	**1**	**1**	2	45554	1		3	55728, 55729, and 55730	1
2.5833	**4.0833**	**1**	**1**	2	45554	1	841405	3	55728, 55729, and 55730	1
5.8833	**7.3833**	**1**	**1**	1	-1	1	841405	0	-	1
3.99	**7.41**	**1**	**1**	1	-1	1		0	-	1
0.5	**2.5**	**1**	**1**	1	-1	1	841405	0	-	1
0.35	**0.65**	**1**	**1**	2	45554	1		3	55728, 55729, and 55730	1
0.2333	**0.4333**	**1**	**1**	2	45554	1		3	55728, 55729, and 55730	1
2.5833	**4.0833**	**1**	**1**	2	45554	1	841405	3	55728, 55729, and 55730	1

4.8.5 Generating non-home activity locations

Each activity in the matched synthetic household must have a specific location within the region under consideration. The following steps describe how the location generation algorithm assigns such a location. First, a *discrete choice model* is used to select an appropriate zone for each activity. This is carried out based on the *zone data* that specifies the attractiveness of each activity type within each zone. Second, a location for each particular activity is randomly specified within the selected zone using the weights given in the *activity location table.*

4.8.5.1 Selecting the zones

The procedure to select the zone for each non-home activity location is conducted in two steps. The first step selects the zones for *work* activities. The second step selects the zones for all *other* types of activities, which are assumed to occur between a pair of:

- *home*-to-*work* activities, or
- *work*-to-*home* activities, or
- *work*-to-*work* activities, or
- *home*-to-*home* activities.

STEP 1: Selecting the zones for work activities

For each *work* activity, the Activity Generator Module randomly selects a location zone L according to the following probability distribution:

$$P_L = \frac{(at)_L \exp(b_m (t_{H,L})^k)}{\sum_{L'} (at)_{L'} \exp(b_m (t_{H,L'})^k)}$$

where L = destination zone for the particular work activity,
H = home location,
$(at)_L$ = attractor for work activity in zone L,
$(t_{H,L})$ = travel time on the link from the home location H to the work location in zone L, and
b_m = coefficient for travel-mode choice m used for arriving at the work location.
k = constant. Default is 1.0.

Note that the sum L' is over all zones within the region encompassing the possible zones for work activities. The attractor $(at)_L$ is obtained from the zone data (see Section 4.4.4). The travel time $(t_{H,L})$ is obtained from the feedback derived from the Traffic Microsimulator and Route Planner Modules, which is updated and corrected by the Selector Module. At the beginning of the simulation, there is no feedback input from these modules to the Activity Generator Module. In that case, the free-flow speed and the length of each link are used to estimate the travel times on the links by dividing the link length by the admissible free-flow speed. The link length and the free-flow speed information are obtained from the *link table* in the TRANSIMS network.

The travel-mode coefficient $\boldsymbol{b_m}$ is obtained from the file containing the coefficients for each activity type and for each travel mode used for arriving at the activity location. Note that travel mode (*m*) is taken directly from the skeletal pattern (see Section 4.8.1). Table-4.21 provides an example of the file, which has coefficients for seven activity types. Each activity type has two travel modes (*walk* and *car*) used for arriving at the activity location.

Table-4.21: An example of the file containing the travel-mode coefficients

Coefficient value (b_m)	Activity type	Travel mode (*m*) used for arriving at the activity location
-0.00008	0 (home)	1 (walk)
-0.00008	1 (work)	1
-0.00008	2 (shop)	1
-0.00008	3 (school)	1
-0.00008	4 (visit)	1
-0.00008	5 (other)	1
-0.00008	6 (serve passenger)	1
-0.00004	0	2 (car)
-0.00004	1	2
-0.00004	2	2
-0.00004	3	2
-0.00004	4	2
-0.00004	5	2
-0.00004	6	2

STEP 2: Selecting the zones for other activities

To select zones for the other activities, the Activity Generator Module uses a *trip-chaining discrete choice model.* This model basically uses the same concept as the model used for selecting zones for the work activities, but considers a *trip chain* (obtained from the skeletal pattern in Section 4.8.1) rather than a single trip.

For example, consider a trip chain that consists of two locations within zones L_1 and L_2 to be visited between work and home. From the work location in zone L, it is required to "visit" a location in zone L_1 via mode m_1, and then to make a second stop to "shop" at a location in zone L_2 via mode m_2, and finally, to return home H via mode m_3. The first zone L_1 for the visiting-activity type is selected with a probability:

$$P_{L1} = \frac{(at)_{L1,v} \exp((b_{m1}(t_{L,L1})^k + b_{m3}(t_{L1,H})^k)}{\sum_{L1'} (at)_{L1',v} \exp((b_{m1}(t_{L,L1'})^k + b_{m3}(t_{L1',H})^k)}$$

where the attractor $(at)_{L1}$ is defined for zone L_1 and activity type v (*visit*), and where $t_{L,L1}$ and $t_{L1,H}$ are the travel times from the *work* location in zone L to the *visit* location in zone L_1, and from the *visit* location in zone L_1 to the *home* location H respectively. The coefficient b_{m1} and b_{m3} are obtained from files such as the one provided in Table-4.21.

After zone L_1 is chosen, it replaces the work location in zone L as the first anchor location that is subsequently used for choosing the "shop" location in zone L_2. This choice is based on the probability distribution:

$$P_{L2} = \frac{(at)_{L2,s} \exp((b_{m2}(t_{L1,L2})^k + b_{m3}(t_{L2,H})^k)}{\sum_{L2'} (at)_{L2',s} \exp((b_{m2}(t_{L1,L2'})^k + b_{m3}(t_{L2',H})^k)}$$

where the different parameters are defined similar to those for P_{L2} above except now "*s*" stands for *shop* instead of "*v*" for *visit.*

4.8.5.2 Choosing destinations within the selected zones

After the zones are selected for all the activities, the specific destinations within the zones are selected at random according to weights given in the *activity location table* (see Section 4.4.5).

For example, for a work zone L, if an employment-based weight w_A is ascribed to each location A, then location A would be selected with the probability $p(A) = \frac{w_A}{\sum_{A' \in L} w_{A'}}$.

This process is repeated until the locations for all of the activities in the matched synthetic household are ascertained.

Example for selecting destinations

For the sake of illustration, we provide in Figure-4.27 a flow chart for the activity pattern of the matched synthetic household ID 26931 corresponding to the skeletal activity pattern shown in Table-4.22 as previously discussed.

