Alternative Methodology To Household Activity Matching In TRANSIMS

Rajan Paradkar

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
Master of Science
in
Civil Engineering

Dr. A.G. Hobeika, Chair

Dr. H. Rakha

Dr. H. Baik

June, 2001
Blacksburg, Virginia Tech

Key Words: TRANSIMS, Activity based modeling, CART

Alternative Methodology To Household Activity Matching In TRANSIMS

by

Rajan Paradkar

Dr. A.G. Hobeika

(Chairman)

(Abstract)

TRANSIMS (Transportation Analysis and Simulation System) developed at the Los Alamos National Laboratory, is an integrated system of travel forecasting models designed to give transportation planners accurate and complete information on traffic impacts, congestion, and pollution. TRANSIMS is a micro-simulation model which uses census data to generate a synthetic population and assigns activities using activity survey data to each person of every household of the synthetic population. The synthetic households generated from the census data are matched with the survey households based on their demographic characteristics. The activities of the survey household individuals are then assigned to the individuals of the matched synthetic households. The CART algorithm is used to match the households. With the use of CART algorithm a classification tree is built for the activity survey households based on some dependent and independent variables from the demographic data. The TRANSIMS model assumes activity times as dependent variables for building the classification tree.

The topic of this research is to compare the TRANSIMS approach of using times spent in executing the activities as dependent variables, compared to match the alternative of using travel times for trips between activities as dependent variables i.e. to use the travel time pattern instead of activity time pattern to match the persons in the survey households with the synthetic households. Thus assuming that if the travel time patterns are the same then we can match the survey households to the synthetic population i.e. people with similar demographic characteristics tend to have similar travel time patterns.

The algorithm of the Activity Generator module along with the original set of dependent variables, were first used to generate a base case scenario. Further tests were carried out using an alternative set of dependent variables in the algorithm. A sensitivity analysis was also carried out to test the affect of different sets of dependent variables in generating activities using the algorithm of the Activity Generator. The thesis also includes a detailed documentation of the results from all the tests.

ACKNOWLEDGMENTS

I would like to acknowledge Dr. A.G. Hobeika, chairman of my committee, for his guidance and extended effort devoted to this research and for his financial support, during my involvement in this research. Without his support and vision this research would not have been possible. I would also like to deeply thank Dr. Hesham Rakha and Dr. Hojong Baik for being my committee members and for providing me with their valuable inputs.

In addition, my colleagues and partners in research Srinivas Jillella and Sasikul Kangwalklai deserve special mention for their help and guidance in understanding the complex algorithms of TRANSIMS modules. I would also like to thank a number of my special friends and colleagues at the department, who have helped me during my stay at Virginia Tech.

This thesis is incomplete without the mention of my mother, Ms. Nimisha Paradker, who provided moral and intellectual support throughout this effort. Her dedication and hard work is the primary cause for all my achievements. She has always been the epitome of strength and courage and has set an example throughout my life. I would like to dedicate this thesis to her.

Table of Contents

List of Illustrations

List of Tables

1.0 INTRODUCTION

1.1 TRANSIMS

TRANSIMS (Transportation Analysis and Simulation System) is an integrated system of travel forecasting models designed to give transportation planners accurate and complete information on traffic impacts, congestion, and pollution (TRANSIMS documentation by Los Alamos National Laboratory). TRANSIMS is sponsored by the U.S. Department of Transportation, the Environmental Protection Agency, and the U.S. Department of Energy. Los Alamos National Laboratory is leading this major effort to develop new, integrated transportation and air quality forecasting procedures necessary to satisfy the Intermodal Surface Transportation Efficiency Act and the Clean Air Act and its amendments.

TRANSIMS models create a virtual metropolitan region with a complete representation of the region's individuals, their activities, and the transportation infrastructure. Trips are planned to satisfy the individuals' activity patterns. TRANSIMS then simulates the movement of individuals across the transportation network, including their use of vehicles such as cars or buses, on a second-by-second basis. This virtual world of travelers mimics the traveling and driving behavior of real people in the region. The interactions of individual vehicles produce realistic traffic dynamics from which analysts using TRANSIMS can estimate vehicle emissions and judge the overall performance of the transportation system.

Previous transportation planning involved, surveying people about elements of their trips such as origins, destinations, routes, timing and forms of transportation used, or modes. TRANSIMS starts with data about people's activities and the trips they take to carry out those activities, and then builds a model of household and activity demand. The model forecasts how changes in transportation policy or infrastructure might affect those activities and trips.

The goal of the TRANSIMS project has been to conduct major research and development of fundamentally new approaches to travel forecasting and to develop technologies that can be used by transportation planners in any urban environment.

There are mainly 6 modules in TRANSIMS, the Population Synthesizer, the Activity Generator, the Route Planner, the Microsimulator, the Emissions Estimator and the Selector module. They can be easily replaced or modified without redoing the entire TRANSIMS framework. Additionally, new modules may be added without much trouble

The Activity Generator module takes as major input the households in the synthetic population, local area surveys, non-residential travel data, TRANSIMS networks, and land use data. The Activity Generator produces a list of activities for each traveler in the system and for each freight-hauling truck. For travelers contained in the synthetic population, activity patterns and mode choice preferences are derived from surveys. This derivation depends on demographic information contained in the synthetic households.

Different versions of each module have been developed during the research process. The differences range from minor changes in factors or values to completely different techniques. TRANSIMS is designed to accommodate and encourage the use of different modules both during the research process and in later commercial versions. This design, or Framework, will facilitate the development and use of new modules and, ultimately, a stronger modeling package. An example of two completely different techniques for activity generation is discussed in chapter 3.0

1.2 Activity Generator in TRANSIMS

The Activity Generator module, the second module in the TRANSIMS framework, develops a list of activities for each member of a synthetic household over a 24-hour horizon.

The Activity Generator module requires principal inputs from the synthetic households (obtained from the Population Synthesizer), the survey households, the TRANSIMS

network data, and the land-use data. Using these inputs, the Activity Generator module produces a list of activities for each traveler in the system (Activity Generator course manual).

The synthetic households are generated by the Population Synthesizer module using census data to represent the real households in the metropolitan area under study. The demographic data of the synthetic households must match the demographic data of the survey households that are part of the Activity Generator's database. These demographic data are used to select a suitable survey household to match any given synthetic household. This matching is then used to produce the activity patterns for the members of this synthetic household.

The survey households are comprised of a set of households that have been surveyed to yield the input database for the Activity Generator module. The data from the survey household comprise of information related to household demographic data and household activity 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. The activity survey obtained from the survey household is a representative sample of the household activities, including travel and event-participation information for each household member. Skeletal activity patterns are created by stripping locations from the database.

The Activity Generator module uses household demographic survey data to build a classification tree by applying the CART (Classification And Regression Tree) algorithm. Each survey household is effectively placed by the CART algorithm into one of the tree's terminal nodes. The survey household demographic data pertaining to 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.

After the Activity Generator module has constructed a classification tree using the survey household demographic data, and each survey household has been placed into one of the

tree's terminal nodes, any given synthetic household must be matched with a survey household based on its demographic data. The format of the synthetic household demographic file is the same as the survey household demographic file (Population Synthesizer course manual).

The Activity Generator in TRANSIMS uses a model, which considers the time spent at activities as dependent variables and basis for comparison to match the synthetic households to the survey households. The assumption made in this model is that the activity travel patterns of households would be similar if the time spent at the activities is similar too.

1.3 The problem and an alternative method for household matching

TRANSIMS assumes that any two activities, separated by time and location, require travel between them. Thus a trip is generated when a successive activity has a different location then the previous one. So as the number of trips generated in the network is directly dependent upon the activities assigned to the members of the synthetic households, it becomes imperative that the activities should be assigned to the synthetic households as accurately as possible.

Now the Activity Generator assigns the activities of the survey households to the synthetic households on the basis of certain dependent and independent factors. So choosing these factors correctly is of great importance in order to predict the trips accurately.

The topic of this research is to compare the TRANSIMS approach of using times spent at the activities executed, as dependent variables to match the households with a proposed alternative of using travel times for trips between activities as dependent variables i.e. to use the travel time pattern instead of activity time pattern to match the persons in the survey households with the synthetic households. Thus assuming that if the travel time patterns are the same then we can match the survey households to the synthetic

population i.e. people with similar demographic characteristics tend to have similar travel time patterns.

1.4 Organization of Thesis

This chapter has discussed briefly the simulation program TRANSIMS and its Activity Generator module and the alternative approach for trip assignment suggested at Virginia Tech. The next chapter is the Literature Review, which discusses the Activity based modeling approach for travel demand forecasting. Also discussed in this chapter is the CART algorithm that is used in the Activity Generator module for tree modeling and pruning. And the program SPLUS, which is used in our study to implement the CART algorithm for growing the regression tree for the households. The third chapter discusses the Activity Generator module and the CART algorithm in detail. In the fourth chapter the development and implementation of an alternative model to assign trips is discussed in detail. Chapter five presents the results and analysis of the study conducted in this research. Chapter six includes the sensitivity analysis and chapter seven discusses the conclusions of the study and recommendations for future research.

2.0 LITERATURE REVIEW

2.1 Introduction:

The literature review involves the activity based modeling approach used in TRANSIMS. It also discusses the CART algorithm that is used by TRANSIMS to build a classification tree based on household demographic data. The use of SPLUS, statistical analysis software to get the classification and regression trees from the household demographic data using the CART algorithm is also discussed in this chapter.

2.2 Activity Based Modeling System for Travel Demand Forecasting:

Looking into the past few decades, one can notice that the emphasis of transportation planning has shifted from the construction of new infrastructure to the effective management of travel demand (RDC Inc.). Rising social, environmental, and economic concerns has brought about this shift, coupled with a realization that building one's way out of congestion is only a temporary solution to serving the increasingly complex patterns of travel demand that evolve over time.

In this regard, the decade of the 1980s saw an increased interest in the development and implementation of Travel Demand Management (TDM) strategies. These strategies were aimed at effectively managing and distributing travel demand, both in the spatial and temporal directions.

The activity-based approach conceived in around the 1970s, is a system that uses as inputs socio-demographic information of potential travelers and land use information to create schedules followed by people in their everyday life providing as output, for a given day, detailed lists of activities pursued, times spent in each activity, and travel information from activity to activity (including travel time, mode used, and so forth).

Activity-based approaches explicitly recognize that travel demand is derived from the need to pursue activities that are dispersed in time and space. Moreover, these approaches recognize the inter-dependence among decisions for a series of trips made by an

individual. They also recognize the interactions among various members of the household, that arise when household members allocate resources (such as household vehicles) to themselves, assign and share tasks, and jointly engage in activities. And hence one can see that the activity-based approach provides a theoretically and conceptually stronger framework within which travel demand modeling may be performed.

Motivation:

The motivation for activity based travel models is that travel is derived from the demand for activities. Stated simply, the motivation for activity based travel forecasting is that travel decisions are activity based. Individuals adjust their behavior in complex ways, motivated by a desire to achieve their activity objectives.

Why The Activity Based Approach?

The activity-based approach explicitly recognizes the fact that the demand for activities produces the demand for travel (RDC Inc.). In other words the need or desire to engage in an activity at a different location generates a trip. Then once we understand how activities are engaged in the course of a day or a week, a rigorous understanding of travel demand will follow.

The activity-based approach thus aims at the prediction of travel demand based on a thorough understanding of the decision process underlying travel behavior. In this sense the activity-based approach is entirely different from the approach taken for the development of the four-step procedure where statistical associations, rather than behavioral relationships, were the main focus. Moreover, as the activities engaged in a day are linked to each other, trips made to pursue them are also linked to each other. They cannot be analyzed separately one by one.

In recent times several important advances have taken place in transportation planning contexts such as:

- Accumulation of activity-based research results,

- Advances in survey methods (e.g., stated-preference (SP) and time-use survey methodologies) and statistical estimation methods, and
- Advances in computational capabilities and supporting software (database software, GIS, etc.).

All these changes have created an environment where a model of travel behavior can be developed while adhering to the principles of the activity-based approach.

Activity-based studies of travel behavior have led to the following emphases:

- Constraints which govern activity engagement and travel behavior (e.g., store opening hours, vehicle availability),
- Behavioral changes, or behavioral dynamics which are exhibited when an individual is faced with changes in the travel environment (e.g., switching between driving alone and carpooling to work),
- Adaptation as a special case of behavioral dynamics (e.g., a new baby prompting the acquisition of a large-screen TV set by the parents who gave up evening outings),
- The time dimension which is implicit in the emphasis of behavioral changes as changes taking place over time,
- Day-to-day variability in behavior and demand, as another special aspect of behavioral dynamics (e.g., part-time carpooling),
- Scheduling of activities and trips over a span of time; when to engage in what type of activities, and in what sequence,
- Trip chaining: combining stops into a trip chain,
- In-home/out-of-home activity substitution (e.g., going out for a movie vs. watching TV at home), which is directly related to trip generation,
- Inter-personal linkages, which may take on the form of task and resource assignment (e.g., vehicle allocation within a household) and resource sharing (e.g., carpooling by family members), joint activity engagement (a Sunday family

outing), and activity generation (e.g., a child's ballet lesson generating the parent's activity of chauffeuring the child to ballet school), and

- Household life-cycle stage, which is strongly associated with the level of inter-personal interaction.

Structure of the Activity-Based Model System:

An ideal activity based model system includes full information on the chain of activities each person in the household is involved in throughout the day. This information includes time of day, duration, activity type, location, mode of travel, and travel time for each activity.

Demand for travel is derived from the demand to engage in various activities:

Current transportation forecasting models use the "trip" (usually a vehicular trip) as the basic unit of analysis and prediction (Spear B. D. et. al.). However, most travel behavior research concludes that the decision to make a trip is really derived from a complex series of decisions including: what activities to engage in; whether the activity can be accomplished without vehicular travel; and whether the activity can be combined with other activities. Consequently, by starting with the trip, current transportation forecasting models are unable to explicitly address issues related to: the substitution of non-travel options (e.g., tele-commuting and tele-shopping), the substitution of non-vehicular trips (e.g., walk trips); trip chaining behavior; and induced demand for travel.

The current "trip-based" model framework should be replaced by an "activity- based" framework in which the demand for travel is derived from a more basic demand to engage in various activities. Activities represent various "goal-satisfying behaviors engaged in by household members". Activities can be grouped into various categories (e.g., work, shop, recreation, mandatory, flexible, optional), and can be described in terms of where, when, and how long the activity takes place. However, an activity does not necessarily result in a trip. By modeling activities rather than trips, it should be possible to forecast a number, which represents a realistic upper bound on the total number of trips that could be generated under a particular scenario.

Model System Requirements:

The following is the list of requirements, which an activity-based travel forecasting model system should ideally satisfy. First, it should be theoretically sound, both behaviorally and mathematically. Without these issues we can not rely on the results. Second, the scope must be complete enough to make the model useful. If important dimensions of the activity scheduling decision are missing, the model prediction will be incomplete and of limited use. Enough resolution of the daily schedule alternatives is required to capture behavior which affects the aggregate phenomena in which we're interested. Third, the resource requirements of the model must allow it to be implemented. In addition to the data we need for estimating the model parameters, we need to validate the model using a different set of data. To use the model for prediction we must also be able to generate reliable forecasts of the exogenous variables used by the model. The model must also be simple enough so that the logic and computation required make it technically and financially feasible to develop, maintain and operate. Finally, the model must produce valid results.

2.3 CART Algorithm:

The word CART means Classification And Regression Tree. Tree-based modeling (Nagpaul P. S.) is an exploratory data analytic technique for uncovering structure in large data sets. This technique is quite useful for:

- Constructing and evaluating multivariate predictive models
- Screening variables
- Summarizing large multivariate datasets
- Assessing the adequacy of linear models

Tree-based models are useful for both classification and regression problems. In these problems, there is a set of classification or predictor variables (Xi) and a dependent variable (Y). In classification trees the dependent variable is categorical, whereas in regression trees the dependent variable is quantitative.

As with all regression techniques we assume the existence of a single response variable and one or more predictor variables (Bell J. F.). If the response variable is categorical then classification trees are created (equivalent to discriminant analysis or logistic regression) and if the response variable is continuous then regression trees (equivalent to multiple regression) are produced. The predictor variables can be a mixture of continuous and categorical variables. The final output is a decision tree where we decide which branch to follow after applying some test to one or more variables.

In certain circumstances they have an advantage over more common methods such as discriminant analysis. In particular, they do not have to conform to the same distributional restrictions and there is no assumption of a linear model. CART is particularly useful when you consider that your predictors may be associated in some non-linear fashion.

Decision points are called nodes, and at each node the data are partitioned. Each of these partitions is then partitioned independently of all other partitions. This could carry on until each partition consisted of only one case. This would be a tree with a lot of branches and as many terminal segments (leaves) as there are cases. Normally some 'stopping rule' is applied before we arrive at this extreme condition. Inevitably this may mean that we have some 'impure' partitions, but it is necessary to balance accuracy against generality. A tree which produced a perfect classification of training data would probably perform poorly with new data.

The theory of regression trees was developed by Breiman et al. (1984). CART methodology consists of three parts. First, we grow a regression tree which overfits the data. Secondly we prune from the overfitting tree a sequence of subtrees and lastly we try to select from the sequence of subtrees a subtree which estimates the true regression function as best as possible.

