NOTE TO USERS

This reproduction is the best copy available.

A Comparison of Neural Networks and Statistical Methods in Tour-Based Travel Demand Modeling

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Applied Science (M.A.Sc.)

By

Hootan Pourkhorsand

B.Sc (Civil) 2002

Baha'i Institute for Higher Education, Iran

Department of Civil and Environmental Engineering
Carleton University
Ottawa, Ontario, Canada

The Master of Applied Science program in Civil and Environmental Engineering is a joint program with the University of Ottawa, administered by the Ottawa-Carleton Institute for Civil Engineering

August 2009

Your file Votre référence
ISBN: 978-0-494-60225-6
Our file Notre référence
ISBN: 978-0-494-60225-6

To

the Baha'i youth in Iran

who are deprived of higher education

Acknowledgments

I would like to express my sincere appreciation to my supervisor, Professor Ata M. Khan, for his invaluable guidance, support, and encouragement throughout the course of this research.

I wish to acknowledge my deepest appreciation to the Baha'i community of Canada and several individuals in this community for their kind assistance and support which was far beyond my expectation and imagination. This study could not have been done without such high quality support.

Lastly, and most importantly, I am most grateful to my dearly loved parents for their encouragement and loving support.

Abstract

Traffic congestion due to its harmful effects on economy, society, and environment is becoming a serious issue around the world. Traffic congestion is on the rise as a consequence of expansion of cities, rapid growth of population, and resulting travel demand. For system planning and traffic management, reliable predictions of demand are required. In order to improve the demand prediction process, transportation researchers are trying to enhance travel demand models by introducing new methodologies. Attempts are being made to use tour-based modeling approach as an alternative to the conventional trip-based approach. This thesis research:

- Combines the trips into tours and eliminates the insignificant stops during the main tour to enable the use of disaggregate models.
- Investigates the capability of probabilistic neural network model (PNN) with radial basis function in classification problems as an alternative to nested logit (NL) models of car ownership prediction.
- Explores the efficiency of feedforward backpropagation neural network approach versus linear regression method for the development of tour production and attraction models.

The findings reported provide new insights into the urban transportation demand modeling process.

Table of Contents

List of Tables

List of Figures

Glossary of Acronyms and Terms

ANN	Artificial Neural Network
BRT	Bus Rapid Transit
FFBP	FeedForward BackPropagation
GEV	Generalized Extreme Value
GLA	Gross Leasable Area
Ha	Hectare
HH	Household
IIA	Independent from Irrelevant Alternative
IID	Independent and Identically Distributed
LRT	Light Rail Transit
MATLAB	MATrix LABoratory
MDC	Multinomial Discrete Choice
MLP	MultiLayer Perceptrons
MLR	MultilLinear Regression
MNL	MultiNomial Logit
MSE	Mean Square Error
NCR	National Capital Region
NL	Nested Logit
OD	Origin-Destination
PDF	Probability Density Function
PNN	Probabilistic Neural Network
RBF	Radial Basis Function
RUF	Random Utility Function
SAS	Statistical Analysis Package
SPSS	Statistical Package for Social Sciences
TAZ	Traffic Analysis System
TOD	Time Of Day

List of Symbols

a	neuron output
Acc	normalized accessibility of zone
b	Bias
B	balanced total tour
C	function choice set
e	mathematical constant equal to 2.71828...
f	transfer function
H	number of households
i	traffic zone index
j	traffic zone index of the primary destination
L	maximum of the log-likelihood function
L_0	maximum of the log-likelihood function when all coefficients, except for an intercept term, are zero
ln	natural algorithm
m	index of independent variables
N	number of observed produced tours
p	inputs to the neural network
P	produced tours
R	daily household tour production rate
S_{au}	production balance factor
S_{pu}	attraction balance factor
T	free flow travel time between traffic zones
t	household type from 1 through 4
U	utility function
u	tour purpose from 1 through 5
V	deterministic part of utility function
w	connection weights of neural network
X	independent variable
Y	predicted value
$\hat{Y}$	observed value
β	regression coefficient for independent variables
β'	vector of coefficients of independent variables
η	location parameter
μ	scale parameter
ξ	stochastic part of utility function

1 Introduction

1.1 Background

In recent years, urban regions have been faced with growing traffic congestion and its subsequent environmental effects such as air pollution. In order to be better prepared to handle severe imbalances of travel demand and capacity of the road network in terms of demand and traffic management strategies and also to plan public transit facilities and services, an important step is to produce improved travel demand forecasts. With enhanced multimodal demand forecasts, the performance of a regional transportation system in handling travel demand can be studied in a reliable manner.