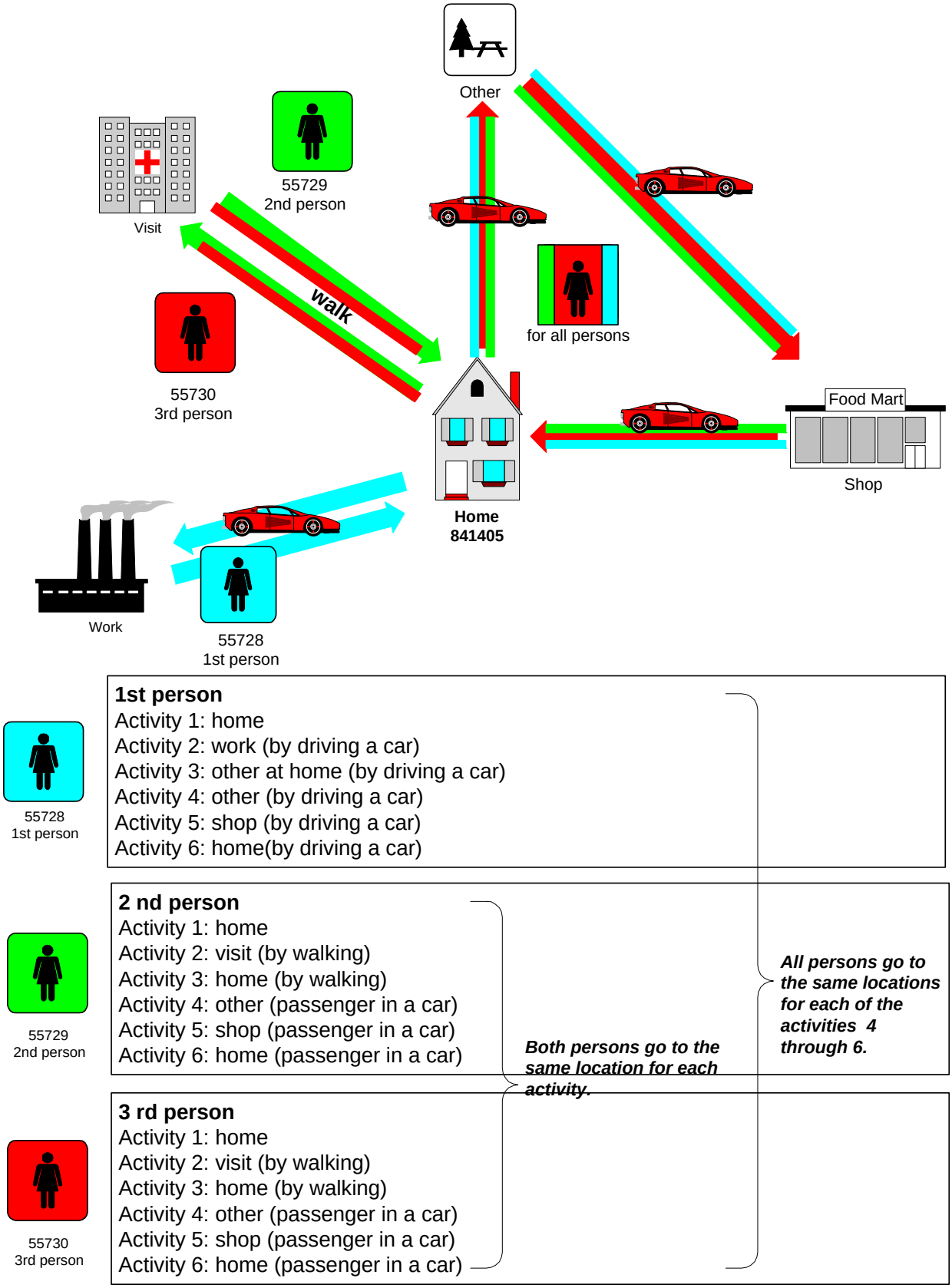


Figure-4.27: Flow chart of the activity pattern for the matched synthetic household ID 26931

As shown in Figure-4.27, the first person (ID 55728) is the only worker in the household, and executes the trip to work by car independently, which is represented by activity 2. The first person shares the fourth, fifth, and sixth activities together with the second and third person in the household, which are the *other, shop,* and *home* activities. These three activities are conducted by using a car. The second and third person also share the first three activities together, which are *home, visit,* and *home* activities, which are done by walking.

The specific location of the home is obtained directly from the synthetic household demographic data (Table-4.22). Hence, we need to locate only the following:

1. **work** location for the second activity of the first person,
2. **visit** location for the second activity of the second and third persons,
3. **other** location, for the fourth activity of each person in the household, and
4. **shop** location for the fifth activity of each person in the household.

Determining these four locations will be discussed in the following sections; zone selection followed by activity location selection.

First, we have to select the zone for each activity location (except the home zone). Suppose that we have **ten** zones within the study area and that we are given the attractor variables $(at)_L$ for each activity type (see Table-4.22), as well as the layout of the ten zones as shown in Figure-4.28.

Table-4.22: Attractor variables $(at)_L$ for each of activity type within ten zones

ZONE	EASTING	NORTHING	$(at)_L$ for each ACTIVITY TYPE					
			WORK	SHOP	SCHOOL	VISIT	OTHER	SERVE_PASS
1	50	16050	2	2	0.0001	0.0001	2	1
2	50	19050	2	2	0.0001	0.0001	2	1
3	16050	35050	2	2	0.0001	0.0001	2	1
4	19050	35050	2	2	0.0001	0.0001	2	1
5	35050	19050	2	2	0.0001	0.0001	2	1
6	35050	16050	2	2	0.0001	0.0001	2	1
7	19050	50	2	2	0.0001	0.0001	2	1
8	16050	50	2	2	0.0001	0.0001	2	1
9	50	50	0.0001	0.0001	0.0001	0.0001	0.0001	1
10	17500	17500	0.0001	0.0001	3.71	3.71	0.0001	1