CART is powerful (Taylor P. et. al.) because it can deal with incomplete data, multiple types of features (floats, enumerated sets) both in input features and predicted features, and the trees it produces often contain rules which are humanly readable.

Decision trees contain a binary question (yes/no answer) about some feature at each node in the tree. The leaves of the tree contain the best prediction based on the training data. Decision lists are a reduced form of this where one answer to each question leads directly to a leaf node. A tree's leaf node may be a single member of some class, a probability density function (over some discrete class), a predicted mean value for a continuous feature or a Gaussian (mean and standard deviation for a continuous value).

Theoretically the predicted value may be anything for which a function can be defined that can give a measure of impurity for a set of samples, and a distance measure between impurities.

The basic algorithm is given a set of samples (a feature vector) to find the question about some feature which splits the data minimizing the mean "impurity" of the two partitions. Recursively apply this splitting on each partition until some stop criteria is reached (e.g. a minimum number of samples in the partition.

2.4 SPLUS:

SPLUS is a statistical analysis software developed by Insightful Corporations © which makes it possible to create CART graphs using the SPLUS program. Once the demographic data is imported into the software, the Tree Model function in the software allows the user to build a classification and regression tree and even prune the tree using the desired variables and split criteria. Thus a tree similar to that created in TRANSIMS can be reproduced using the SPLUS software.

2.5 Conclusions:

The basic task is to propose an alternative methodology for the activity based model used in TRANSIMS. The methodology developed here attempts to use the travel time pattern instead of activity time pattern to match the persons in the survey households with the synthetic households.

3.0 ACTIVITY GENERATOR IN TRANSIMS

3.1 Introduction

As discussed in chapter one, the role of the Activity Generator in TRANSIMS is to develop a list of activities for each member of a synthetic household over a 24-hour horizon. The *Fig. 3.1* (Hobeika et. al.) details the inputs and outputs of this module, along with its interactions with the other relevant modules of TRANSIMS.

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 discusses in detail the first three of the above major algorithmic steps and thus provides a better understanding of the Activity Generator module and it's functioning. The last two steps are not discussed here, as they do not serve any purpose in our study. We are concerned with only the steps where we match the households and assign the activities to the synthetic households. To understand more about the complete algorithm of the Activity Generator please refer to the references.

3.2 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. Activities are grouped sequentially for each member in the household as collected from the household activity survey form (Activity Generator course manual by Hobeika et. al.).

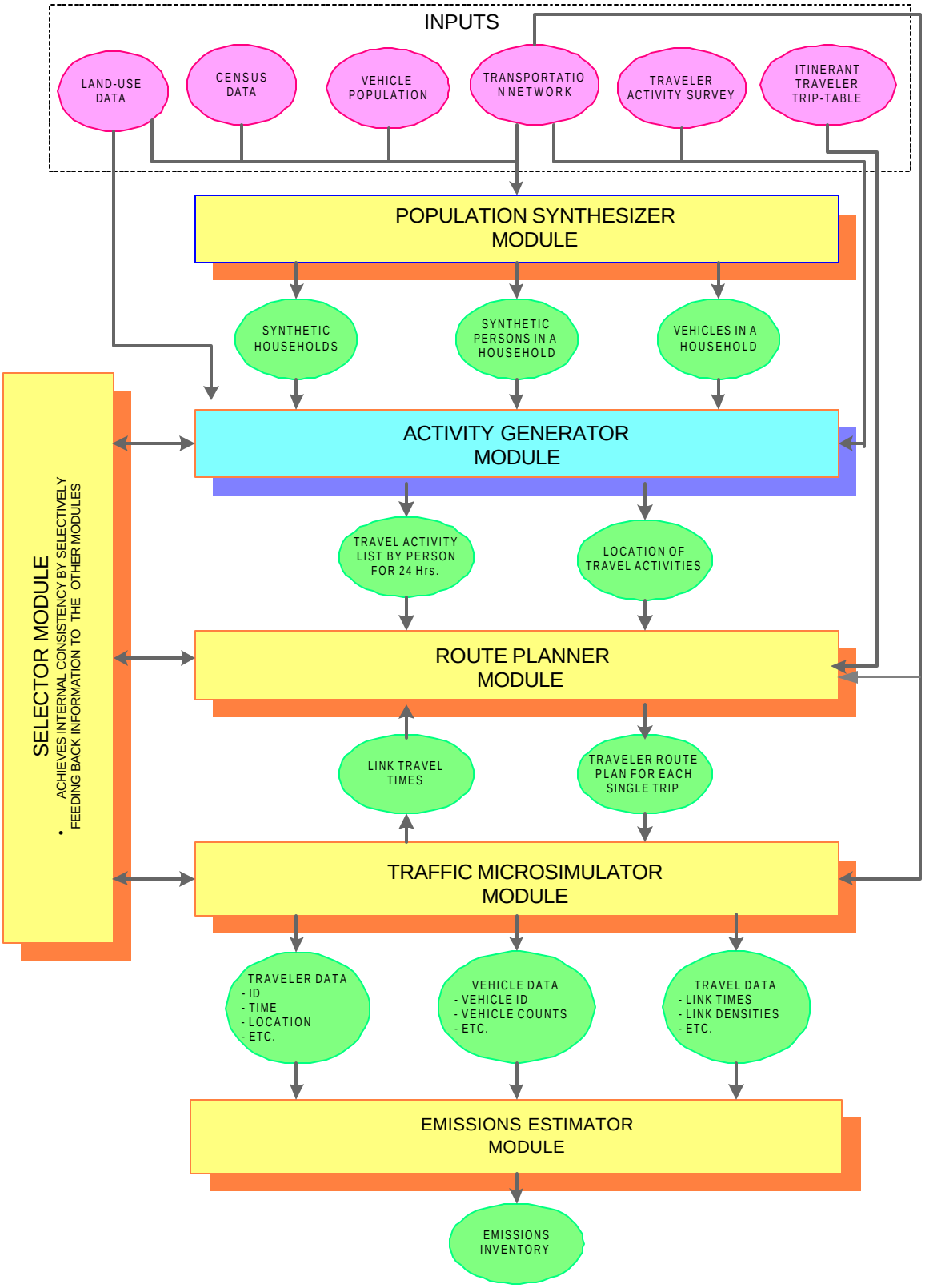


Figure 3-1: The Activity Generator module within the TRANSIMS framework.

The format of the skeletal activity pattern is shown in the *Table 3.1*, and an example is given in *Fig. 3.2*.

Table 3-1: Survey Activity file format

Field	Description	Allowed Values
Survey Household ID	Each survey household has a unique ID.	Integer
Person Number	Each person in the household has a unique ID, starting with 1. Numbers are only unique within the household.	integer: 1 through household size
Activity Number	The activity number for each person.	integer: 0 through n: 0 = initial at-home activity of the day if necessary
Activity Type	Definitions may vary.	integer: 0 through n: Example: 0 = at-home activity 1 = work 2 = shop 3 = school 4 = visit 5 = other 6 = serve passenger
At Home	Coded 1 for activity at-home, 2 for out of home.	integer: 1 or 2
Were-you-there	Coded 1 if person was already at the location, 2 if not.	integer: 1 or 2
Mode for arriving at activity	Integer code for mode. Modes must correspond to modes in TRANSIMS mode map file used by the Route Planner and the Traffic Microsimulator.	integer: 1 through n: 1 = Walk 2 = Car 3 = Transit 4 = Light Rail 5 = Park and Ride outgoing 6 = Park and Ride incoming 7 = Bicycle 8 = Magic Move 9 = School Bus
Driver	Coded 1 if person was the driver, 2 if passenger. Otherwise, 0.	integer: 1, 2, or 0
Number in Vehicle	The number of people in vehicle	integer:1 through n
Activity Start Time	The time of start of activity in minutes after midnight.	integer: 0 through 2400
Activity End Time	The time of end of activity in minutes after midnight.	integer: 0 through 2400

Field	Description	Allowed Values
Geocode x	The Easting geocoordinate. Must be in units agreeing with mode coefficients.	decimal
Geocode y	The Northing geocoordinate. Must be in units agreeing with mode coefficients.	Decimal

The example given in *Figure 3.2* specifies the skeletal activity pattern for a household numbered 200087 that contains two persons. The **first** person has nine activities including an initial at-**home** activity, **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 (passenger)	0	180	460	7637347.0000	687428.0625
200087	2	1	3 (school)	2	2	2	2	2	480	1023	7648077.5000	713295.7500
200087	2	2	0 (home)	1	2	1	0	0	1050	1440	7637347.0000	687428.0625

Figure 3-2: An example of the skeletal activity pattern for the survey household numbered 200090.

3.3 Using the CART (Classification and Regression Tree) algorithm to build a classification tree based on household demographic data

The main purpose of the CART algorithm, used in the Activity Generator Module, is to produce an accurate classification of household characteristics based on households travel behaviors (Hobeika et. al.). Indeed, 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 characteristics. This tree is chosen to be sensitive to the principal characteristics of household behavior, 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.

This algorithm is comprised of two steps: the tree growing step and the cross-validation step for reducing or pruning the tree (Hobeika et. al.).

3.3.1 Tree-growing step

The CART algorithm is a technique for modeling a regression relationship between one or more dependent variables Y and independent variables $X_1, X_2, \ldots, X_K$. In the Activity Generator, a classification tree is constructed using the total times household spend in some 7 broadly classified activity types, obtained from the skeletal activity patterns, as the dependent variables Y_1, Y_2, …, Y_7. For example, Y_1 = total time spent by the household in working, etc. 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

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

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

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

$\boldsymbol{Y_j, j}$ **= 1, …, *p*:** given set of p dependent variables used to measure the performance characteristics of the household.

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

T **=** classification tree, comprised of nodes and branches.

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 $j = 1, .., p$:

$$D_j(N) = \sum_{i \in N} (Y_{ij} - \bar{Y}_{Nj})^2, \quad \text{where} \quad \bar{Y}_{Nj} = \frac{\sum_{i \in N} Y_{ij}}{|N|}.$$

$D(N)$ = total deviance of node N:

$$D(N) = \sum_{j=1}^{p} s_j D_j(N)$$

where s_j = 1/variance (Y_j)

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

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

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

N_{1kt} and 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 $X_{ik} > t \ \forall \ i \in N_{2kt}$.

$D_{kt}(N)$: 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 $i = 1, \ldots, n$, 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 i^{th} survey observation) for $k = 1, \ldots, K$, obtained from the survey household demographic data in Section 4.2, and Y_{ij} (value of measure Y_j for the i^{th} survey observation) for $j = 1, \ldots, 7$, which represent the set of dependent variables as described above. The information on the Y_j-variables are obtained from the skeletal activity patterns as shown in Section 3.1.
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 $D(N_i) > \boldsymbol{b}$, then also add N_i to $\tilde{T}_A$, for each i = 1, 2. Repeat the Main Step.

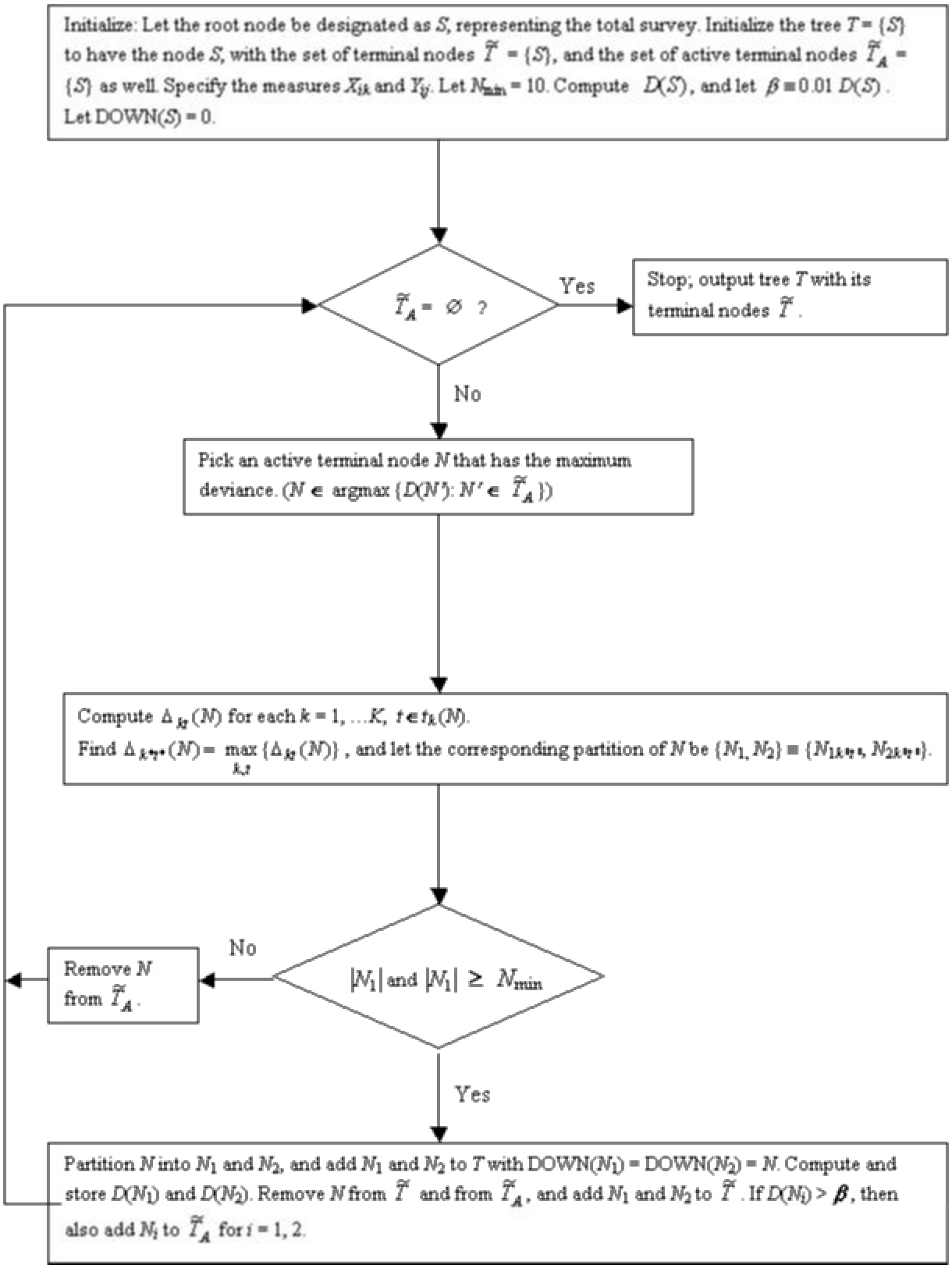


Figure 3-3: Flow-Chart fro the Tree-Growing Procedure.

3.3.2 Cross-validation step for reducing or pruning the tree

The binary tree produced by the tree-growing algorithm generally over fits the data that is 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 (called **S-PLUS**) of the former scheme, which is the one recommended by Breiman et al. (1984). This approach is described below.

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

v **= 1, …, 10**: index for the partitioned data sets.

$S^{(v)}$ = subset of the observations in *S* containing roughly 90% of the overall sample size pertaining to partition n .

$\boldsymbol{T^{(v)}}$ = a tree, which is built from the observation in $S^{(v)}$ alone.

$\boldsymbol{t}$ = any arbitrary tree selected from the set of trees $\{T^{(v)}, v = 1, \ldots, 10\}$

$\tilde{\boldsymbol{t}}$ = the set of terminal nodes of $\boldsymbol{t}$.

$\tilde{\boldsymbol{t}}(\boldsymbol{i})$ **for any observation** $\boldsymbol{i}$ = the **terminal node** from $\tilde{t}$ that this observation i would be classified into based on its demographics (this is called the *classification node* of *i*).

$\boldsymbol{e_i(t)}$ = a *prediction error* for observation i with respect to the tree $\boldsymbol{t}$:

$e_i(\boldsymbol{t}) = \sum_{j=1}^{p} s_j (Y_{ij} - \overline{Y}_{\tilde{t}(i),\, j})^2$, where $\overline{Y}_{\tilde{t}(i),\, j} = \dfrac{\sum_{i \in \tilde{t}(i)} Y_{ij}}{|\tilde{\boldsymbol{t}}(i)|}$ is the average of the values of the measure j for the observations that currently reside in the terminal node $\tilde{\boldsymbol{t}}(i)$ of the tree $\boldsymbol{t}$.

Likewise, for any subset $S_v \subseteq S$, define $e_{S_v}(\boldsymbol{t}) = \sum_{i \in S_v} e_i(\boldsymbol{t})$,

and for the entire set of observations in S, let $e(\boldsymbol{t})$ be the *prediction error* for the tree $\boldsymbol{t}$ as given by $e(\boldsymbol{t}) \equiv e_S(\boldsymbol{t}) = \sum_{i=1}^{n} \sum_{j=1}^{p} s_j (Y_{ij} - \overline{Y}_{\tilde{t}(i),\, j})^2$.

INITIALIZATION STEP

1. From the given survey sample S comprised of the n household data observations, randomly partition S into ten roughly equal subsets $S_1, \ldots, S_{10}$.
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, using the tree growing step.

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

3. Then, the *cross-validation method* **S-PLUS** 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.

3.3.3 Matching the given synthetic household with a survey household

This procedure involves three steps (Hobeika et. al.):

3.3.3.1 Assign each synthetic household (obtained from the Population Synthesizer Module) to a unique terminal node of the tree built from the survey households.
3.3.3.2 Select a survey household within the terminal node to match with the given synthetic household that is assigned to this node.
3.3.3.3 Match the assigned synthetic household members with the selected survey household members based on the following four demographic variables: relate, work, gender, and age.