For many years, a four-stage demand forecasting model has been extensively used to forecast travel behavior and congestion, and predict traffic volumes in each link of the transportation network. The conventional travel demand modeling framework is based on four sequential steps of trip generation, trip distribution, modal split, and traffic assignment. The trip generation model uses various characteristics of land use and each member of a household (i.e., socio-economic data) for the study of taken trips in one day. Relevant attributes such as origin, destination, time of day, and travel mode are obtained from household surveys. In traditional practice, the observations are aggregated at the zone level. In the conventional model, the fundamental unit of analysis is a person trip, defined as the travel required from an origin location to access a destination for the purpose of performing some activity (McNally, 2000).Thus, the conventional four-step model is also called the trip-based model.

The first three steps of the four step modeling framework (i.e., trip generation, distributions and modal split) produce trip tables of person trips between origins and destinations. In the fourth step, trips and resulting traffic units are assigned to the transportation network. In other words, the first three steps provide the produced and attracted trips and their relative proportions by alternative modes between zonal pairs. In the final stage this demand is loaded onto the links of the transportation network. There are a number of criteria for assessing the reasonableness of the traffic assignment, such as the user equilibrium. The four step modeling framework is implemented by using sophisticated software such as EMME/3.

Although, conventional travel demand models have been and continue to be widely used, there is a growing concern about the quality of forecasts as a basis for decision making. It has been shown that the conventional four-step travel demand method has many limitations due to its trip-based sequential structure and lack of behavioral detail. The weaknesses and limitations of trip-based models have been noted in the literature as follows (Meyer & Miller, 2001; Pas, 1996; Dickey 1983; Domencich & McFadden, 1975;):

1- In the first step, the generation of trip is considered to be independent of the transportation supply characteristics. A fixed travel demand estimate is unable to take into account technological and service improvements and it is taken to the following steps of the travel demand forecasting process.
2- These trip generation models (also called aggregate models) are based on data that represent zonal aggregates of trips and socioeconomic conditions and the person

and household characteristics as captured by the disaggregate data are not used. Thus, the aggregate models are not capable of studying the choice processes of each person or a household in making travel decisions.

3- Most of the four-step models ignore the time-of-day dimension and only are capable of estimating either morning or afternoon peak traffic hours. But, the results for an entire day are required for estimating the full impacts of travel demand and policies in terms of emissions.

Due to the above limitations and weaknesses, another method is required to improve the accuracy of the models. The new model should be able to take into account the travel behavior for all of the households throughout the entire day and new policy decisions that affect travel decisions. For this purpose, the disaggregate modeling approaches were advanced for capturing the choice processes which the household or individual considers in making travel and activity decision.

Thus, the methodology of disaggregate models is based on the basic idea that the demand for travel is derived from people's decision to carry out different activities with regard to temporal-spatial constraints (Bowman & Ben-Akiva, 2000). For this reason, this method is called activity-based and is developed to overcome the deficiencies of conventional trip-based models.

Some of the activity-based models consider the chain of activities each member of the household is taking throughout the day. These models, called tour-based or trip-chaining models, consider individual's daily commute trips as part of the trip chain or a tour. They try to combine a number of shorter trips with different purposes into one longer trip

called tour. In this case, a tour is defined as a sequence of trips that originate at home and finally end at home, while between the start and end, they go to other activity locations such as work (Bowman, 1995; Shiftan, 1998). The tour-based analysis method is a sophisticated approach for forecasting travel behavior through considering the interrelationships between the numbers of trips, their purposes (destinations), time of day and their locations (temporal and spatial constraints).

1.2 Analysis Methods

Although a conceptual and theoretical foundation in improving travel demand modeling methodology is important, nonetheless, selecting the appropriate and practical analysis technique for calibrating the models has a significant effect on their accuracy for predicting future travel behavior and thereby, demand of regional transportation system. In other words, choosing the best analysis method corresponding to the purpose of each model has a substantial effect on the output results in terms of efficiency.

1.3 Research Objectives

The overall purpose of this research is to overcome the weaknesses and limitations of conventional four step models and improve the accuracy of the travel demand model by using the concept of activity-based modeling. Moreover, artificial neural network (ANN) was proposed as an alternative for statistical methods in travel demand forecasting. In doing so, five objectives have been defined and explained as follows:

- The first objective is to combine the trips into tours to enable models to forecast number of tours instead of individual trips. Moreover, this enables the models to

consider the person and household characteristics as well as socio-economic and demographic data at the zonal level and present a suitable framework in which travel is viewed in the context of traveler's behavior and daily activity pattern.

- The second objective of this research is to develop disaggregate models in combination with aggregate models in tour generation process. Tour production and attraction models should be capable of operating with tours (instead of trips) and determine total generated tours in each traffic zone. Moreover, the disaggregate model (tour production) should be able to consider household attributes in addition to socio-economic and land-use variables.

- The third objective is to explore the effectiveness of feedforward backpropagation neural network as an alternative method for multiple linear regression analysis for tackling the prediction problem.

- The fourth objective is to deploy the capability of probabilistic neural network model (PNN) with radial basis function in solving classification problems as an alternative to the nested logit (NL) models.