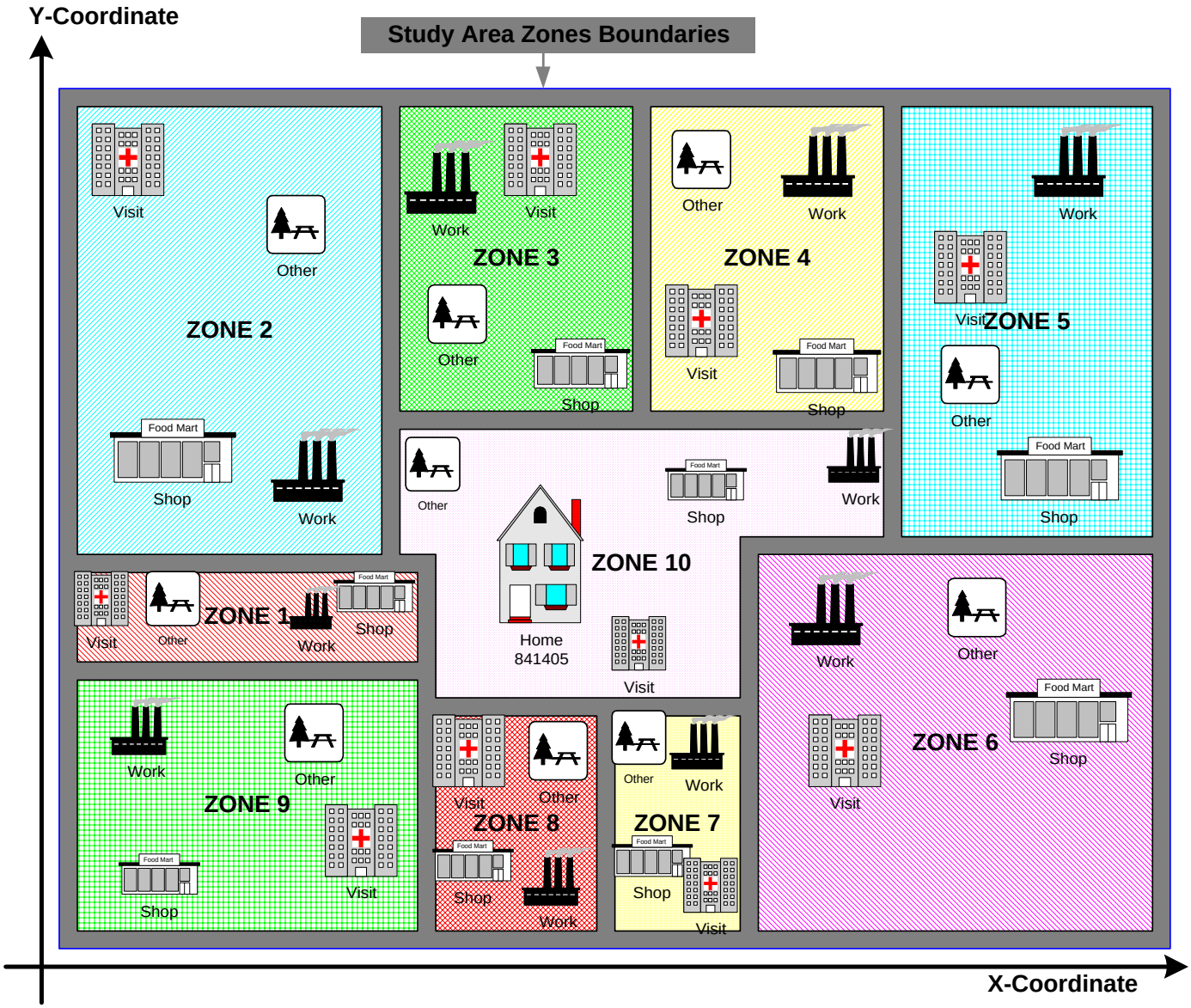


Figure-4.28: Layout of the ten zones within the study area

Note that each zone has its own locations for the various activities, and that the home location is located in zone 10. We are also given the information regarding the coefficient $\boldsymbol{b_m}$ for the travel-mode choice m used for arriving at the activity location (see Table-4.23). Moreover, the information regarding the travel time $\boldsymbol{t_{i,\,j}}$ for each pair of activity locations, which are derived from the Selector Module, is given in Table-4.23.

Table-4.23: Travel times $t_{i,j}$

Starting location, *i*	Ending location, *j*	Travel time $t_{i,j}$ (minutes)									
		ZONE 1	ZONE 2	ZONE 3	ZONE 4	ZONE 5	ZONE 6	ZONE 7	ZONE 8	ZONE 9	ZONE 10
Home	Work (by car)	7	10	15	17	27	16	9	11	16	15
Home	Other (by car)	65	100	110	112	120	96	110	109	99	50
Other	Home (by car)	40	70	75	80	88	71	73	72	76	35
Other*	Shop* (by car)	Not available at this time	Not available at this time	Not available at this time	Not available at this time	Not available at this time	Not available at this time	Not available at this time	Not available at this time	Not available at this time	Not available at this time *
Shop	Home (by car)	5	7	6	8	9	8	8	7	7	6
Home	Visit (by walk)	12	32	15	10	35	31	11	9	13	5
Visit	Home (by walk)	12	32	15	10	35	31	11	9	13	5

Note that currently the travel times from **other location to **shop location** are not available at this step because we have to locate the zone for the **other location** first.*

Location #1. Selecting a zone for the work activity.

According to Tables 4.21 through 4.23, we can compute the probability for selecting the work zone for the ten zones as follows using the equation:

$$P_{Li} = \frac{(at)_{Li}\exp(b_{car}t_{H,Li})}{\sum_{i=1}^{10}(at)_{Li}\exp(b_{car}t_{H,Li})} \qquad \forall\, i.$$

	ZONE 1 ***i* = 1**	ZONE 2 ***i* = 2**	ZONE 3 ***i* = 3**	ZONE 4 ***i* = 4**	ZONE 5 ***i* = 5**	ZONE 6 ***i* = 6**	ZONE 7 ***i* = 7**	ZONE 8 ***i* = 8**	ZONE 9 ***i* = 9**	ZONE 10 ***i* = 10**
$(at)_{Li}\exp(b_{car}(t_{H,Li}))$	1.99902*	1.99860	1.99790	1.99762	1.99622	1.99776	1.99874	1.99846	0.00010	0.00010
P_{Li}	0.12506*	0.12503	0.12499	0.12497	0.12488	0.12498	0.12504	0.12502	0.00001	0.00001
$\sum_{i=1}^{10}(at)_{Li}\exp(b_{car}(t_{H,Li}))$ = **15.98453**										

From Table 4.22 From Table 4.21 From Table 4.23

*For example, $P_{L1} = \dfrac{2\times\exp(-0.00004\times 7)}{15.98453} = \dfrac{1.99902}{15.98453} = 0.12506.$

We **randomly** select a zone L_i according to the calculated probabilities by using the associated cumulative distribution (see Figure-4.29). This is done by generating a random number between 0 and 1, and selecting that zone for which the random number falls within its defined interval.

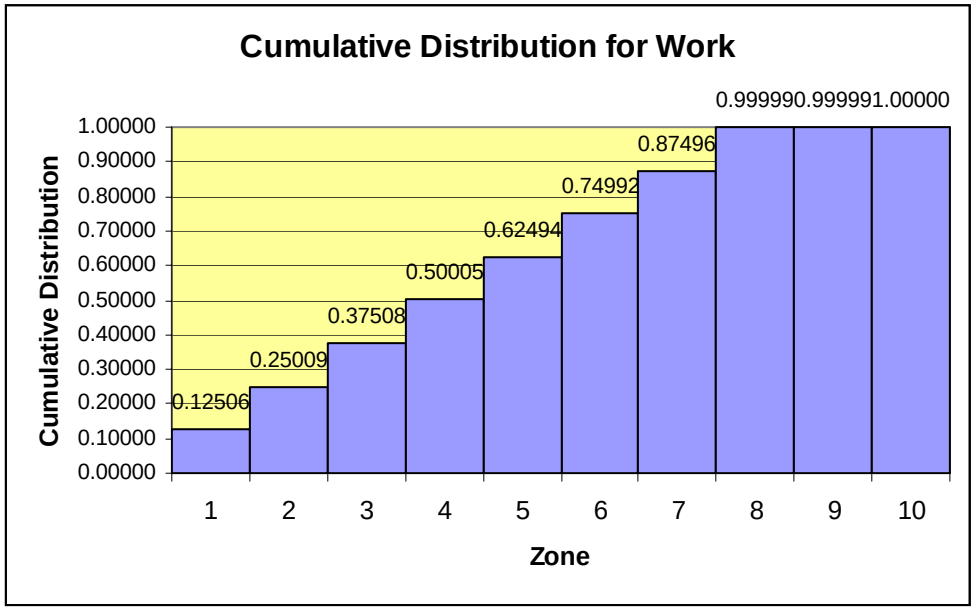


Figure-4.29 Cumulative distribution for selecting a work zone

For example, suppose that we select a random number **0.2025**. Then, the selected **work location** would be in **Zone 2**.

Location #2. Selecting the zone for *visit, other, and shop* activities.