3.3.3.1 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.

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

What this means is that for each survey household classified to an end node, a cumulative distribution probability is assigned by using the associated cumulative distribution. So if there are 20 households in one end node then the first household will be assigned a probability of 1/20 = 0.05 and the next would be assigned 0.05 + 1/20 = 0.1 and so on.

Once all the survey households are assigned these probability values, the synthetic households in that end node are then matched to them one by one using random numbers. A uniformly distributed random number between 0 and 1 is generated for every synthetic household. Then the first survey household with a probability greater than the random number is assigned to that synthetic household. Thus say a random number .067 is generated for a synthetic household, then the second household having probability equal to 0.1 will be assigned to it. And so the activity pattern of this survey household is assigned to the synthetic household.

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

For our study we calculate directly the total number of trips after this point. We will assume that the number of trips will remain almost the same even if we don't assign individual members of Household.

The Activity Generator algorithm also generates activity times and durations for the activities assigned and also locations for the activities. These steps are described in detail in the Activity Generator module developed at Virginia Tech by Hobeika et. al.

4.0 DEVELOPMENT AND IMPLEMENTATION OF MODEL

4.1 Introduction:

This chapter deals with the model development process for the Activity Generator module of TRANSIMS. In this model an attempt has been made to replicate the original algorithm of the Activity Generator model, so as to give flexibility in using the dependent variables of our choice.

The input data used by the Activity Generator is first converted to a format that could be used in excel worksheets. With the use of Visual Basic macros, this data was then modified to suit the algorithm. The survey activity data contains only activity times. Hence a code was written to calculate the travel times between these activities and the total number of trips for each household. These steps are discussed in detail in the later sections.

To get the classification trees, the program SPLUS was used which follows the CART algorithm like TRANSIMS. These trees were then pruned in accordance with the method used in TRANSIMS.

A code was written using Visual Basic to distribute each household at the end nodes of the classification tree. Each synthetic household as well as survey household was assigned an end node. After the households are assigned to a specific node, a synthetic household is chosen at random within this terminal node to obtain a matching survey household.

Each survey household is first assigned a weight $w_i = 1$. This means that each survey household is equally important. Hence each survey household in the assigned terminal node for the synthetic household is chosen with probability of 1/N (where N = the total number of survey households in that end node). We randomly select a survey household according to this probability by using the associated cumulative distribution. This is done by generating a random number between 0 and 1 for each synthetic household, and then

selecting that survey household for which the random number falls within its defined interval.

Thus each synthetic household has now been assigned activity data. From this we can get the total number of trips for each household. We can analyze this data to compare the total number of trips generated using the TRANSIMS method of using activity times, with our proposed method of using travel times.

4.2 Input Data:

As discussed earlier the Activity Generator uses the following files as input files to generate the activities:

1. Household activity survey data
2. Survey household demographic data
3. Synthetic households
4. Land-Use data
5. TRANSIMS network data

Our study includes the algorithm only up till the synthetic households are assigned the activities of the survey households. This step requires only the first three data. The last two files are used to locate the activities on the TRANSIMS network. Hence only the first three data are discussed in detail here.

In our study an attempt has been made to select data, which can to the nearest, represent a real-life example.

1. Household activity survey data:

The household activity survey data that is used in our study is taken from the Portland, Oregon Activity and Travel Survey of 1994/95 (Activity Generator course manual). 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 coded.

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) that are selected in TRANSIMS are included in the survey file to represent the activity travel pattern of each household. These are 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).

Our study proposes another set of five dependent variables (Y_j, j = 1, …, 5) that we calculate from the original file to represent the activity travel pattern of each household.. They are given below:

Y_1 = total travel times for household in minutes for trips starting from Home and ending at Work (THomeBasedWork),

Y_2 = total travel times for household in minutes for trips starting from Work and ending at Home (TWorkBasedHome),

Y_3 = total travel times for household in minutes for trips starting from Home to destinations other than Work and back to Home (THomeBasedOther),

Y_4 = total travel times for household in minutes for trips starting from Work to destinations other than Home and back to Work (TWorkBasedOther),

Y_5 = total number of trips per household (TTrips),

To model the activity travel pattern of an average or typical weekday 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 weekday. This limited the sample to 3,505 households. The final sample, excluding those with missing data, includes **3,470** households (i = 1, …, 3470).

2. Survey household demographic data

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 (DENSITY).

This same set of variables are used in our study except the last variable which is the residential density of block group in housing units/acre, as this data was unavailable in the survey activity data file used.

The dependent variables along with the independent variables are basically used as the basis for classification of the households into a binary tree using the CART algorithm. In

our study we try to compare the output from using the original independent variables (activity times Ys) with the output from using the new independent variables (travel times Ys).

3. Synthetic households

The synthetic households data consists of the output of the Population Synthesizer module in TRANSIMS. The Population Synthesizer uses the census data from that region and forecasts a synthetic population that represents the actual population in that region.

The population thus generated will be assigned activities from the survey data. Hence it is very important that the survey population should be able to represent the synthetic population accurately.

The Bignet Network, which is an example network in the current version of TRANSIMS, was run to generate the required synthetic population for our study. This sample network has been designed to demonstrate the computational resources required for a complete run of TRANSIMS using a small community. It is also designed to demonstrate the principal components of TRANSIMS: Population Synthesizer, Activity Generator, Route Planner, Traffic Microsimulator and Emissions Estimator. The Bignet Network is approximately one-tenth the size of Portland, Oregon.

In order to get the synthetic population data, it was necessary to run the Bignet scenario in TRANSIMS. The scripts of the TRANSIMS modules were run on the Bignet scenario to get the required output.

The output of the Population Synthesizer consists of 3 files containing synthetic population data for family households, Non-Family households and Group quarters. For our study we take only the family households data. A total of **112,782** households were generated in the family household category and this data was used in our study.

4.3 Approach:

The output from the Population Synthesizer is not in the same format as the survey data. Even the variables required for matching the households need to be calculated from the output data. Hence a code was written to extract the required demographic data from this file and arrange it in another file. The final file will contain the following values for each household:

X_1 = number of persons in household (HHSIZE),

X_2 = number of vehicles in household (VEHICLES),

X_3 = total income in $ (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), and

X_8 = number of employed household members (WORKERS)

The survey activity file used is the same as used in the sample network of Bignet. The required data for our analysis had to be extracted from this file too.

The variables (Ys) discussed previously were calculated from this survey data file. The survey file contains the information for types of activities, their starting times and their ending times. To get the travel times, we calculated the difference between the ending time of one activity and the starting time of the immediate next activity. For example, say person 1 in household 200070 has the following activity pattern: The first activity was a Home activity starting at 0 mins and ending at 708 mins, and the next activity was a Work activity starting at 720 mins and ending at 870 mins. TRANSIMS uses the variable THOME, which is the time spent for in-home activity, which in this case would be equal to 708 – 0 = 708 minutes. Similarly TWORK = 870 – 720 = 150 minutes. For our study we calculate the TWork (THomeBasedWork) = 720 – 708 = 12 minutes. Thus person 1 of household 200070 spends 12 minutes in travel time from Home to Work. The final file will contain the values for Xs and Ys for each household as mentioned earlier in the section.

This file is then used as an input to the SPLUS program. Where the Xs are used as the independent variables and the Ys as the dependent variables. After running the program on our survey activity data, we noticed that the first split for the classification tree was not that of "Number of Workers". But in accordance with the LANL theory we modified our approach to suit their method of splitting the households first with "Number of Workers in the Household".

The tree used by the LANL is described below in 4 parts:

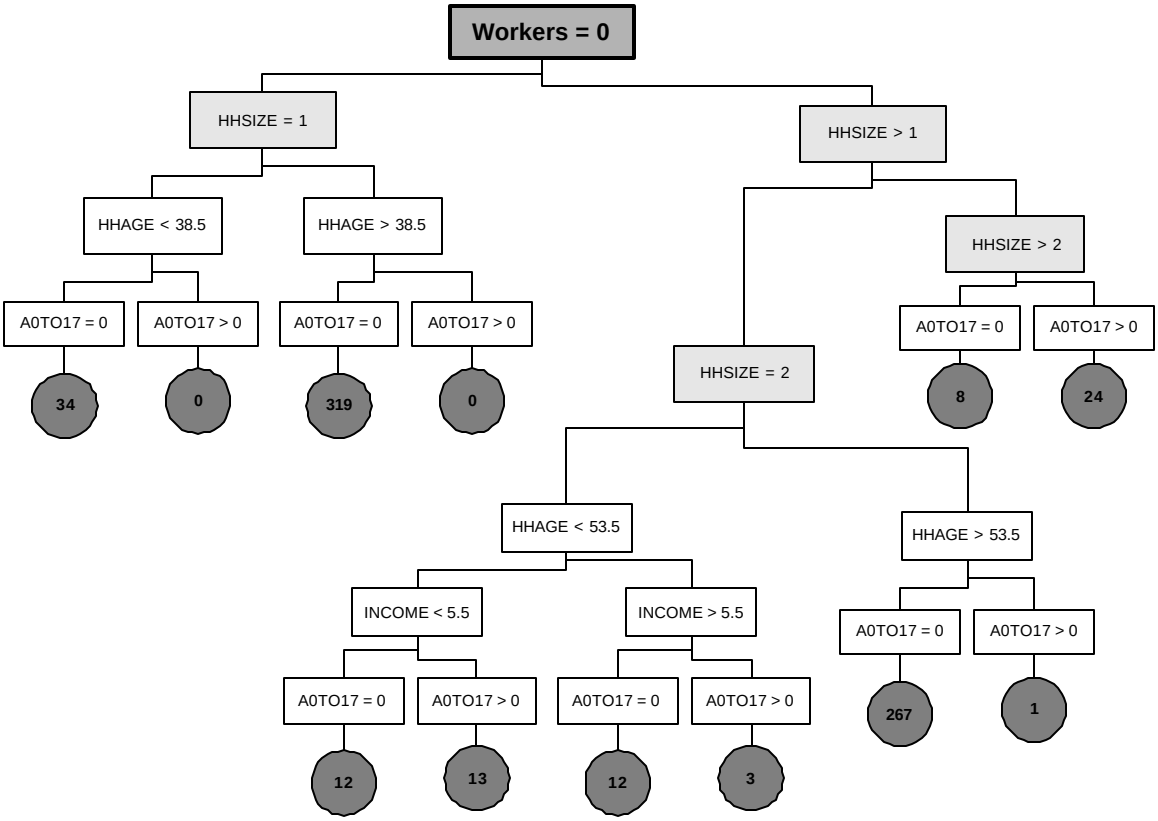


Figure 4-1 Number of Workers equal to zero

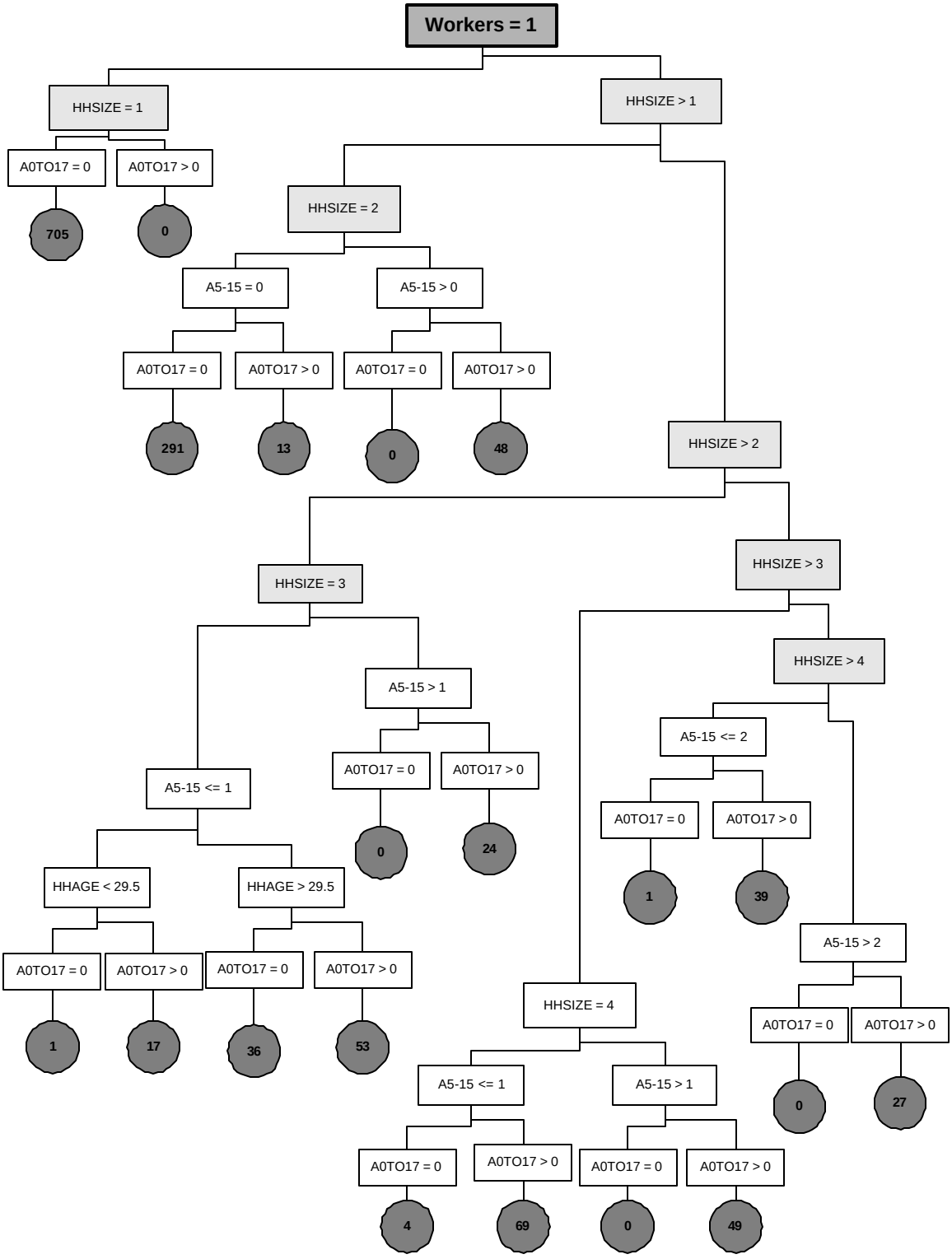


Figure 4-2: Number of Workers equal to 1

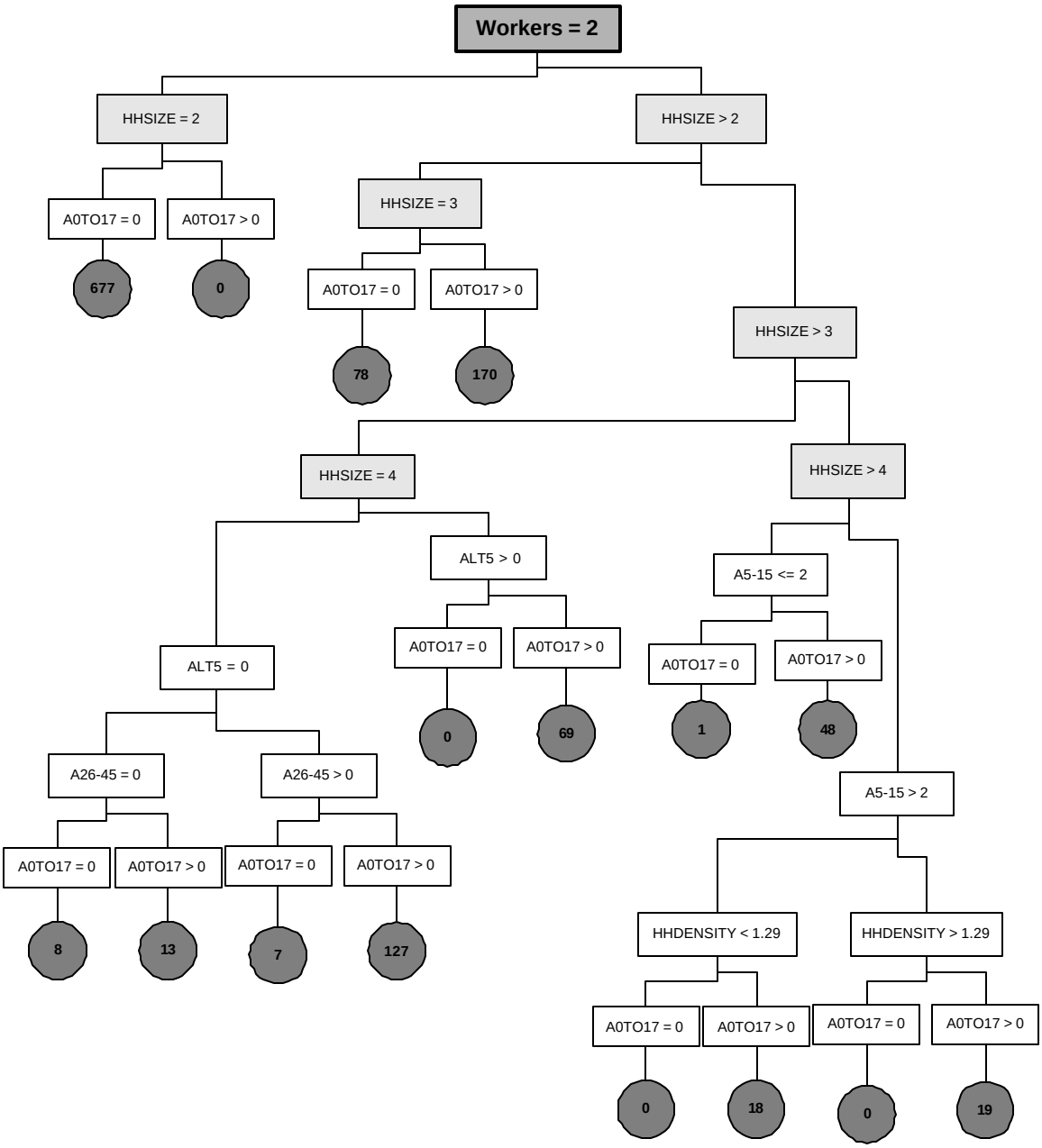


Figure 4-3: Number of Workers equal to 2

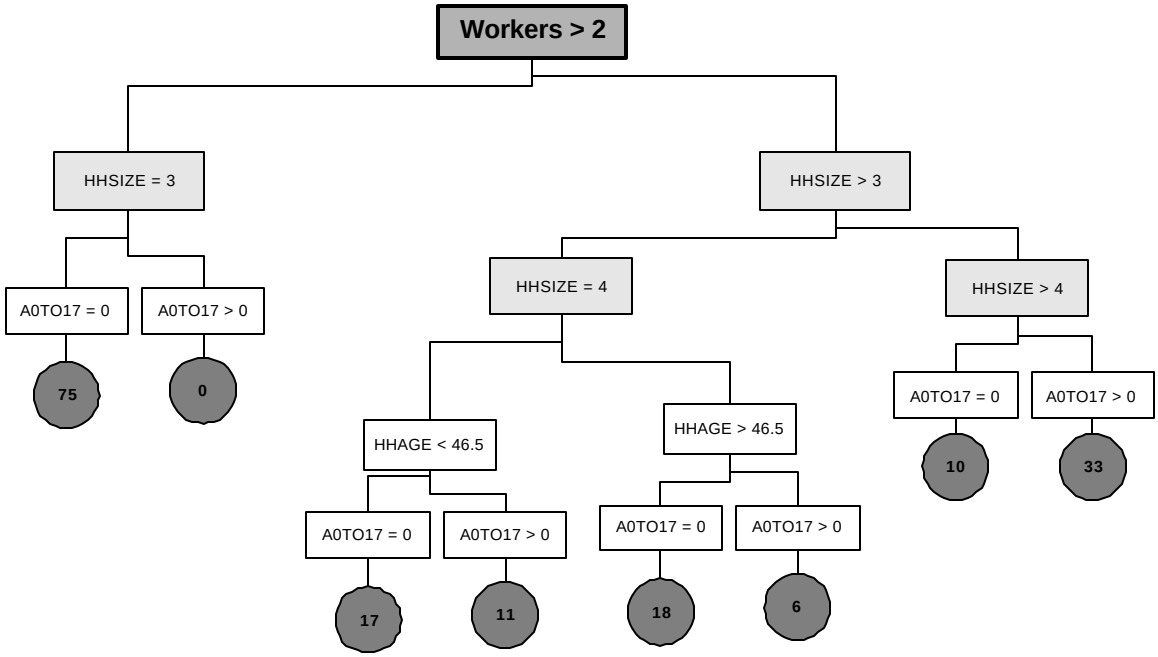


Figure 4-4: Number of Workers greater than 2

To achieve this, the activity data as well as the synthetic population data was first split into 4 separate files:

1) Households with workers = 0
2) Households with workers = 1
3) Households with workers = 2
4) Households with workers > 2

These files were then separately run in SPLUS to get the classification trees out of them. For each file, two separate runs were made. First using the original Ys (Activity times) and the second using the new Ys (Travel times). The trees were then pruned using the algorithm described in Section 3.3.2. The final trees that we arrived at are shown below:

4.3.1 Tree for Workers = 0, and original Ys:

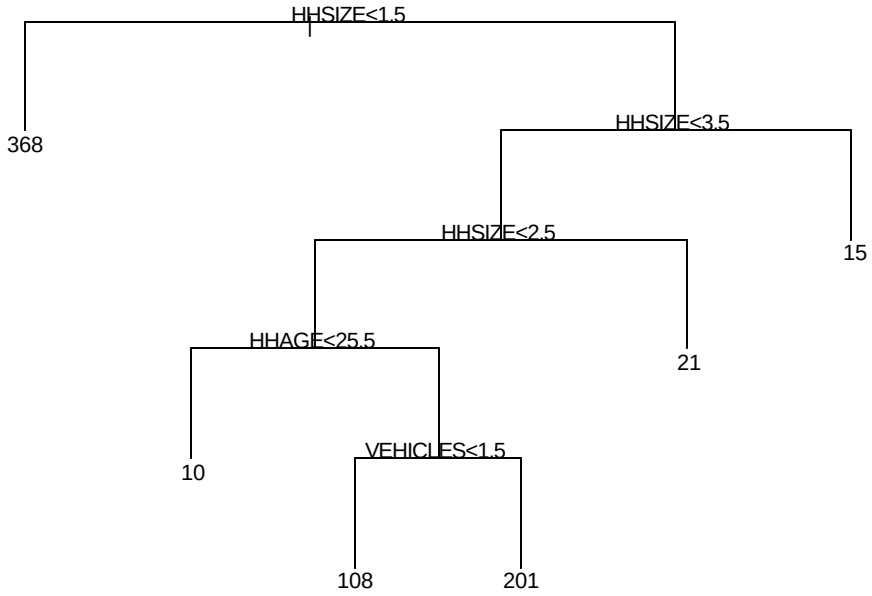


Figure 4-5: Tree for Workers = 0, and original Ys

4.3.2 Tree for Workers = 1, and original Ys:

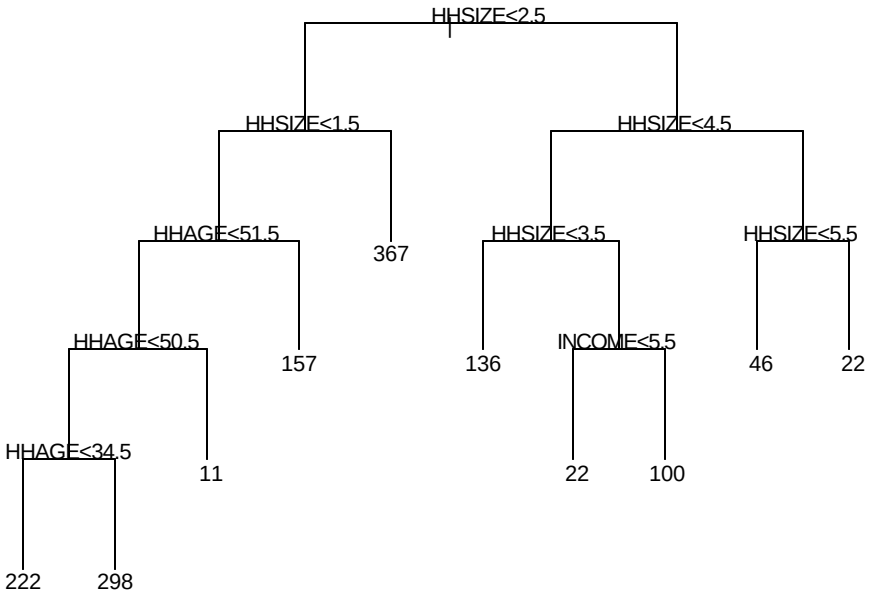


Figure 4-6: Tree for Workers = 1, and original Ys

4.3.3 Tree for Workers = 2, and original Ys:

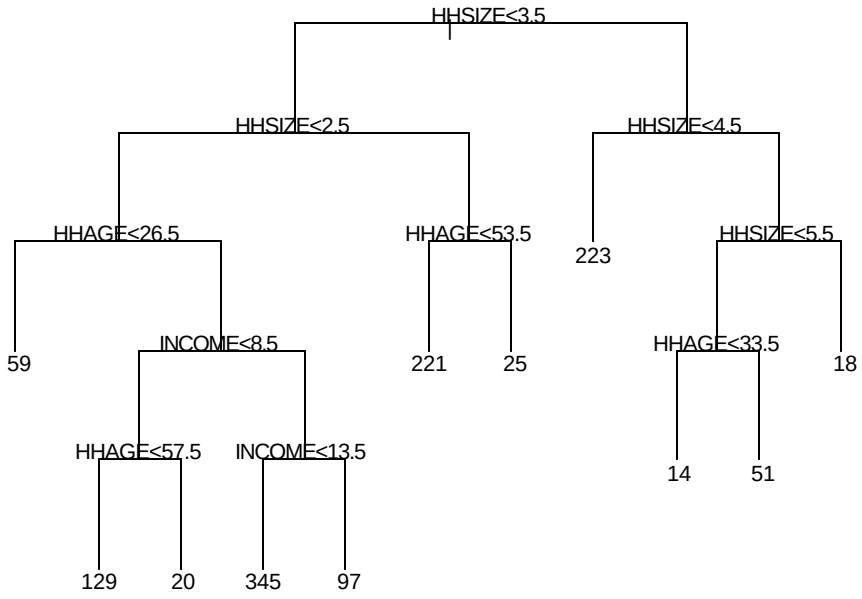


Figure 4-7: Tree for Workers = 2, and original Ys

4.3.4 Tree for Workers > 2 and original Ys:

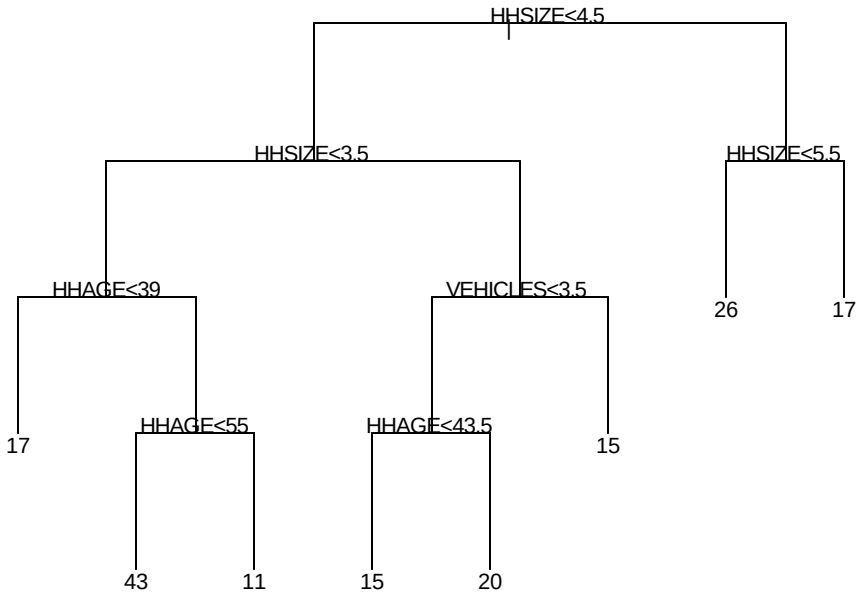


Figure 4-8: Tree for Workers > 2, and original Ys

4.3.5 Tree for Workers = 0, and new Ys:

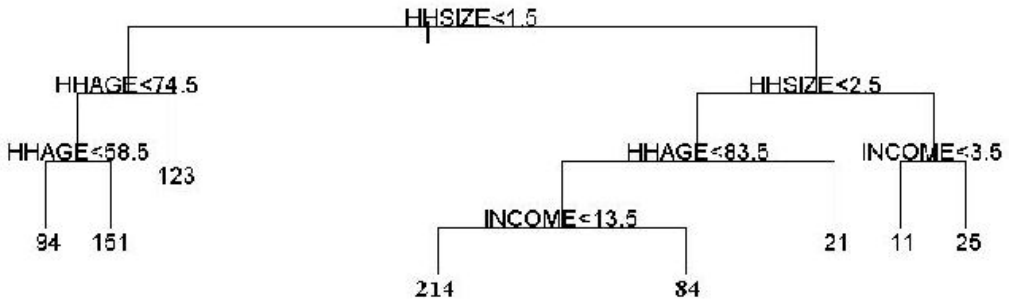


Figure 4-9: Tree for Workers = 0, and new Ys

4.3.6 Tree for Workers = 1, and new Ys:

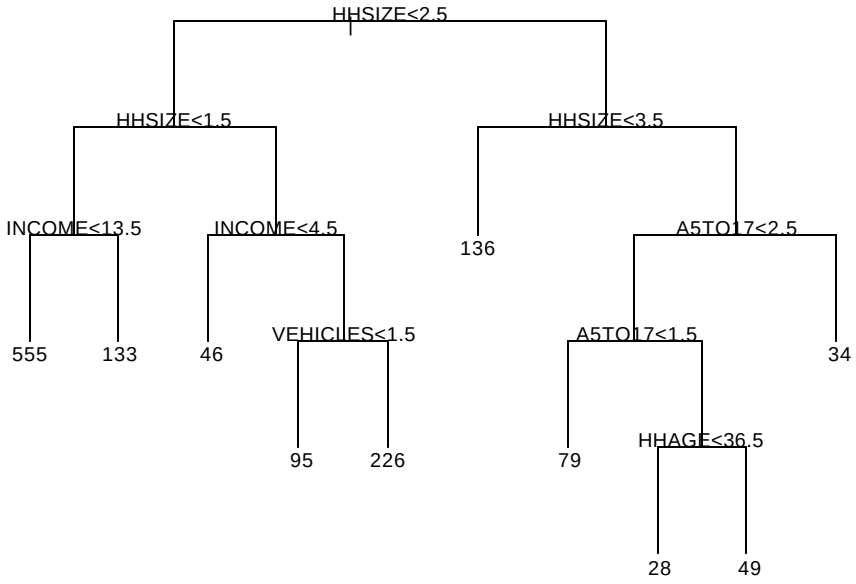


Figure 4-10: Tree for Workers = 1, and new Ys

4.3.7 Tree for Workers = 2, and new Ys:

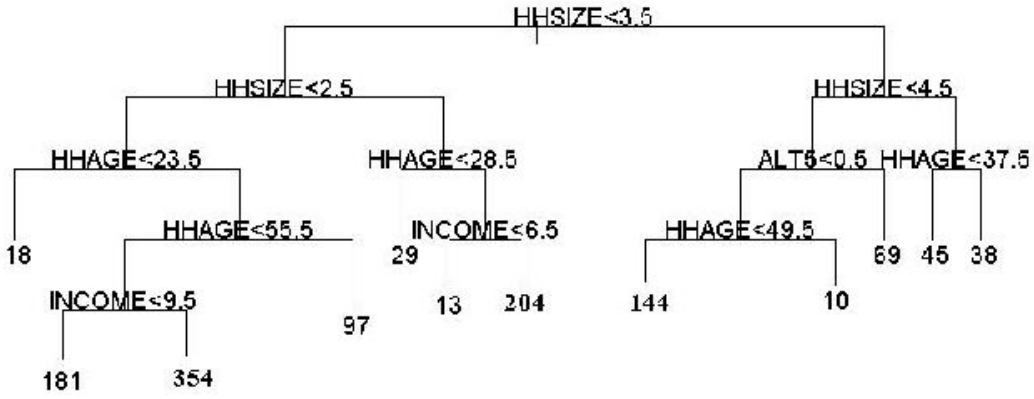


Figure 4-11: Tree for Workers = 2, and new Ys

4.3.8 Tree for Workers > 2, and new Ys:

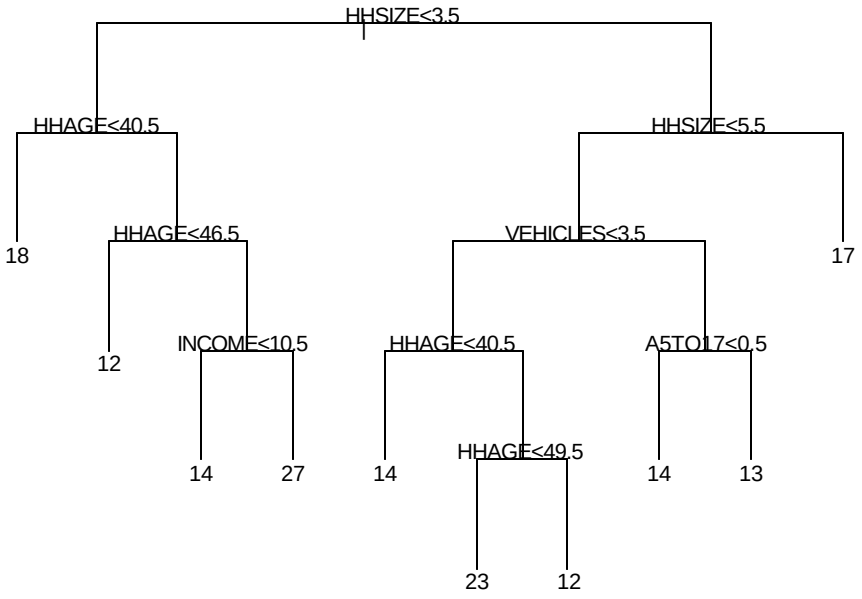


Figure 4-12: Tree for Workers > 2, and new Ys

These are the final trees that we arrived at after pruning. Note that this is just one single binary tree, which has been represented in 4 parts for the sake of convenience.