Location #3. Selecting a zone for the *visit location.*

According to the skeletal activity pattern, the second and the third person **walk** to the same **visit location**. The trip to the visit location occurs between a pair of **home** to **home** activities. Hence, the probability distribution used for selecting the zone for the **visit** location is defined via:

$$P_{Li} = \frac{(at)_{Li,visit} \exp(b_{walk} t_{H,Li} + b_{walk} t_{Li,H})}{\sum_{i=1}^{10} (at)_{Li,visit} \exp(b_{walk} t_{H,Li} + b_{walk} t_{Li,H})} \qquad \forall i.$$

According to Tables 4.21 through 4.23, we can compute the probability for selecting the **visit** zone in each of the ten zones as follows.

	ZONE 1 $i = 1$	**ZONE 2** $i = 2$	**ZONE 3** $i = 3$	**ZONE 4** $i = 4$	ZONE 5 $i = 5$	ZONE 6 $i = 6$	ZONE 7 $i = 7$	ZONE 8 $i = 8$	ZONE 9 $i = 9$	ZONE 10 $i = 10$
$(at)_{Li,visit}$ exp(...)	9.98E-05*	9.95E-05	9.98E-05	9.98E-05	9.94E-05	9.95E-05	9.98E-05	9.99E-05	9.98E-05	3.70703
$P(L_i)$	0.00003*	0.00003	0.00003	0.00003	0.00003	0.00003	0.00003	0.00003	0.00003	0.99976
$\sum_{i=1}^{10}(at)_{Li,visit}\exp(b_{walk}t_{H,Li}+b_{walk}t_{Li,H})*10^4 =$ **3.70793**										

Note that Zone 10 has a very high probability of being selected as compared with the other zones because of its relatively higher value of the attractor parameter $at_{Li,Visit}$ (3.71 as compared with 0.0001 for the other zones).

*For example, $P_{L1} = \dfrac{0.0001\times\exp(-0.00008\times12-0.00008\times12)}{3.70793} = \dfrac{9.98E-05}{3.70793} = 0.00003.$

A zone L_i is then selected randomly according to the calculated probabilities by using the associated cumulative distribution (see Figure-4.30). This is done by generating a random number between 0 and 1, and selecting that zone for which the random number falls within its defined interval.

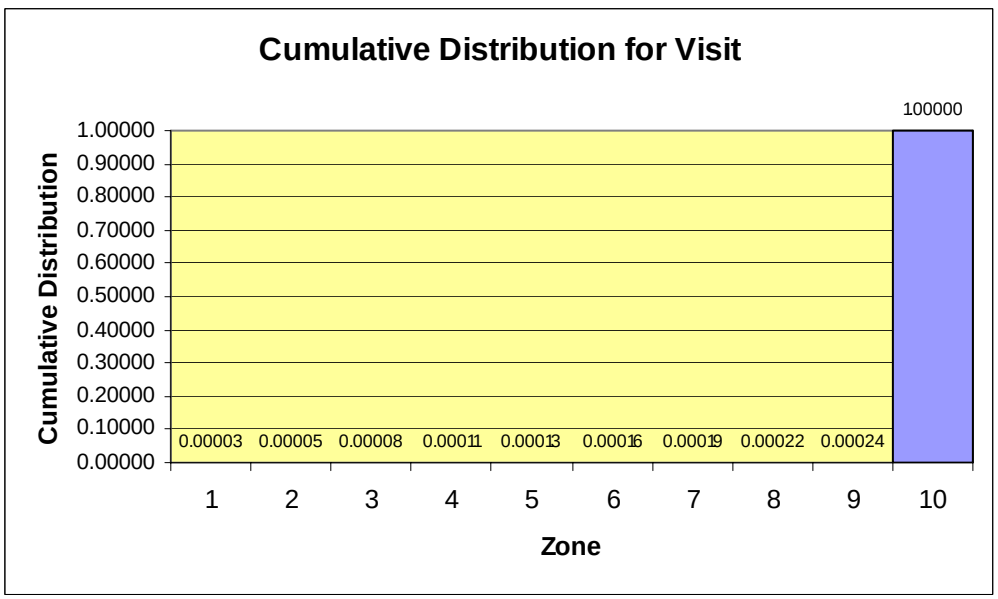


Figure-4.30: Cumulative distribution for selecting a visit zone

For example, suppose that we select a random number **0.7916**. Then, the selected **visit location** would be in **Zone 10**.

Location #4. Selecting a zone for the *other* and *shop locations.*

According to the skeletal activity pattern, every person in the household **uses the same car** to traverse the same trip chain, which is comprised of going to the **other location** and then to the **shop location**. The trips to the **other** and **shop** locations occur between a pair of **home** to **home** activities. Hence, the probabilities used for selecting a zone for the **other** location are given by:

$$P_{Li} = \frac{(at)_{Li,other}\ \exp(b_{car}t_{H,Li} + b_{car}t_{Li,H})}{\sum_{i=1}^{10}(at)_{Li,other}\ \exp(b_{car}t_{H,Li} + b_{car}t_{Li,H})} \qquad \forall\, i,$$

and the probabilities used for selecting a zone for the **shop** location are given by:

$$P_{Li} = \frac{(at)_{Li,shop}\ \exp(b_{car}t_{Other,Li} + b_{car}t_{Li,H})}{\sum_{i=1}^{10}(at)_{Li,shop}\ \exp(b_{car}t_{Other,Li} + b_{car}t_{Li,H})} \qquad \forall\, i.$$

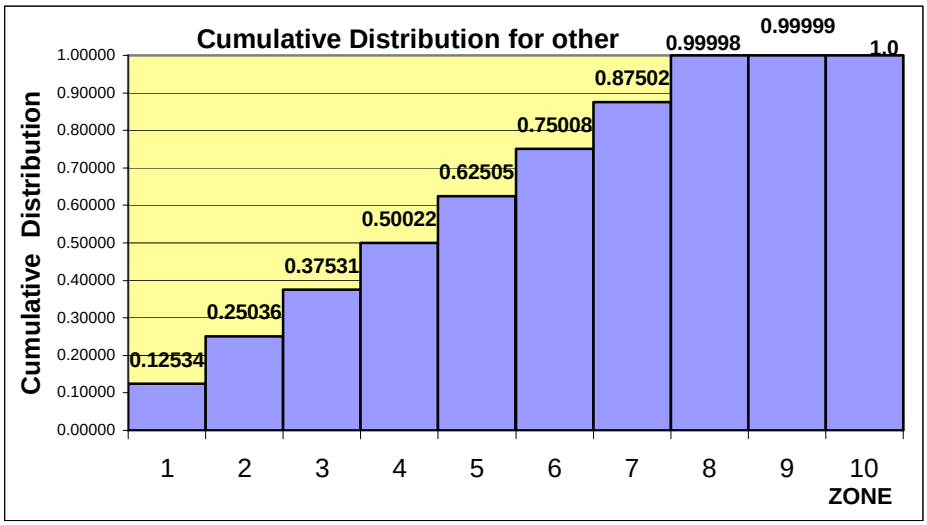
According to Tables 4.21 through 4.23, we can calculate probability for selecting the **other** zone for the ten zones as follows.