The next step is to assign each survey household to an end node of the respective tree. A code was written to classify the households in accordance to the tree. Each end node was given a unique 3-character identity. The first digit represents the Ys that were used i.e. 1 is used for the original Ys and 2 for the travel time Ys. The second digit represents the number of workers in the household i.e. 0 stands for 0 workers, 1 for 1 worker, 2 for 2 workers and 3 for workers greater than 2 in a household. The last character/s represents the number of the end node in that particular tree. Like in the tree in section 4.3.2 the end node with total households = 46 is given the identity of 119.

Following this method we carry on the distribution of the survey households into each of the end nodes. Each survey household is also assigned a weight w_i = 1. This means that

each survey household is equally important. Hence, each survey household in the assigned terminal node (with N households) for the synthetic household is chosen with the probability 1/N. Hence a cumulative distribution is assigned to each node with unit equal to 1/N, where N is the total number of survey households in the end node.

Once this is done, the synthetic households are also assigned to these end nodes in the same manner. Each synthetic household is also assigned a uniformly distributed random number between 0 and 1.

Our final step would be to match these households by selecting that survey household for which the random number of the synthetic household falls within the defined interval of the cumulative distribution. For example, say the end node 119 contains 46 survey households. So the first household is assigned a number 1/46 = 0.0217. Similarly the second household would have 0.0217 + 1/46 = 0.0434 and so on. Thus the last household would always be assigned 1. Referring to our results in chapter 5, the total number of synthetic households assigned to 119-end node is 407. And each synthetic household will be assigned a random number between 0 and 1. This number is then compared with the survey household cumulative distribution. The first survey household, for which the random number is greater than or equal to its cumulative distribution, is assigned the total trips of that survey household.

Thus we can calculate the total number of trips made by each household for both scenarios. We can then observe the difference in the total trips generated in both methods. The result achieved from our study is listed in the next chapter.

5.0 Results and Analysis

5.1 Introduction:

The deployment of an alternative method to match the households in the Activity Generator, described in the previous chapter was applied to the input data. The original method was also applied to the same data and the results from both these runs were then compared. Further, sensitivity analysis carried out for some other alternative methods are explained in the next chapter.

5.2 Results from Scenario 1:

Using the original set of Ys (activity times as dependent variables), classification trees were obtained as shown in Chapter 4 from sections 4.3.1 to 4.3.4. Following the algorithm described in the earlier chapters the results achieved are given below.

The table below lists the total number of trips per end node that were calculated for each end node of the original trees. Also it lists the total trips of the synthetic households that were calculated after assigning them the survey household trips.

Table 5-1: Total number of Trips per end-node for Scenario 1

End Node	Total Trips Assigned/EndNode for the Synthetic HHs	Total Trips/EndNode for the Survey HHs
101	0	1694
102	0	70
103	0	852
104	0	1655
105	0	271
106	0	319
111	0	1052
112	0	1428
113	0	57
114	0	752
115	108355	3371
116	22084	1820
117	3824	371
118	6392	2015
119	9716	1095
1110	11474	615
121	6513	593
122	6160	1224
123	1545	134
124	10766	3331
125	88717	831
126	90100	2963
127	17357	328
128	101911	4074
129	17945	350
1210	25711	1054
1211	43256	543
131	181647	231
132	212537	663
133	95559	139
134	241769	384
135	69596	276
136	26714	294
137	141342	606
138	92632	478
	Total Trips Assigned = 1633622	Total Trips = 35933

Thus we can see that the total trips obtained from 3470 survey households are 35933 and these are assigned to the synthetic households randomly. The total number of trips thus obtained for our synthetic household data is 1,633,622 trips.

Table 5-2: Total number of HHs per end-node for Scenario 1

End Node	Total HHs/EndNode for the Synthetic HHs	Total HHs/EndNode for the Survey HHs
101	0	368
102	0	10
103	0	108
104	0	201
105	0	21
106	0	15
111	0	222
112	0	298
113	0	11
114	0	157
115	11807	367
116	1625	136
117	236	22
118	326	100
119	407	46
1110	454	22
121	670	59
122	654	129
123	180	20
124	1120	345
125	10412	97
126	6718	221
127	4871	25
128	5532	223
129	719	14
1210	1267	51
1211	1434	18
131	13375	17
132	15983	43
133	7532	11
134	13164	15
135	3568	20
136	1372	15
137	6067	26
138	3289	17
	Total Households = 112782	Total HHs = 3470

The table above classifies the survey households and the synthetic households to the end-nodes. This gives us an idea as to how the households were matched. Thus in all 3470 households were matched to the 112,782 households i.e. the trips of 3470 households were assigned to 112,782 households.

5.3 Results from Scenario 2:

Using the new set of *Y*s (travel times between activities as dependent variables), classification trees were obtained as shown in Chapter 4 from sections 4.3.5 to 4.3.8. Following the algorithm described in the earlier chapters the results achieved are given below:

The table below lists the total number of trips per end node that were calculated for each end node of the new trees. Also it lists the total trips of the synthetic households that were calculated after assigning them the survey household trips.

Table 5-3:Total number of Trips per end-node for Scenario 2

End Node	Total Trips Assigned/EndNode for the Synthetic HHs	Total Trips/EndNode for the Survey HHs
201	0	500
202	0	804
203	0	390
204	0	1749
205	0	728
206	0	100
207	0	214
208	0	376
211	0	261
212	0	679
213	5384	455
214	53551	1289
215	45589	2127
216	21855	1820
217	14711	1559
218	5541	558
219	3936	1096
2110	5693	883
221	3383	402
222	7639	1639
223	45389	3408
224	70551	954
225	21916	337
226	10857	293
227	77553	2721
228	45987	2818
229	6224	125
2210	44570	1325
2211	49926	1436
2212	30028	1085
231	201482	560
232	193132	483
233	3518	170
234	152718	351
235	368977	903
236	39308	452
237	86086	312
238	2889	51
239	1566	42
2310	92001	478
	Total Trips Assigned = 1711960	Total Trips = 35933

The total trips for the 3470 survey households are still the same 35933 and these are assigned to the synthetic households randomly. The total number of trips thus obtained for our synthetic household data is 1,711,960 trips.

Table 5-4:Total number of HHs per end-node for Scenario 2

End Node	Total HHs/EndNode for the Synthetic HHs	Total HHs/EndNode for the Survey HHs
201	0	94
202	0	151
203	0	123
204	0	214
205	0	84
206	0	21
207	0	11
208	0	25
211	0	555
212	0	133
213	540	46
214	6433	95
215	4833	226
216	1635	136
217	753	79
218	274	28
219	169	49
2110	218	34
221	293	18
222	1750	181
223	6585	354
224	7112	97
225	1860	29
226	612	13
227	5826	204
228	2918	144
229	501	10
2210	2700	69
2211	2165	45
2212	1255	38
231	13916	18
232	12502	12
233	291	14
234	11608	27
235	16433	14
236	1926	23
237	4142	12
238	170	14
239	73	13
2310	3289	17
	Total Households = 112782	Total HHs = 3470

The table above classifies the survey households and the synthetic households to the end-nodes of the new trees. This gives us an idea as to how the households were matched at every end-node. The same 3470 households were matched to the 112,782 households i.e. the trips of 3470 households were assigned to 112,782 households, but their distribution is different for this case now which means that the households that are matched are different than the ones in the previous scenario. This is the reason for the difference in the total number of trips of the synthetic households.

5.4 Further analysis of the results:

Thus we notice that there is a change in the total number of trips in these two methods. By using the original trees we get 1,633,622 trips for the network. But if we use the new trees we get 1,711,960 trips for the same network. This is an increase of 78,338 trips for the network. This is an increase of approximately 5 % trips. Now this 5 % increase is not only a change in the number of trips, but also the types of trips are bound to change with the change in assignment i.e. this change in trips could be distributed over the whole network or could also be a change in a small part of the network. In the latter case it would create a significant difference to the analysis of a network.

Also the thing to note here is that the synthetic household data that was used was of the family households. Which could be the reason why the total number of households with no workers is 0. While the number of households with workers greater than 2 is much higher. In case of non-family households or group-quarters this could be the opposite. But the survey data does not comply with such details. It is not classified into family, non-family or group-quarters. So what happens is that some households are left unmatched while some households are matched to many households. For example, for workers greater than 2, each survey household is matched to approximately 200 or more synthetic households. This is a very high number when we are trying to achieve as much accuracy as we can.

The table below shows the difference in the rate at which the households were matched for workers 1 and 2 and workers greater than 2.

Table 5-5: Comparison of results of Workers 1 and 2

Workers 1 and 2:	
Original Ys:	
Total HHs	*Assigned HHs*
2583	48432
Total Trips	*Assigned Trips*
28001	571826
New Ys:	
Total HHs	*Assigned HHs*
2583	48432
Total Trips	*Assigned Trips*
27270	570283

Table 5-6: Comparison of results for Workers > 2

Workers > 2:	
Original Ys:	
Total HHs	*Assigned HHs*
164	64350
Total Trips	*Assigned Trips*
3071	1061796
New Ys:	
Total HHs	*Assigned HHs*
164	64350
Total Trips	*Assigned Trips*
3802	1141677

Here we can see that for workers equal to 1 and 2 the total households that are matched are 2,583 to 48,432. While for workers greater than 2 the activities of 164 survey households are assigned to 64,350 synthetic households.

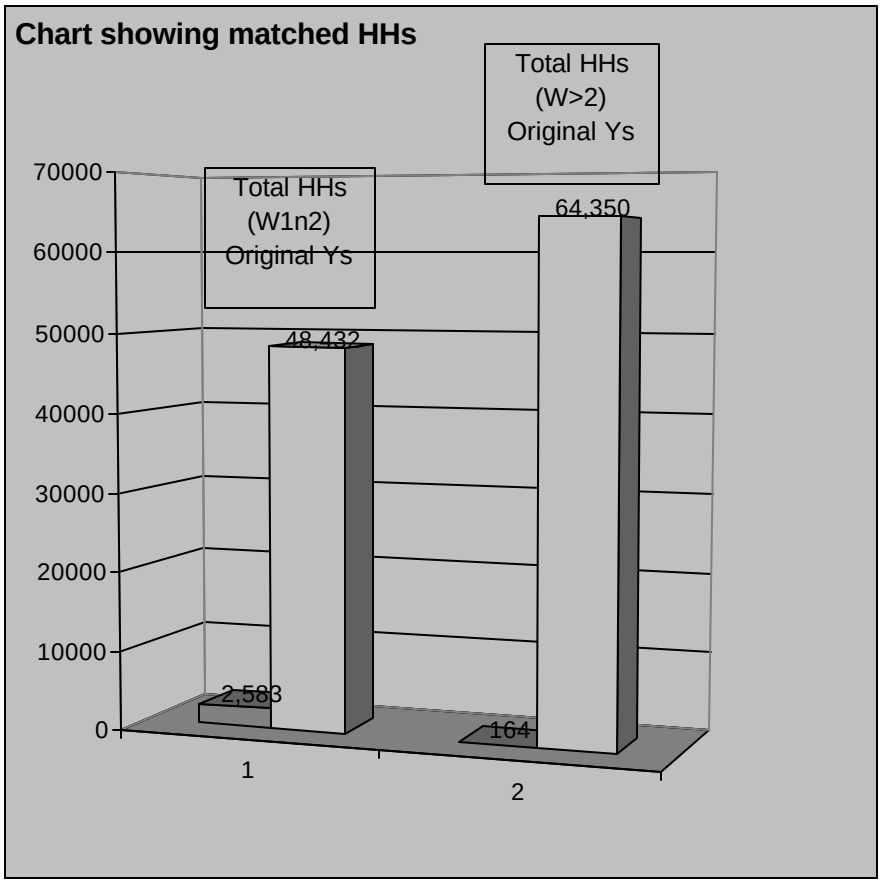


Figure 5-1: Chart showing matched Households

This could significantly affect the number of trips that are generated. And the trips thus generated cannot be validated as accurate.

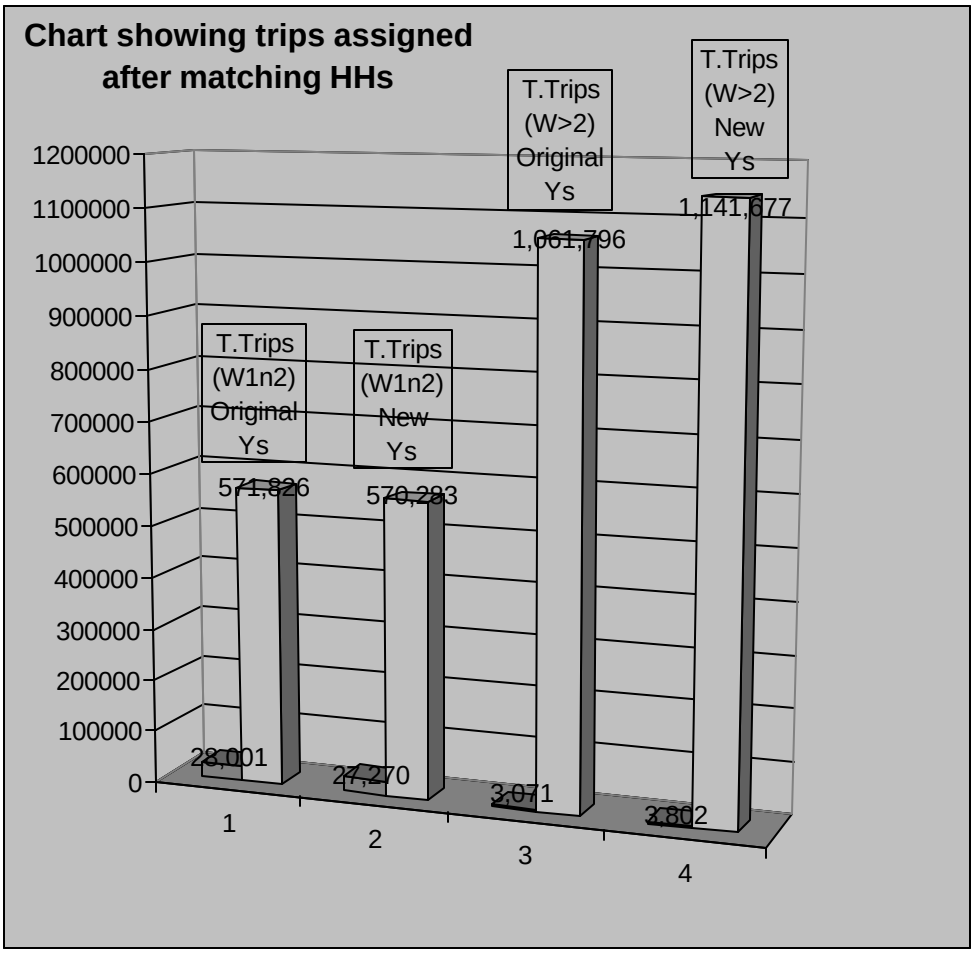


Figure 5-2: Chart showing trips assigned after matching Households

Thus looking at the above charts we can see that the significant difference in the number of trips generated using both methods is mainly in the households with workers greater than 2. For workers 1 and 2 there is a decrease of about 1500 in the total number of trips generated, while for workers greater than 2 there is an increase of almost 140,000 trips. This again proves the point that the initial split of households based on number of workers in the household makes our results inconclusive because of the uneven split of number of households.

5.5 Conclusions:

The main purpose of the CART algorithm, used in the Activity Generator Module, is to produce an accurate classification of household characteristics based on household travel

behaviors. The binary tree structures 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. These splits are performed until all leaf nodes can be declared to be terminal nodes. The first split that SPLUS makes is based on the household size. So further research could be carried out on the same subject but without using the number of workers as the first split. This could give results that can prove to be more conclusive.

6.0 Sensitivity Analysis

6.1 Introduction:

Sensitivity analysis was carried out in the study to compare the results of altering some of the dependent variables as well as the independent variables. The following analyses were carried out and will be discussed in this chapter.

A - Altering the Independent Variables (X):

Case-1: Combining ALT5 and A5To17 as one *X* and using this with all original *Y.*

Case-2: Discarding A26To45 and using the rest *X* variables with all original *Y.*

B - Altering the Dependent Variables (Y):

Case-3: Using only one Y = Total number of trips with the *X* variables.

Case-4: Combining HbasedWork and WbasedHome trips and using its travel times as one dependent variable.

C - Altering the Survey data set.

Case-5: The survey household data set is split and only partial data set is used for tree building. The partial data set is then matched with the entire set.

The classification trees for all the cases are included in *Appendix A.* The trips assignment tables below list the total number of trips per end node that were calculated for each end node of the classification trees. It also lists the total trips of the synthetic households that were calculated after assigning them the survey household trips. The household assignment tables below list the total number of households assigned per end node for every case.

6.2 Case-1: Combining ALT5 and A5To17 as one X and using this with all original Y.

The demographic data was first modified to group the two sets of age categories, ALT5 (Age less than 5) and the A5To17 (Age between 5 and 17). The combined value was used as one independent variable. Using the new *X*s with the original set of *Y*s, classification trees were obtained following the algorithm described in the earlier chapters. The results achieved are given below. See *Appendix A* for the classification tree.

Table 6-1: Trip Assignment for Case-1

End Node	Total Trips Assigned/EndNode	Total Trips/EndNode
201	0	1694
202	0	35
203	0	885
204	0	1657
205	0	271
206	0	165
207	0	154
211	0	1050
212	0	1421
213	0	57
214	0	752
215	79575	3371
216	17954	1820
217	3433	371
218	5503	2015
219	8096	1095
2110	9807	615
221	5541	593
222	5034	1224
223	1584	228
224	8247	3331
225	62528	831
226	67515	2963
227	14647	328
228	73190	4074
229	12836	350
2210	13972	1054
2211	9641	137
2212	21078	406
231	50877	136
232	77294	95
233	99985	578

234	58548	139
235	106530	181
236	103410	655
237	5206	118
238	104756	606
239	34879	277
2310	33514	201
	Total Trips Assigned =	Total Trips =
	1095180	35933

Thus we can see that the total trips obtained from 3470 survey households are 35933 and these are assigned to the synthetic households randomly. The total number of trips thus obtained for our synthetic household data is 1,095,180 trips.

Table 6-2: Household Assignment for Case-1

End Node	Total HHs/ EndNode	Total HHs/EndNode
201	0	368
202	0	5
203	0	112
204	0	202
205	0	21
206	0	7
207	0	8
211	0	222
212	0	298
213	0	11
214	0	157
215	8725	367
216	1156	136
217	167	22
218	313	100
219	418	46
2110	587	22
221	426	59
222	360	129
223	127	20
224	793	345
225	7724	97
226	6228	221
227	1605	25
228	7064	223
229	1340	14
2210	1307	51
2211	739	5
2212	1904	13

231	5651	10
232	8587	7
233	11388	43
234	7729	11
235	10863	11
236	11869	34
237	752	5
238	9231	26
239	3043	10
2310	2684	7
	Total Households =	Total HHs =
	112782	3470

Table 6.2, classifies the survey households and the synthetic households to the end-nodes. This gives us an idea as to how the households were matched using the classification for this case. Thus in all 3470 households were matched to the 112,782 households i.e. the trips of 3470 households were assigned to 112,782 households.

Table 6-3: Trip distribution by workers in household for Case-1

	Survey HHs	*Synthetic HHs*	*Survey HH %*	*Multiplication Factor*
All Workers	**35,933**	**1,095,180**	**3.28**	**30**
Workers 0	4,861	0	0.00	0
Workers 1	12,567	124,368	10.10	10
Workers 2	15,519	295,813	5.25	19
Workers Gr2	2,986	674,999	0.44	226

6.3 Case-2: Discarding A26To45 and using the rest X variables with all original Y.

The demographic data was modified to discard the age category, A26To45 (Age between 26 and 45). Using the *X*s without A26To45 with the original set of *Y*s, classification trees were obtained following the algorithm described in the earlier chapters. The results achieved are given below. See *Appendix A* for the classification tree.

Table 6-4: Trip Assignment for Case-2

End Node	Total Trips Assigned/EndNode	Total Trips/EndNode
201	0	1694
202	0	35
203	0	885
204	0	1657
205	0	271
206	0	165
207	0	154

211	0	1050
212	0	1421
213	0	57
214	0	752
215	79575	3371
216	17954	1820
217	3433	371
218	5503	2015
219	8096	1095
2110	9807	615
221	5541	593
222	5034	1224
223	1584	228
224	8247	3331
225	62528	831
226	67515	2963
227	14647	328
228	73190	4074
229	12836	350
2210	13972	1054
2211	9641	137
2212	21078	406
231	50877	136
232	77294	95
233	99985	578
234	58548	139
235	106530	181
236	103410	655
237	5206	118
238	104756	606
239	34879	277
2310	33514	201
	Total Trips Assigned =	Total Trips =
	1095180	35933

Thus we can see that the total trips obtained from 3470 survey households are 35933 and these are assigned to the synthetic households randomly. The total number of trips thus obtained for our synthetic household data is 1,095,180 trips. This is the same as that obtained from Case 1. This was expected because the classification trees remained the same for both cases. This proves that discarding the A26To45 variable did not affect the classification at all.

Table 6-5: Household Assignment for Case-2

End Node	Total HHs/EndNode	Total HHs/EndNode
201	0	368
202	0	5
203	0	112
204	0	202
205	0	21
206	0	7
207	0	8
211	0	222
212	0	298
213	0	11
214	0	157
215	8725	367
216	1156	136
217	167	22
218	313	100
219	418	46
2110	587	22
221	426	59
222	360	129
223	127	20
224	793	345
225	7724	97
226	6228	221
227	1605	25
228	7064	223
229	1340	14
2210	1307	51
2211	739	5
2212	1904	13
231	5651	10
232	8587	7
233	11388	43
234	7729	11
235	10863	11
236	11869	34
237	752	5
238	9231	26
239	3043	10
2310	2684	7
	Total Households =	Total HHs =
	112782	3470

Table 6.5, classifies the survey households and the synthetic households to the end-nodes. In all 3470 households were matched to the 112,782 households i.e. the trips of 3470 households were assigned to 112,782 households. This classification remains exactly the same as that for Case 1. Thus the distribution of not only the total number of trips but also the households was same as in Case 1.

Table 6-6: Trip distribution by workers in household for Case-2

	Survey HHs	*Synthetic HHs*	*Survey HH %*	*Multiplication Factor*
All Workers	**35,933**	**1,095,180**	**3.28**	**30**
Workers 0	4,861	0	0.00	0
Workers 1	12,567	124,368	10.10	10
Workers 2	15,519	295,813	5.25	19
Workers Gr2	2,986	674,999	0.44	226

6.4 Case-3: Using only one Y = Total number of trips with the X variables.

For this case the independent variables were kept the same as that used in TRANSIMS. But for the dependent variables only the "total number of trips per household" was used. The classification or households and trips assigned results achieved are given below. See *Appendix A* for the classification tree.

Table 6-7: Trip Assignment for Case-3

End Node	Total Trips Assigned/EndNode	Total Trips/EndNode
201	0	227
202	0	1077
203	0	390
204	0	134
205	0	650
206	0	191
207	0	585
208	0	172
209	0	337
2010	0	277
2011	0	121
2012	0	110
2013	0	271
2014	0	132
2015	0	187
211	0	3289
212	107788	3371
213	22019	1820

214	7158	1336
215	8672	1050
216	15234	1446
217	8653	261
221	3284	144
222	13179	2660
223	15280	1806
224	7169	683
225	92158	914
226	55760	1546
227	54586	1745
228	41425	220
229	1398	239
2210	19706	2190
2211	30890	169
2212	43320	1082
2213	7122	174
2214	42645	1404
2215	43392	543
231	10269	53
232	4125	389
233	4380	342
234	530689	164
235	181771	442
236	0	189
237	8417	116
238	86032	201
239	129980	179
2310	0	163
2211	37118	373
2212	37174	369
	Total Trips Assigned =	**Total Trips =**
	1670793	**35933**

Thus we can see that the total trips obtained from 3470 survey households are 35933 and these are assigned to the synthetic households randomly. The total number of trips thus obtained for our synthetic household data is 1,670,793 trips.

Table 6-8: Household Assignment for Case-3

End Node	Total HHs/EndNode	Total HHs/EndNode
201	0	56
202	0	189
203	0	123
204	0	20
205	0	77

206	0	29
207	0	62
208	0	15
209	0	51
2010	0	33
2011	0	9
2012	0	23
2013	0	21
2014	0	5
2015	0	10
211	0	688
212	12425	367
213	1710	136
214	420	73
215	420	49
216	659	61
217	238	7
221	290	12
222	1515	293
223	1556	177
224	706	63
225	11166	105
226	4705	124
227	4028	122
228	2980	15
229	87	14
2210	998	105
2211	1930	10
2212	2766	65
2213	603	14
2214	2090	65
2215	1509	18
231	1251	6
232	300	27
233	373	28
234	33989	10
235	10242	24
236	0	13
237	374	5
238	3604	8
239	6849	9
2310	0	5
2211	1752	17
2212	1245	12
	Total Households =	**Total HHs =**
	112782	**3470**

Table 6.8, classifies the survey households and the synthetic households to the end-nodes. In all 3470 households were matched to the 112,782 households i.e. the trips of 3470 households were assigned to 112,782 households.

Table 6-9: Trip distribution by workers in household for Case-3

	Survey HHs	Synthetic HHs	Survey HH %	Multiplication Factor
All Workers	**35,933**	**1,670,793**	**2.15**	**46**
Workers 0	4,861	0	0.00	0
Workers 1	12,573	169,524	7.42	13
Workers 2	15,519	471,314	3.29	30
Workers Gr2	2,980	1,029,955	0.29	346

6.5 Case-4: Combining HbasedWork and WbasedHome trips and using its travel times as one dependent variable.

For this case the independent variables were kept the same as that used in TRANSIMS. The dependent variables used are given below:

Y_1 = total travel times for household in minutes for trips starting from Home to Work and then back to Home (THomeWorkHomeBased),

Y_2 = total travel times for household in minutes for trips starting from Home to destinations other than Work and back to Home (THomeBasedOther),

Y_3 = total travel times for household in minutes for trips starting from Work to destinations other than Home and back to Work (TWorkBasedOther),

Y_4 = total number of trips per household (TTrips),

They are the same as those used for the alternative scenario in base case except, the travel times for the Home-based-Work and Work-based-Home trips were combined and used as one dependent variable. The results achieved for classification of households and trips assigned are given below. See *Appendix A* for the classification tree.

Table 6-10: Trip Assignment for Case-4

End Node	Total Trips Assigned/EndNode	Total Trips/EndNode
201	0	500
202	0	804
203	0	390
204	0	443
205	0	123

206	0	333
207	0	233
208	0	413
209	0	204
2010	0	264
2011	0	144
2012	0	320
2013	0	100
2014	0	233
2015	0	245
2016	0	112
211	0	2610
212	0	679
213	5243	455
214	53391	789
215	46268	2127
216	21740	1820
217	14557	1559
218	482	183
219	7462	762
2110	4054	1307
2111	4065	315
221	3404	206
222	7570	1639
223	45720	3408
224	69730	954
225	21133	337
226	11041	233
227	77568	2721
228	45724	2818
229	0	0
2210	44889	1131
2211	12110	1094
2212	878	672
2213	573	78
2214	47831	103
231	200985	240
232	81229	187
233	64585	280
234	29233	191
235	56690	70
236	112432	176
237	116433	178
238	20126	403
239	97104	215
2310	22450	271

2211	8389	164
2212	9466	219
2213	91650	478
	Total Trips Assigned =	**Total Trips =**
	1456205	**35933**

Thus we can see that the total trips obtained from 3470 survey households are 35933 and these are assigned to the synthetic households randomly. The total number of trips thus obtained for our synthetic household data is 1,456,205 trips.

Table 6-11: Household Assignment for Case-4

End Node	Total HHs/EndNode	Total HHs/EndNode
201	0	94
202	0	151
203	0	123
204	0	56
205	0	13
206	0	46
207	0	25
208	0	52
209	0	22
2010	0	28
2011	0	18
2012	0	38
2013	0	21
2014	0	11
2015	0	16
2016	0	9
211	0	555
212	0	133
213	605	46
214	7204	95
215	5412	226
216	1820	136
217	843	79
218	37	11
219	390	36
2110	181	55
2111	132	9
221	328	18
222	941	181
223	5350	354
224	7964	97
225	2083	29
226	685	13

227	6524	204
228	2610	144
229	0	0
2210	3024	69
2211	560	46
2212	39	26
2213	45	5
2214	3134	6
231	16808	18
232	5868	12
233	5413	21
234	2693	14
235	5443	6
236	7181	10
237	8868	12
238	1109	20
239	3696	17
2310	1196	13
2211	564	7
2212	347	7
2213	3683	17
	Total Households =	**Total HHs =**
	112782	**3470**

Table 6.11, classifies the survey households and the synthetic households to the end-nodes. In all 3470 households were matched to the 112,782 households i.e. the trips of 3470 households were assigned to 112,782 households.

Table 6-12: Trip distribution by workers in household for Case-4

	Survey HHs	*Synthetic HHs*	*Survey HH %*	*Multiplication Factor*
All Workers	**35,933**	**1,456,205**	**2.47**	**41**
Workers 0	4,861	0	0.00	0
Workers 1	12,606	157,262	8.02	12
Workers 2	15,394	388,171	3.97	25
Workers Gr2	3,072	910,772	0.34	296

6.6 Case-5: The survey household data set is split and only partial data set is used for tree building and matching.

The survey household data set used for this study consists of 3470 households. The activities of these households are matched with those of the synthetic population data containing 112,782 households. In this research the matching is done using two different

methods. But in order to calibrate the methodology we will need actual data for the synthetic households. This data is not available, but alternatively we could split the survey households into a set which contains 10% of the total data. This data should be treated as survey data and the 100% of the sample should be treated as synthetic households. The matching should be carried out using both the methods and then we can calibrate the output by comparing it to the actual data available for the 100% survey household population.

The survey household data set used in our research contains a list of activities for 3470 households. This data was split to create another set of data with 10% of the population contained in the original data set. The new set of data contains approximately 347 households. This data set was treated as survey household data while the 3470 household data set was treated as the synthetic households.

According to both the methodologies a classification tree should be grown for the 347 households. The first split would be for workers = 0, 1, 2 and Greater than 2. This split makes the data sets in each end nodes very small. SPLUS does not grow any trees beyond this split because of the lack of non-homogeneity between the households at the end nodes. It is necessary to have a certain degree of diversity between the household sets so as to split and group them into further end nodes. This makes it impossible to carry out household matching using the current data set.

This calibration is possible if we use a larger data set for survey households.
It is thus recommended to approach this scenario with a larger survey household data set.

6.7 Summary for the Sensitivity Analysis:

The results of the sensitivity analyses are listed in Table 6.13. The total number of trips for all cases is different than that of the base case. The difference in total trips from the base case is listed in the table.

Table 6-13: Results from all Cases

Case	Trips	Difference from Base Case
Base Case - Original Ys	1,633,622	0
Alt Case - New Ys	1,711,960	78,338
Case 1	1,095,180	-538,442
Case 2	1,095,180	-538,442
Case 3	1,670,793	37,171
Case 4	1,456,205	-177,417

We can see that the number of trips in Case 1 and 2 is far less than the base case. So when we altered the independent variables by combining ALT5 and A5To17 the total number of trips went down by 538,442 trips. The exclusion of independent variable A26To45 did not affect this number. The classification tree remained the same as in Case 1, even when one of the independent variables was discarded. So it is observed here that the variables ALT5 (age less than 5) and A5To17 (age 5 to 17) play a much more important role in the building of the classification tree than the variable A26To45 (age between 26 and 45).

In Case 3, when we used only one dependent variable (total number of trips of household per household), the results were very close to the base case results. The difference in the total number of households matched was very small and the trips assigned increased by only 37,171. Our observations suggest that the classification of households by the total number of trips they make is almost the same as the classification done by using the original variables used in TRANSIMS. Though the households matched will be different because of the different classification trees for both cases. This will result in different trips being assigned to the synthetic households for both cases. And so even though the difference in the total number of trips being assigned is not high, the model needs to be calibrated in order to conclude its validity.

For Case 4, the dependent variables used were the travel time variables and the home-based travel times and the work-based travel times were combined and used as one variable. The result was a decrease of 177,477 trips from the base case and a decrease of 255,755 trips from alternative suggested for the base case.

The sensitivity analysis was carried out to observe the results of altering the dependent and independent variables used in building the classification tree. The classification tree determines how the households are grouped and matched. The activities of the survey households are then assigned to the matched synthetic households. This makes the classification tree a very important part of the simulation process, as it will determine the total number of trips and type of trips on the network. It is thus important to develop an accurate methodology to build a classification tree that most homogeneously groups the households.

7.0 Conclusions and Recommendations for Further Research

7.1 Conclusions:

The TRANSIMS software developed by the Los Alamos National Laboratory is still in a stage of development. Changes are still being made to the original algorithm as research continues.

The Activity Generator module that we deal with in our study is also one of the modules which require further refinement, considering the magnitude of data handled by TRANSIMS to generate accurate data to imitate a real-life network.

One of the areas in the Activity Generator, which could be further refined, is the household matching and the trip assignment method adopted by TRANSIMS. The purpose of this study was to analyze the concept of choosing activity times as the basis for matching the households with similar demographic data. And then compare it with a new concept of using cumulative travel times between significant activities and the total number of trips made by the household as the basis for matching.

The results in the previous chapters have shown us that:

- There is definitely a significant difference in the total number of trips that are generated if the dependent variables are changed or altered. Changing the dependent variables gives us a different tree, and thus changing the total number of end-nodes and the split criteria. This causes a change in the households that are grouped in the end-nodes and thus matched. The trips thus assigned will be different compared to the original case. Hence not only will this affect the number of trips generated, but also the types of trips generated will be different. This will significantly affect the analysis of the network. The sensitivity analysis that was carried out also proved that further research is required using accurate real-life data to analyze different approaches to matching the households.

- The bias of number of workers as the first split for building the trees needs to be reconsidered. It could be used if the survey household data can precisely represent the synthetic data (which is primarily the census data of that region). So if the households that are matched are proportional then the bias can be used to match the activities. The tree is built using the survey data and in our study when the first split is done using number of workers, it is hard to predict whether or not the synthetic household data would comply with the split. As mentioned before, if the non-family and group quarters were also included, the picture would have been different. But then it can also be said that the activity survey data that was used in our study, was not sufficient enough to match the high number of synthetic household data. It is very important to use a data that can most accurately represent the real life data.
- The small sample size of the activity survey data also hampered the calibration of the methodology proposed in the study.