	ZONE 1 ***i* = 1**	ZONE 2 ***i* = 2**	**ZONE 3 *i* = 3**	ZONE 4 ***i* = 4**	ZONE 5 ***i* = 5**	ZONE 6 ***i* = 6**	ZONE 7 ***i* = 7**	ZONE 8 ***i* = 8**	ZONE 9 ***i* = 9**	ZONE 10 ***i* = 10**
$(at)_{Li,other}\ \exp(...)$	1.99162*	1.98645	1.98525	1.98470	1.98343	1.98668	1.98541	1.98557	0.00010	0.00010
$p(L_i)$	0.12534*	0.12502	0.12494	0.12491	0.12483	0.12503	0.12495	0.12496	0.00001	0.00001
$\sum_{i=1}^{10}(at)_{Li,other}\ \exp(b_{car}t_{H,Li} + b_{car}t_{Li,H})$ = **15.88932**										

Note that the zones 9 and 10 have a very low probability of being selected as compared with the other zones because of their relatively lower values of the attractor parameter $(at)_{Li,Other}$ (0.0001 as compared with 2 for the other zones).

*For example, $P_{L1} = \dfrac{2\times(\exp(-0.00004\times 65 - 0.00004\times 40)}{15.88932} = \dfrac{1.99162}{15.88932} = 0.12534.$

A zone L_i is randomly selected according to the calculated probabilities by using the associated cumulative distribution (see Figure-4.31). This is done by generating a random number between 0 and 1, and selecting that zone for which the random number falls within its defined interval.



Figur-4.31: Cumulative distribution for selecting an other zone

For example, suppose that we select a random number **0.3010**. Then, the selected location of **other trip** would be in **Zone 3**.

Finally, we will select a zone for the ***shop* location.** In addition to the information given in Table-4.24, we need to find the travel time from the ***other* location** to the ***shop* location**, which can be obtained at this step because we have already selected a zone for the ***other* location.**

Table-4.24 (updated): Travel times t(i, j)

Starting location, *i*	Ending location, *j*	Travel time ***t(i, j)*** (minutes)									
		ZONE 1	ZONE 2	ZONE 3	ZONE 4	ZONE 5	ZONE 6	ZONE 7	ZONE 8	ZONE 9	ZONE 10
Home	Work (by car)	7	10	15	17	27	16	9	11	16	15
Home	Other (by car)	65	100	110	112	120	96	110	109	99	50
Other	Home (by car)	40	70	75	80	88	71	73	72	76	35
Other*	Shop* (by car)	**9**	**4**	**3**	**10**	**21**	**15**	**30**	**27**	**35**	**6**
Shop	Home (by car)	5	7	6	8	9	8	8	7	7	6
Home	Visit (by walk)	12	32	15	10	35	31	11	9	13	5
Visit	Home (by walk)	12	32	15	10	35	31	11	9	13	5

According to Tables 4.21 through 4.23 *(the updated one)*, we can compute the probabilities for selecting the **shop** zone for the ten zones as follows.

	ZONE 1 $i = 1$	ZONE 2 $i = 2$	ZONE 3 $i = 3$	ZONE 4 $i = 4$	ZONE 5 $i = 5$	ZONE 6 $i = 6$	ZONE 7 $i = 7$	ZONE 8 $i = 8$	ZONE 9 $i = 9$	ZONE 10 $i = 10$
exp(b_{car}….)	2.99944*	2.99956	2.99964	2.99928	2.99880	2.99908	2.99848	2.99864	0.99842	0.99962
$p(L_i)$	0.11540*	0.11541	0.11541	0.11540	0.11538	0.11539	0.11537	0.11537	0.03841	0.03846
$\sum_{i=1}^{10} \exp(b_{car} \ldots.)$ = **25.99097**										

Note that the zones 9 and 10 have a very low probability of being selected as compared with the other zones because of their relatively lower values of the attractor parameter $at_{Li,Shop}$ (0.0001 as compared with 2 for the other zones).

*For example, $P_{L1} = \dfrac{2 \times \exp(-0.00004 \times 9 - 0.00004 \times 5)}{25.99097} = \dfrac{2.99944}{25.99097} = 0.11540.$

A zone L_i is randomly selected according to the calculated probabilities by using the associated cumulative distribution (see Figure-4.32). This is done by generating a random number between 0 and 1, and selecting that zone for which the random number falls within its defined interval.

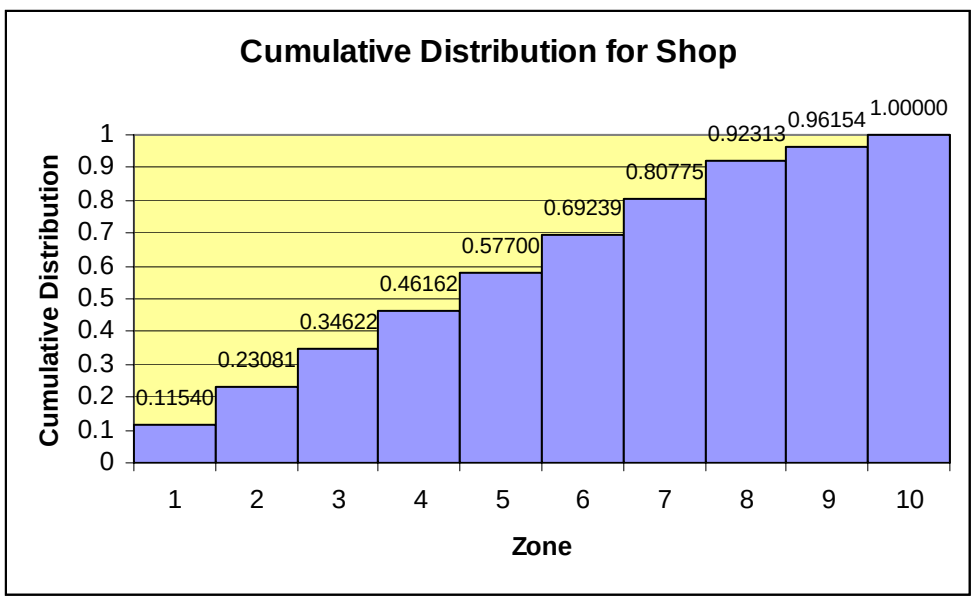


Figure-4.32: Cumulative distribution for selecting a shop zone

For example, suppose that we select a random number **0.6534**. Then, the selected location for the **shop** trip would be in **Zone 6**. Figure- 4.33 shows the results for selecting the zones that correspond to the activity patterns of household ID 26931.