7.2 Recommendations:

The following are the recommendations for future research in this area:

- The bias of number of workers can be modified for family households, non-family households and group quarters i.e. in our study if we combine the number of workers = 0 and 1 into one node and workers = 2 and greater than 2 into another node and use that data to build further trees, then it could give us better results. Thus one recommendation would be to choose different split criteria for different set of population data.
- We can also further analyze the types of trips that are affected by altering the matching process. This could also play a major role in the network analysis process.
- The survey activity data should represent at least 5 % of the total population that is being analyzed to properly conclude something. Which means that every survey household will be matched to approximately 20 synthetic households. Also that data should be able to accurately represent the area under study. The best way to make sure of this is to use the activity data for the same region as under study.

BIBLIOGRAPHY

1. "A System of Activity Based Models for Portland, Oregon", TMIP, USDOT, May 1998
2. "Activity Based Modeling System for Travel Demand Forecasting", Prepared by RDC Inc. for Metropolitan Washington Council of Governments, September 1995
3. Beckman, R., K. Baggerly and M. McKay, "Creating Synthetic Baseline Populations", Los Alamos National Laboratory Unclassified Report, Los Alamos, NM, 1995
4. Bell, J. F., Tree-based methods. The use of classification trees to predict species distributions. Kluwer Associates, 1999
5. Bowman John L. and Moshe Ben-Akiva, "Activity Based Travel Forecasting", Conference Proceedings, June 1996
6. Breiman, L., Friedman, J., Olshen, R. and Stone, C. J., Classification and Regression Trees, Chapman and Hall, New York, 1984
7. Bush B. W. and the TRANSIMS Team Los Alamos National Laboratory, "The TRANSIMS Framework", June 1999
8. Jussi Klemala, Sigbert Klinke and Hizir Sofyan, "Classification and Regression Trees", April 2001
9. Nagpaul P. S., "Guide to Advanced Data Analysis using IDAMS Software", New Delhi, India
10. Paula Stretz and the TRANSIMS team, "TRANSIMS Activity Generator", January 2001
11. Spear Bruce D. and John A., "New Approaches to Travel Forecasting Models: A Synthesis of Four Research Proposals", Volpe National Transportation Systems Center, USDOT, January 1994
12. Taylor P., Caley R., Black A. W., King S., Edinburgh Speech Tools Library, Chapter 10: Classification and Regression Trees, System Documentation Edition 1.2, Centre for Speech Technology, University of Edinburgh, June 1999
13. TRANSIMS 1.1, Volume 3 – Modules, Los Alamos National Laboratory, April 2000
14. TRANSIMS 2.0, Volume 4 – Calibrations, Scenarios and Tutorials, Los Alamos National Laboratory, April 2001

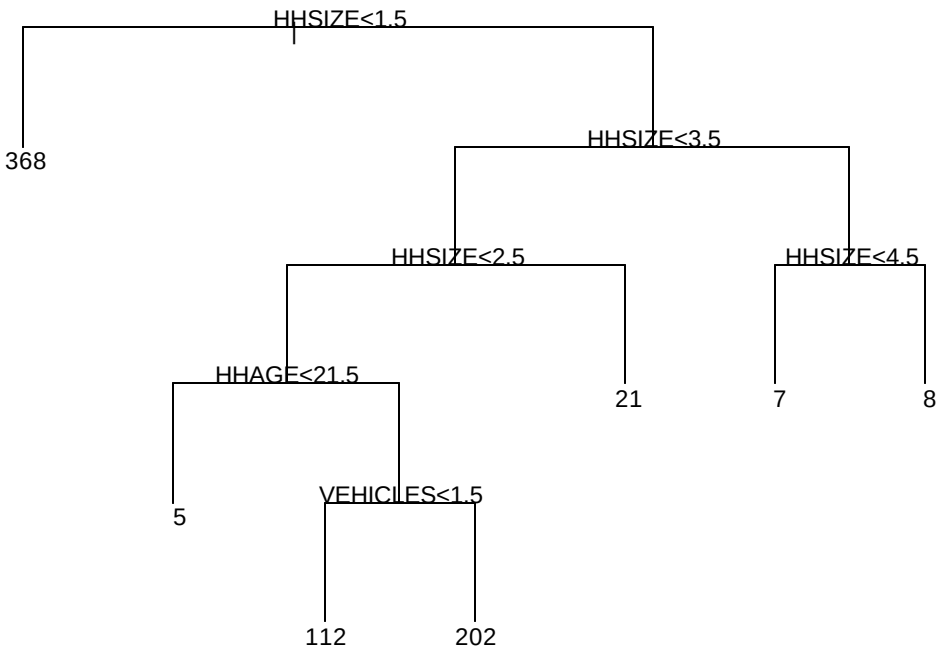
15. TRANSIMS 2.0, Volume 2 – Networks and Vehicles, Los Alamos National Laboratory, March 2001
16. TRANSIMS 2.0, Volume 5 – Software: Interface Functions and Data Structures, Los Alamos National Laboratory, May 2001
17. "TRANSIMS Course Manual – Activity Generator", originally prepared at Virginia Tech by Hobeika et. al., 2000.
18. "TRANSIMS Course Manual – Population Synthesizer", originally prepared at Virginia Tech by Hobeika et. al., 2000.
19. Vaughn K. M., Speckman P. and Pas E. I., "Generating Household Activity-Travel Patterns for Synthetic Populations", December 1997
20. Vaughn K. M., Speckman P. and Sun D., "Identifying Relevant Socio-Demographics for Distinguishing Household Activity-Travel Patterns: A Multivariate Regression Tree Approach", NISS, May 1999