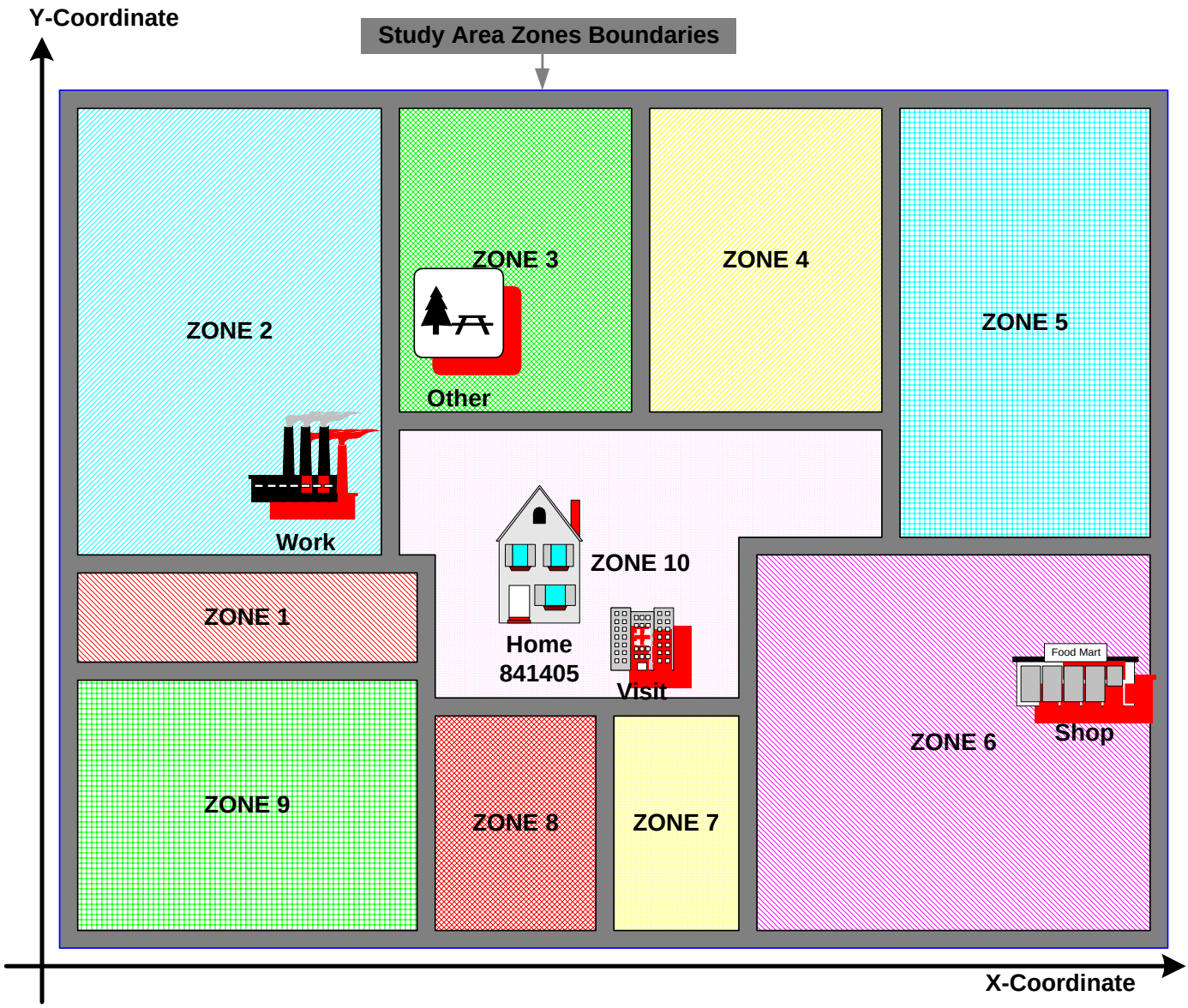


Figure-4.33: Results of selecting the zones corresponding to the activity patterns for household ID 26931

Next we need to specify the destination choice within the respective selected zones. Suppose that the information regarding the relative weights for the locations within each zone are obtained from the same table in the TRANSIMS network (Table-4.25). (Generally, each zone would have its own weight file.)

Table-4.25: Activity location table obtained from the TRANSIMS network

Location ID	Node	Link	Offset	Layer	Easting	Northing	Elevation	**Weights for work**	**Weights for visit**	**Weights for other**	**Weights for shop**	Notes
857532	213	1212	200	Auto	2000	700	1000	**5**	**1**	**3**	**1**	
853363	334	1657	300	Walk	1500	1000	700	**2**	**5**	**2**	**2**	
862272	512	7654	121	Auto	100	600	1000	**1**	**2**	**5**	**3**	
843700	167	2341	156	Auto	50	150	500	**3**	**3**	**1**	**5**	
888999	965	7415	312	Bus	2000	1500	400	**4**	**4**	**4**	**4**	
							SUM for a specific zone:	15	15	15	15	

For the sake of illustration, Table-4.26 displays the computation of the probability of selecting each location within each of the relevant zones.

Table-4.25: Probability for specifying the activity locations within the selected zone

Location ID	**Probability** $p(A_i)$			
	work	**visit**	**other**	**shop**
857532	5/15 = **0.3333**	1/15 = **0.0667**	3/15 = **0.2**	1/15 = **0.0667**
853363	2/15 = **0.1333**	5/15 = **0.3333**	2/15 = **0.1333**	2/15 = **0.1333**
862272	1/15 = **0.0667**	2/15 = **0.1333**	5/15 = **0.3333**	3/15 = **0.2**
843700	3/15 = **0.2**	3/15 = **0.2**	1/15 = **0.0667**	5/15 = **0.3333**
888999	4/15 = **0.2667**	4/15 = **0.2667**	4/15 = **0.2667**	4/15 = **0.2667**

Location #1. Specifying the *work* location (within the selected zone 2).

According to the calculated probabilities, we randomly select a location ID within a zone by using the associated cumulative distribution (see Figure-4.34). This is done by generating a random number between 0 and 1, and selecting that location ID within the zone for which the random number falls within its defined interval.

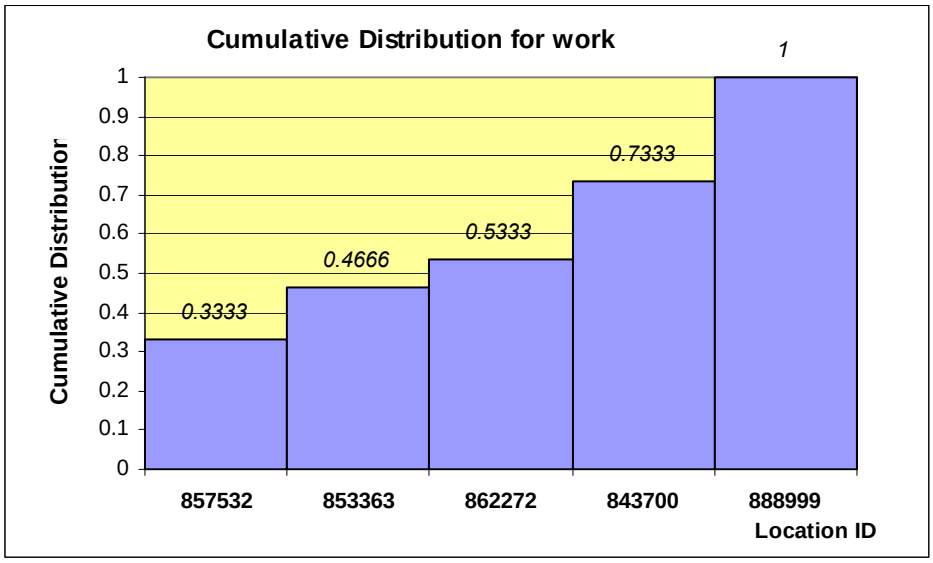


Figure-4.34: Cumulative distribution for specifying the work location within a zone

For example, suppose that we select a random number **0.1231**. Then, the **work** location ID selected would be **857532**.

Location #2. Specifying the *visit* location (within the selected zone 10).

Location ID for the visit trip is selected within the zone according to the calculated probabilities by using the associated cumulative distribution (see Figure-4.35). This is done by generating a random number between 0 and 1, and selecting the location ID for which the random number falls within its defined interval.

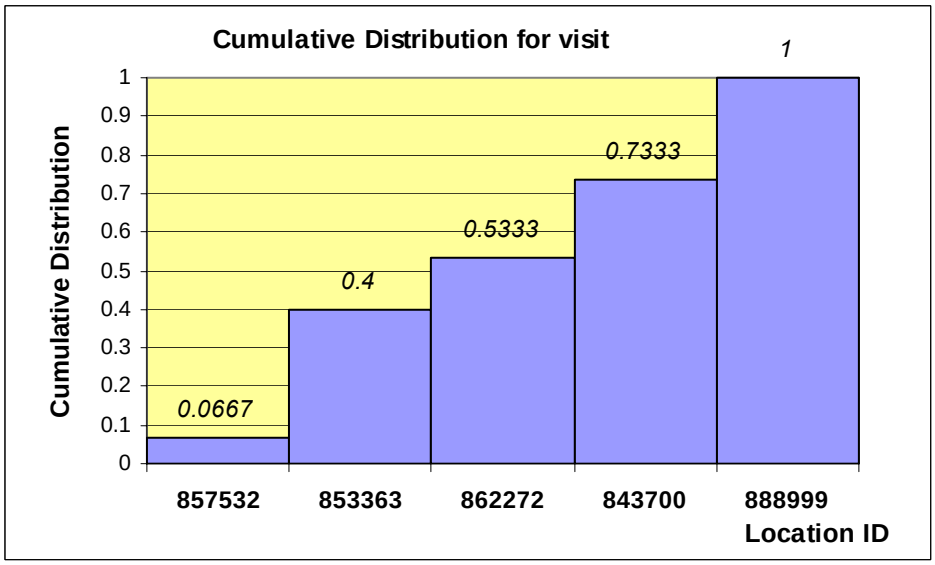


Figure-4.35: Cumulative distribution for specifying the visit location within a zone

For example, suppose that we select a random number **0.3911**. Then, the **visit** location ID selected would be **853363**.

Location #3. Specifying the *other* location (within the selected zone 3).

Similarly, we **randomly** select a location ID according to the calculated probabilities by using the associated cumulative distribution (see Figure-4.36). This is done by generating a random number between 0 and 1, and selecting that location ID for which the random number falls within its defined interval.

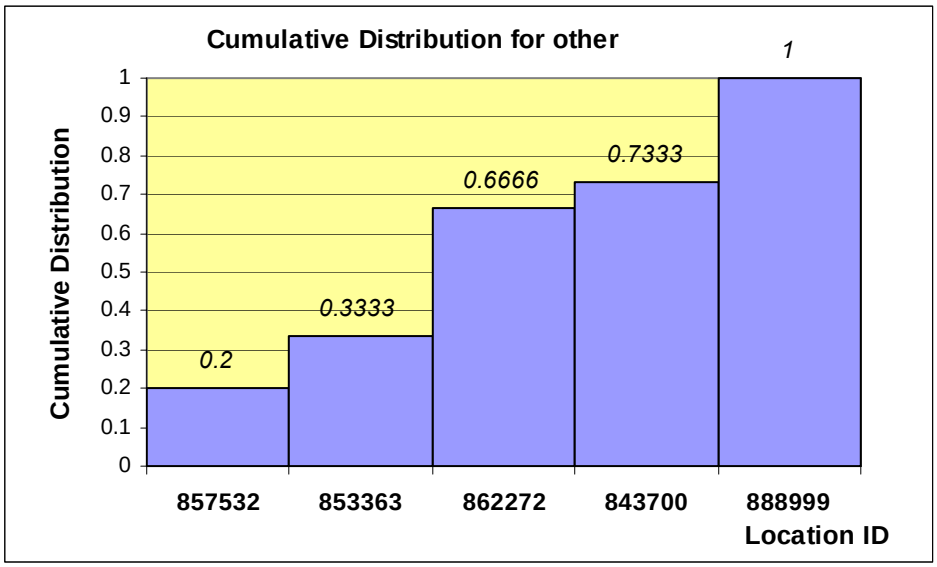


Figure-4.36: Cumulative distribution for specifying the other trip location within a zone

For example, suppose that we select a random number **0.5498**. Then, the **other** trip location ID selected would be **862272**.

Location #4. Specifying the *shop* location (within the selected zone 6).

The same procedure applies for selecting the shop trip location within zone 6 by calculating the associated cumulative distribution as shown in Figure-4.37. This is done by generating a random number between 0 and 1, and selecting the location ID for which the random number falls within its defined interval.

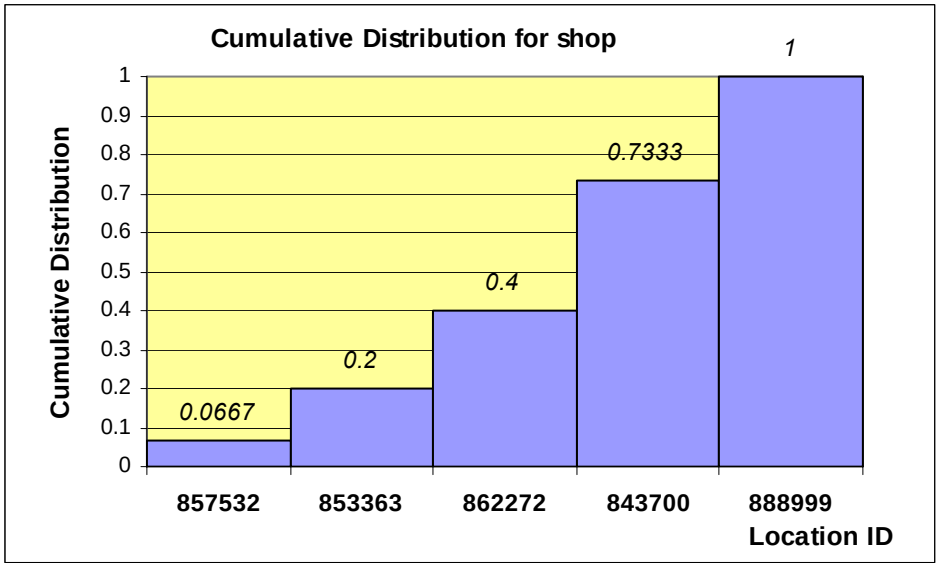


Figure-4.37: Cumulative distribution for specifying the shop location within the zone

For example, suppose that we select a random number **0.70001**. Then, the **shop location ID** selected would be **843700**. Figure-4.38 shows the final results for selecting the various activity locations for household ID 26931.