APPENDIX

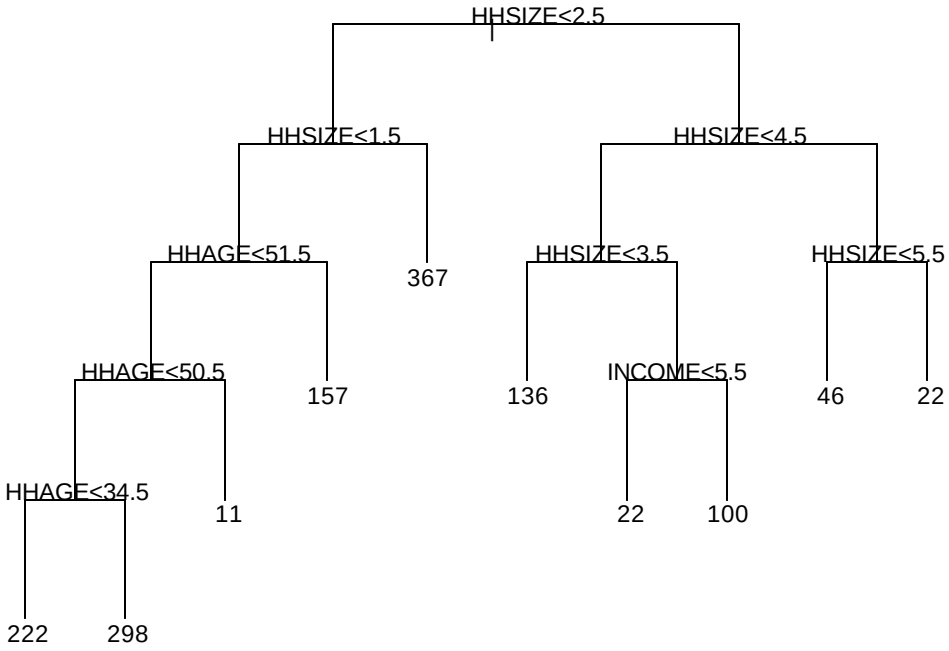
APPENDIX-A

- **Sensitivity Analysis: Case1 Classification Trees**

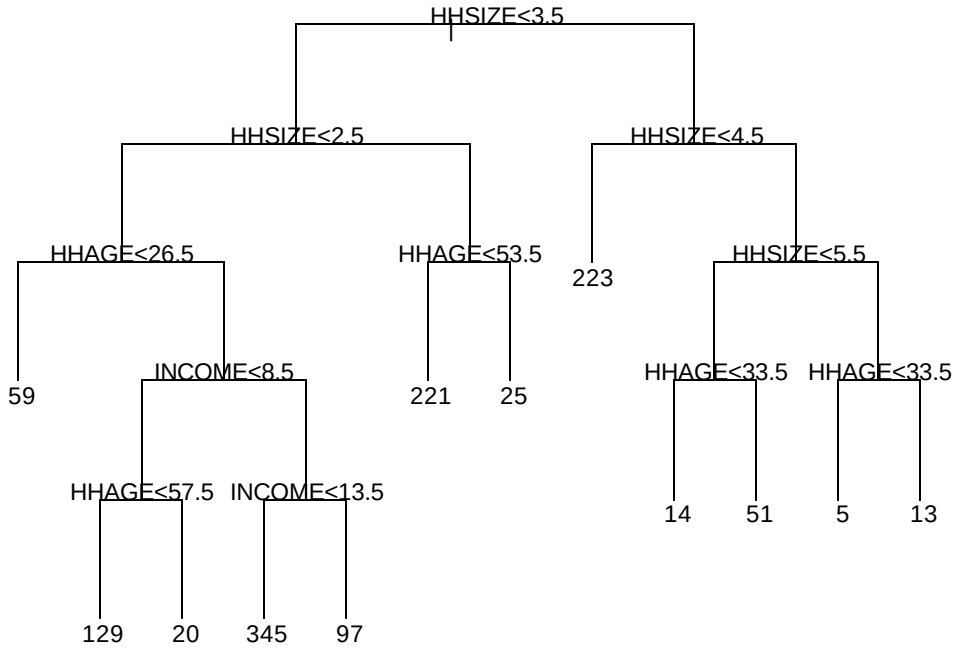
1) Classification tree for Workers = 0



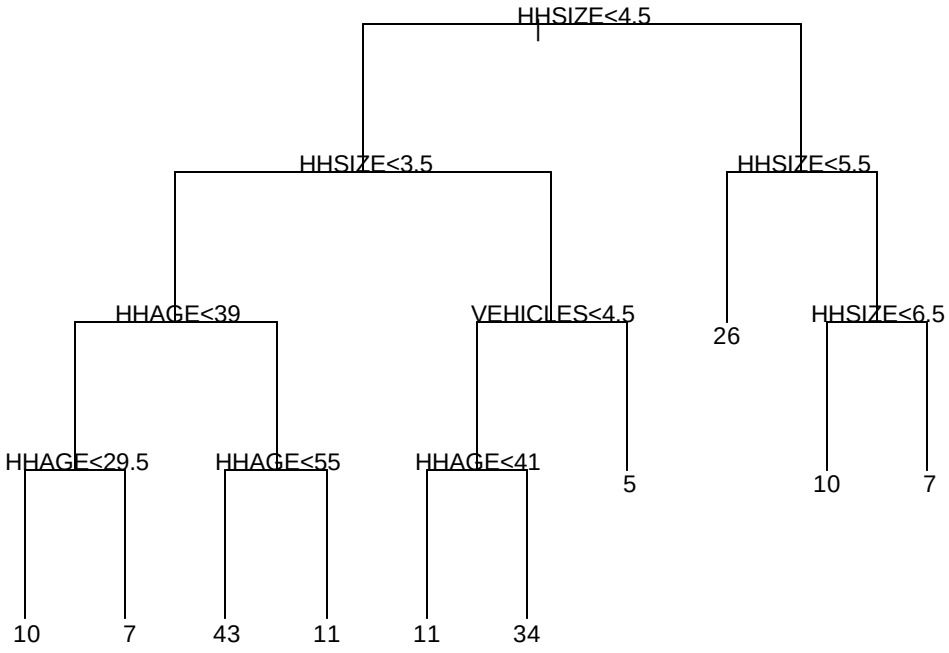
2) Classification tree for Workers = 1



3) Classification tree for Workers = 2

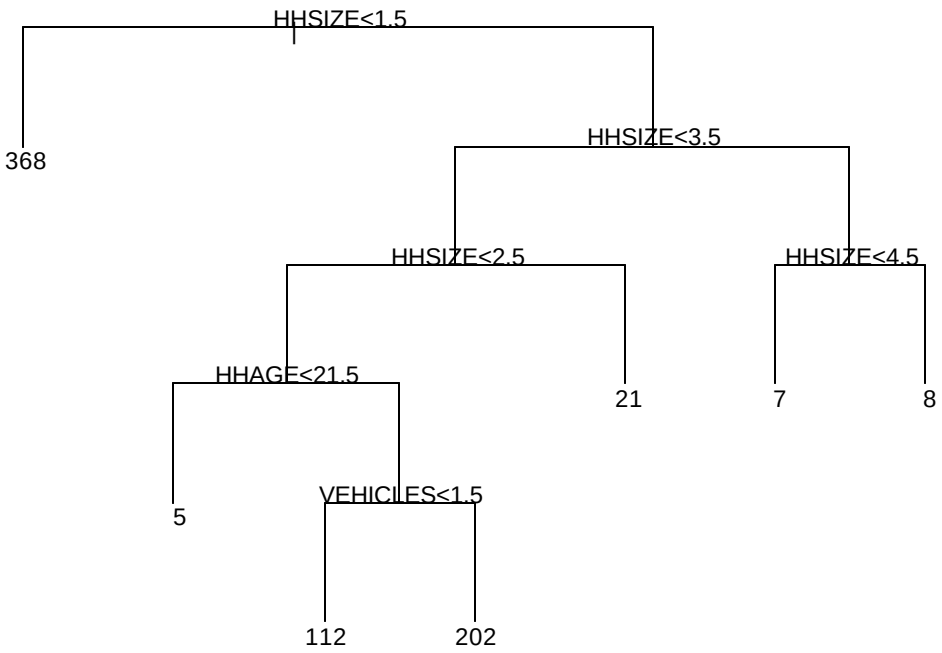


4) Classification tree for Workers > 2

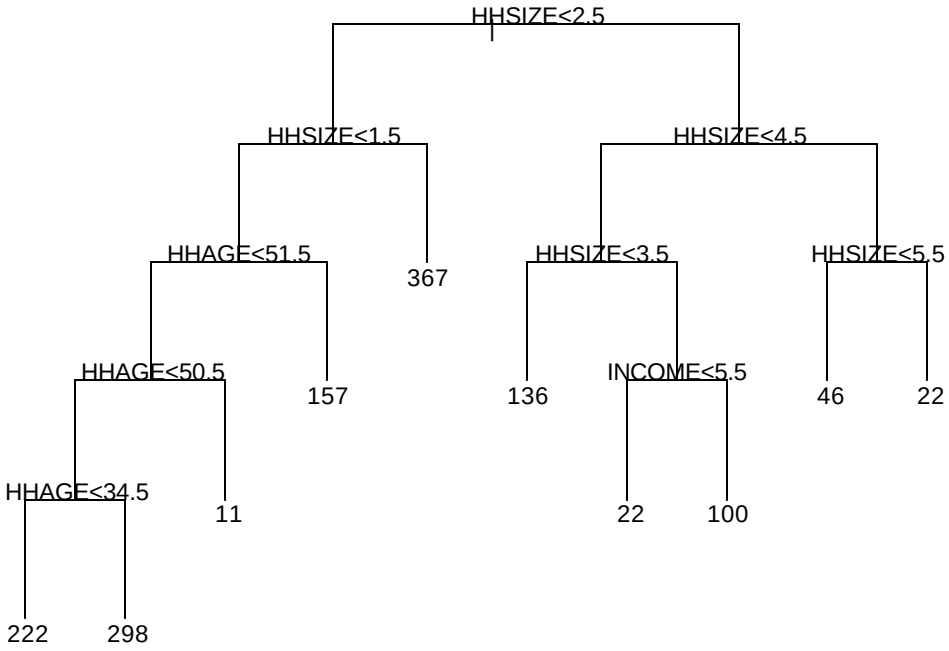


- **Sensitivity Analysis: Case2 Classification Trees**

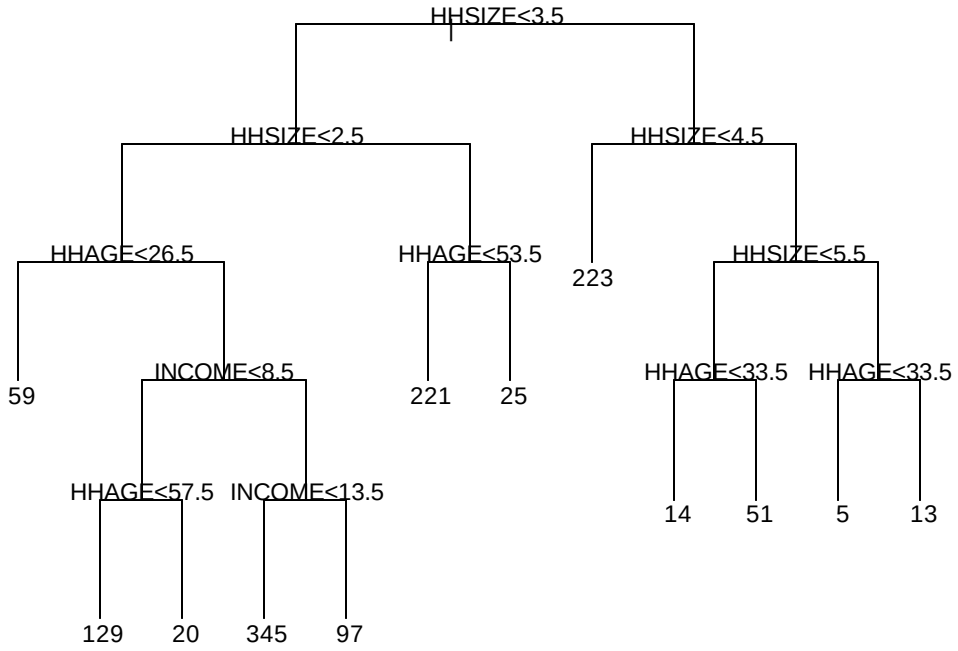
1) Classification tree for Workers = 0



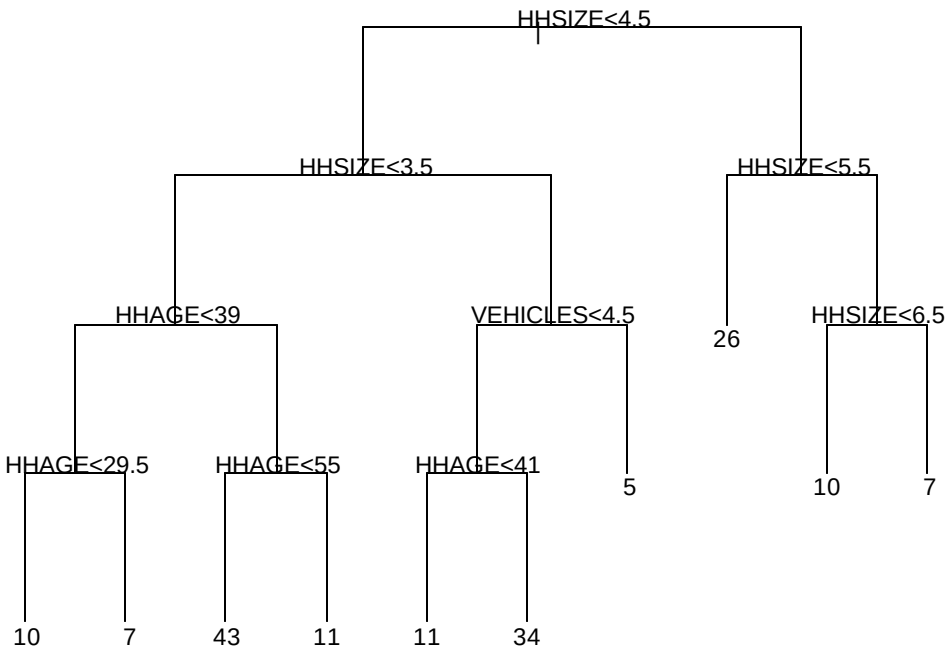
2) Classification tree for Workers = 1



3) Classification tree for Workers = 2

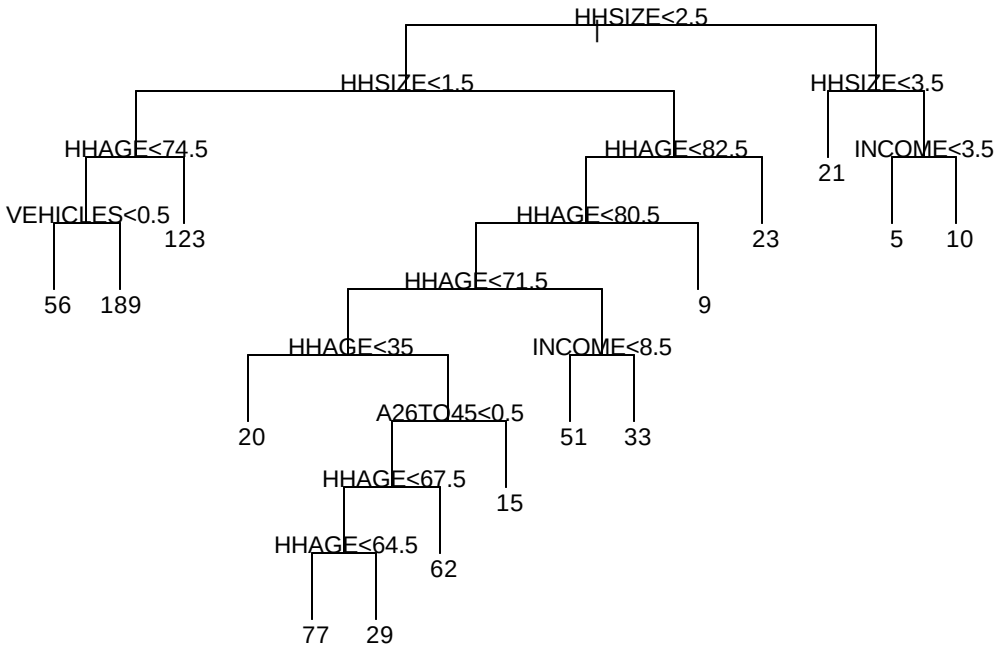


4) Classification tree for Workers > 2

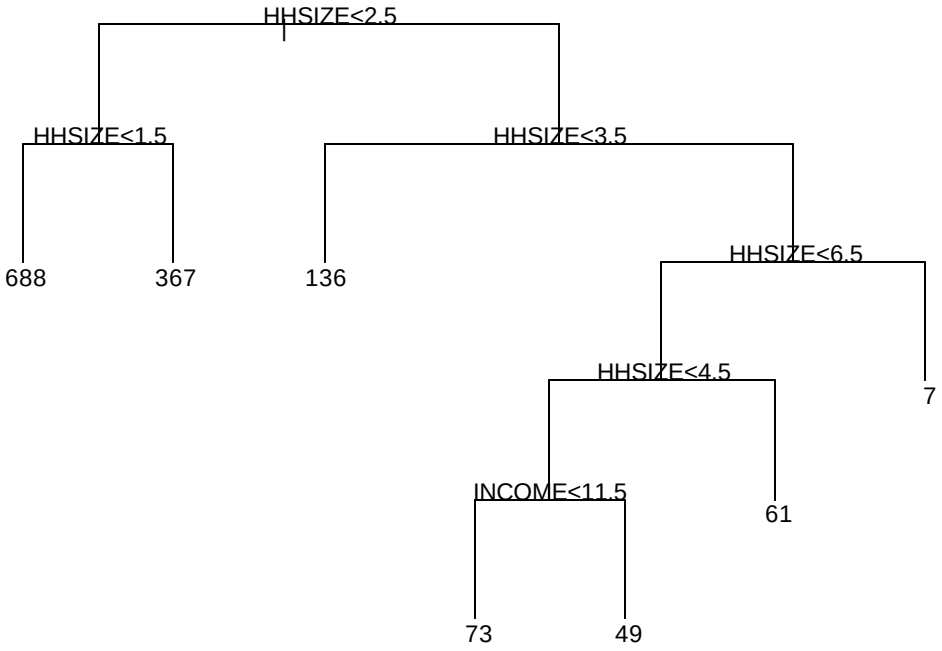


- **Sensitivity Analysis: Case3 Classification Trees**

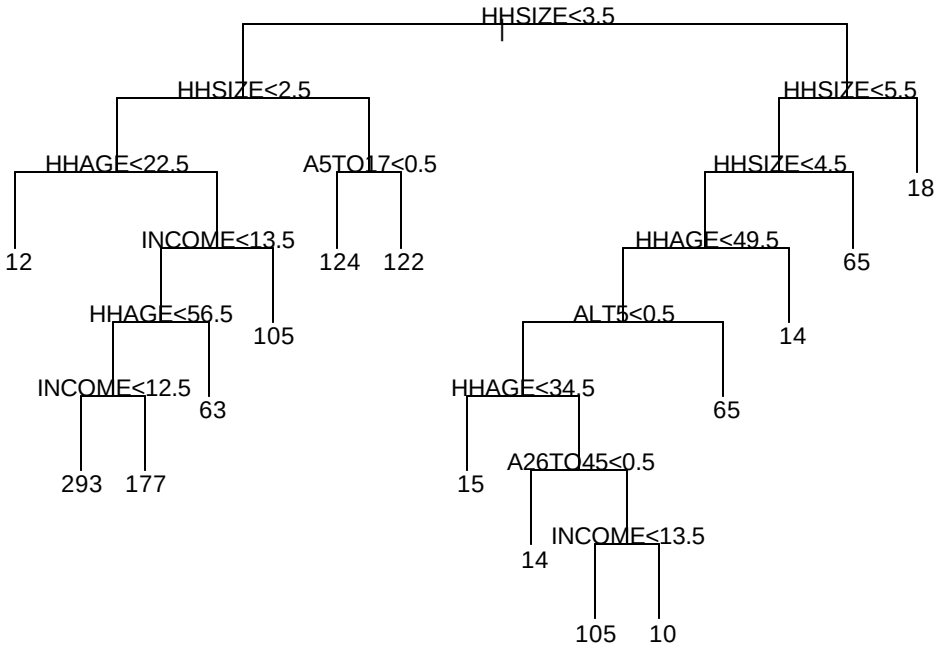
1) Classification tree for Workers = 0



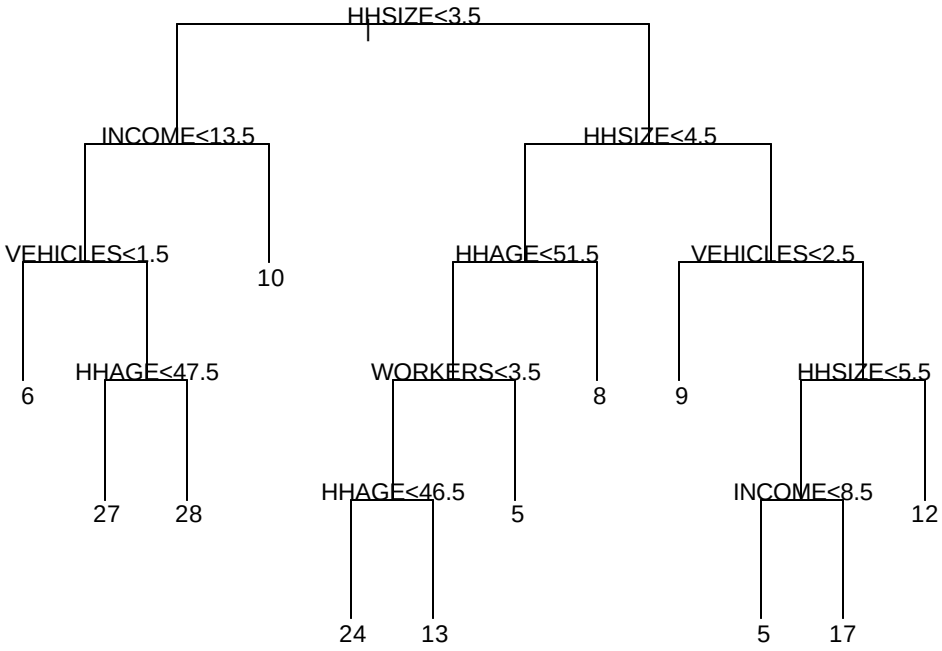
2) Classification tree for Workers = 1



3) Classification tree for Workers = 2

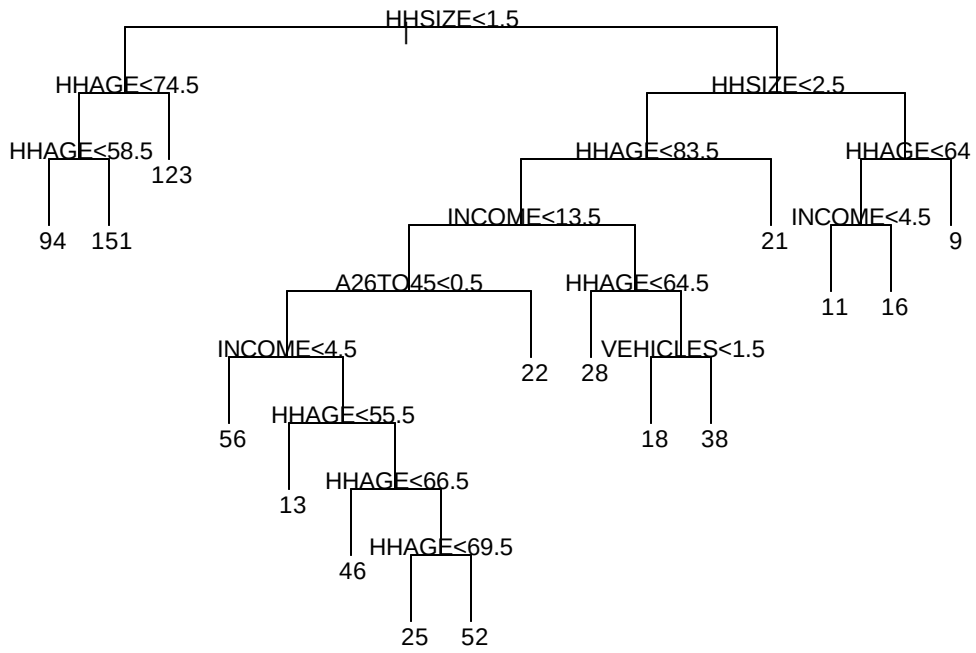


4) Classification tree for Workers > 2

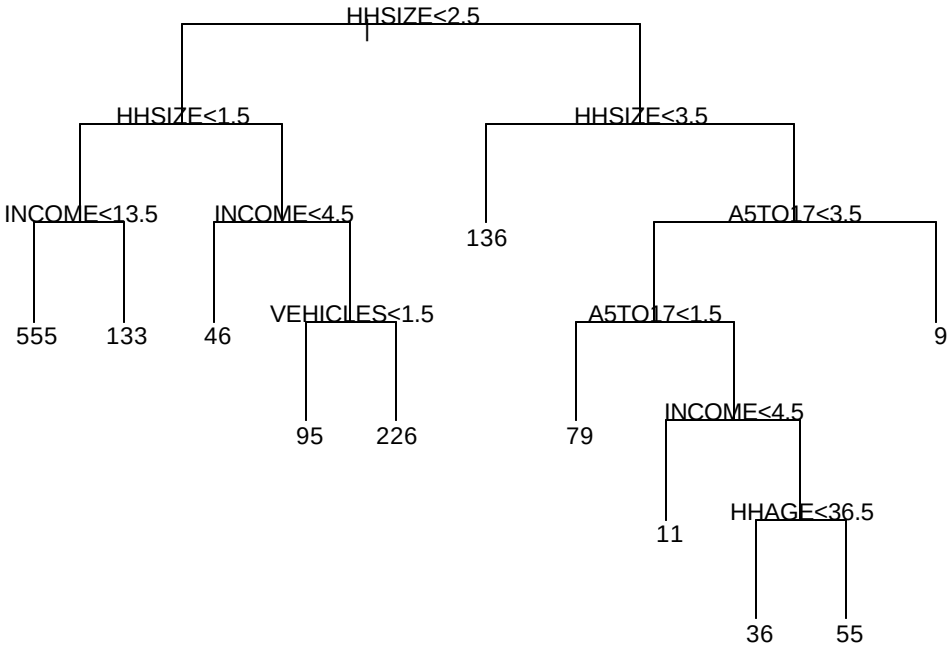


- **Sensitivity Analysis: Case4 Classification Trees**

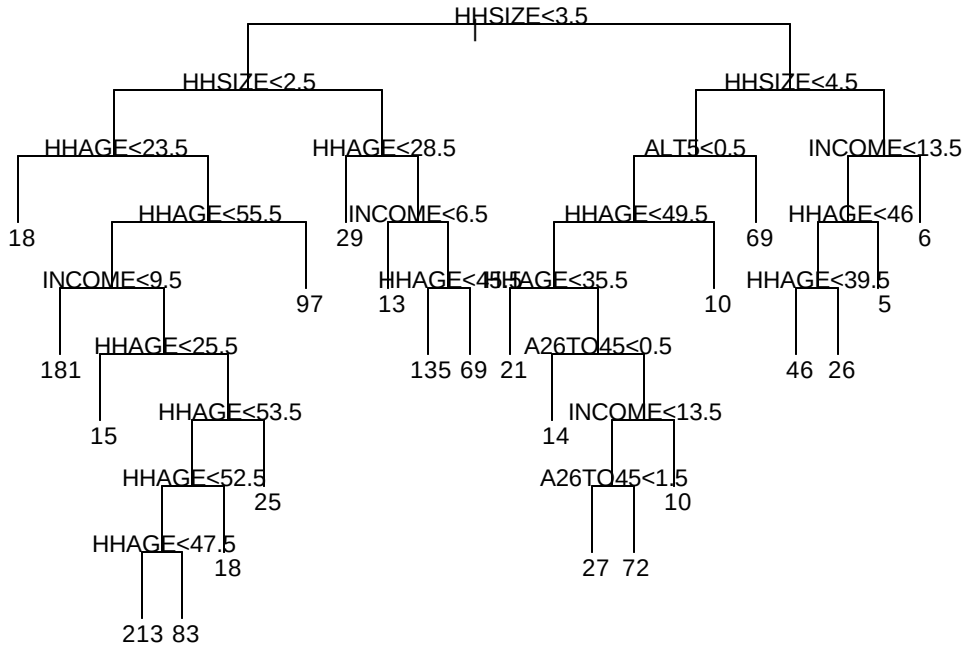
1) Classification tree for Workers = 0



2) Classification tree for Workers = 1



3) Classification tree for Workers = 2



4) Classification tree for Workers > 2