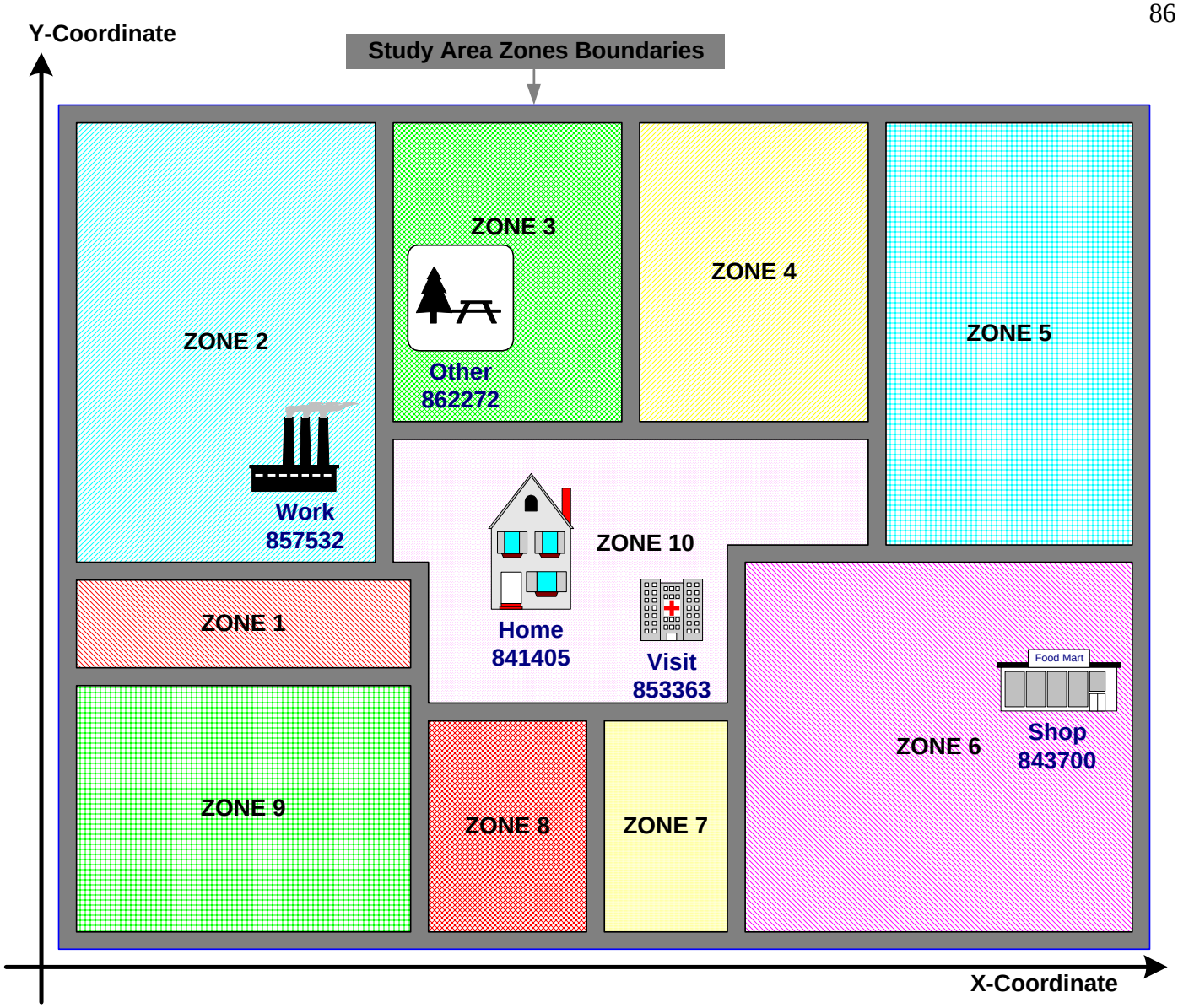


Figure-4.38: Final results of selecting the various activity locations

From all of the examples in Sections 4.8.1 to 4.8.5, the corresponding outputs generated by the Activity Generator Module are shown in Table-4.27. Note that this table displays only the outputs for the household ID 26931.

Table-4.27: Example of a TRANSIMS Activities file (part 1) corresponding to the examples from Sections 4.8.2 to 4.8.5

HHID	PERID	ACTNO	ACTTYP	PRIORITY	St Time-*a*	St Time-*b*	St-*a*	St-*b*	End T-*a*	End T-*b*	Et-*a*	Et-*b*
26931	55728	1	0 (home)	9	0	0	-1	-1	8.75	10.25	1	1
		2	1 (work)	9	9.25	9.75	1	1	13.25	13.75	1	1
		3	5 (other)	9	13	14.5	1	1	16.75	18.25	1	1
		4	5	9	19.1667	20.1667	1	1	19.6667	20.6667	1	1
		5	2 (shop)	9	19.75	20.75	1	1	20.0833	21.0833	1	1
		6	0	9	19.9167	21.4167	1	1	24	24	-1	-1
	55729	1	0	9	0	0	-1	-1	5.8833	7.3833	1	1
		2	4 (visit)	9	6.1333	7.1333	1	1	13.5	14.5	1	1
		3	0	9	15.25	16.75	1	1	16.75	18.25	1	1
		4	5	9	19.1667	20.1667	1	1	19.6667	20.6667	1	1
		5	2	9	19.75	20.75	1	1	20.0833	21.0833	1	1
		6	0	9	19.9167	21.4167	1	1	24	24	-1	-1
	55730	1	0	9	0	0	-1	-1	5.8833	7.3833	1	1
		2	4	9	6.1333	7.1333	1	1	11.8333	12.8333	1	1
		3	0	9	15.25	16.75	1	1	16.75	18.25	1	1
		4	5	9	19.1667	20.1667	1	1	19.6667	20.6667	1	1
		5	2	9	19.75	20.75	1	1	20.0833	21.0833	1	1
		6	0	9	19.9167	21.4167	1	1	24	24	-1	-1

Table-4.28: Example of a TRANSIMS Activities file (part 2) corresponding to the examples from Sections 4.8.2 to 4.8.5

Dur Tim-*a*	Dur Tim-b	D-*a*	D-*b*	Mode	VEHid	N Loc	Location	N Others	OthersID	G Number
8.75	10.25	1	1	1 (walk)	-1 (no vehicle)	1	841405 (home)	0	-	1
3.75	4.25	1	1	2 (car)	45554 (from vehicle file)	1	**857532**	0	-	1
2.75	4.75	1	1	2	45554	1	841405	0	-	1
0.35	0.65	1	1	2	45554	1	**862272**	3	55728, 55729, and 55730	1
0.2333	0.4333	1	1	2	45554	1	**843700**	3	55728, 55729, and 55730	1
2.5833	4.0833	1	1	2	45554	1	841405	3	55728, 55729, and 55730	1
5.8833	7.3833	1	1	1	-1	1	841405 (home)	0	-	1
5.1567	9.5767	1	1	1	-1	1	**853363**	0	-	1
0.5	2.5	1	1	1	-1	1	841405	0	-	1
0.35	0.65	1	1	2	45554	1	**862272**	3	55728, 55729, and 55730	1
0.2333	0.4333	1	1	2	45554	1	**843700**	3	55728, 55729, and 55730	1
2.5833	4.0833	1	1	2	45554	1	841405	3	55728, 55729, and 55730	1
5.8833	7.3833	1	1	1	-1	1	841405	0	-	1
3.99	7.41	1	1	1	-1	1	**853363**	0	-	1
0.5	2.5	1	1	1	-1	1	841405	0	-	1
0.35	0.65	1	1	2	45554	1	**862272**	3	55728, 55729, and 55730	1
0.2333	0.4333	1	1	2	45554	1	**843700**	3	55728, 55729, and 55730	1
2.5833	4.0833	1	1	2	45554	1	841405	3	55728, 55729, and 55730	1