- The last objective is to compare the final results derived from the application of both methods with each other in order to explore the benefits and capabilities of neural network methods versus statistical methods for predicting traveler behavior, particularly in the context of tour generation.

Figure 1-1 shows the relationships between the proposed models. In part A of this figure, the disaggregate model for predicting household car ownership is depicted which has

been studied first. In part B, the disaggregate daily household (HH) tour production rate model and daily zonal tour attraction model of aggregate nature is shown and consequently have been developed. Then, the balance of daily tour production and daily tour attraction is of interest.

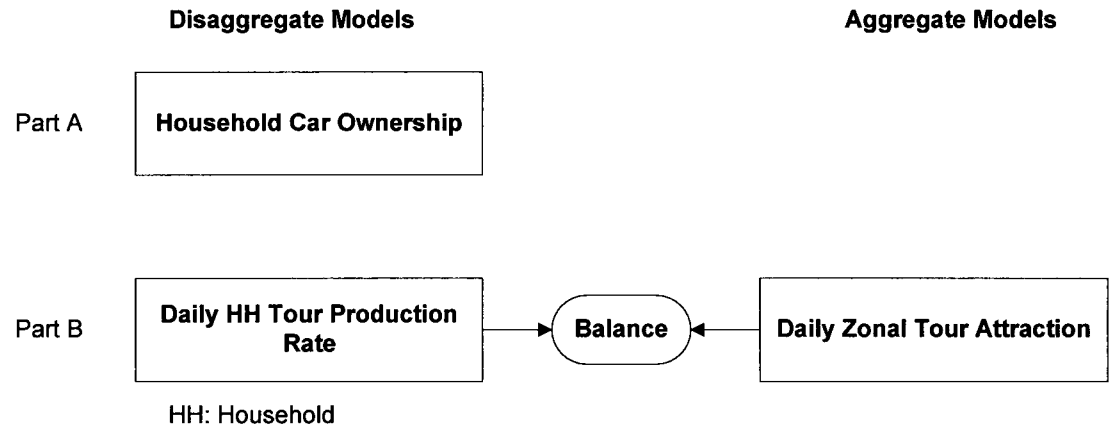


Figure 1-1: Proposed models

1.4 Research Approach

This research aims to solve classification and prediction problems which are two of the major problems in travel demand modeling. For classification problems, discrete choice methods and probabilistic neural networks (PNN) are applied to find the best choice for each household and also, multiple linear regression method and feedforward backpropagation neural network technique are deployed for solving the prediction problem. Finally, the output results of each method are compared to find the best approaches and techniques for solving those problems. However, it should be noted that it would be difficult to compare the accuracy of the techniques and to determine the best

one due to the fact that their performance is significantly related to data characteristics (Kim, 2008).

The research framework is presented as Figure 1-2. The components are shown in the diagram. As a part of the study two models have been implemented in this research: i) household car ownership model, ii) tour generation model.

First model tries to predict the choice of each household for number of cars owned from the specific choice set. Two different analysis methods for calibrating the model have been used: i) discrete choice models, ii) probabilistic neural network. The discrete choice model employed in this study is a three-level nested logit model with four alternatives. Figure 2-5 exhibits the structure of nested logit model and how the alternatives have been categorized in different nests. The utility function and choice probability of each alternative for each household are calculated based on the equations and methods explained in Section 2.4. The second calibration method is a simple probabilistic neural network with three layers. The transfer function of hidden layer is a radial basis function. In this method, output results directly indicate the car ownership choice of each household for the training data set and new (testing) data sets.

The second model employed for forecasting tour generation in each traffic zone. This model is composed of two different types of models for different tour purposes. Aggregate and disaggregate models are employed for predicting tour attraction and tour production, respectively. Again, two different analysis techniques are used for the purpose of calibration: i) multiple linear regression, ii) artificial neural network. Linear regression is popular and is a common technique in the prediction area. It has been used

for forecasting daily zonal tour attraction and household daily tour production rate as the dependent variable. Moreover, the results determine the relationship between independent variables and dependent variable and compute the related coefficients of explanatory variables. The other technique, ANN, employed in this study is a multilayer feedforward network (MLP) trained by a backpropagation algorithm. The number of hidden layers is considered to be just one and the number of neurons is set to either 20 or 100 for tour attraction models, and 100 or 200 for tour production models for different travel purposes.

The initial guess for the number of neurons in hidden layer involves selecting corresponding number of neurons in input and output layers. Then, the best number of neurons can be determined by the trial and error method in order to identify the best ANN structure. The transfer function for hidden layer is a sigmoid function followed by a linear function.

Furthermore, each data source was split into two data sets: i) training, ii) testing. Training data were used as an input data for both statistical methods and ANN models for calibration. The testing data were used to test the capability and efficiency of both models for new datasets. Thus, both methods for each model share the same data set for training and testing.

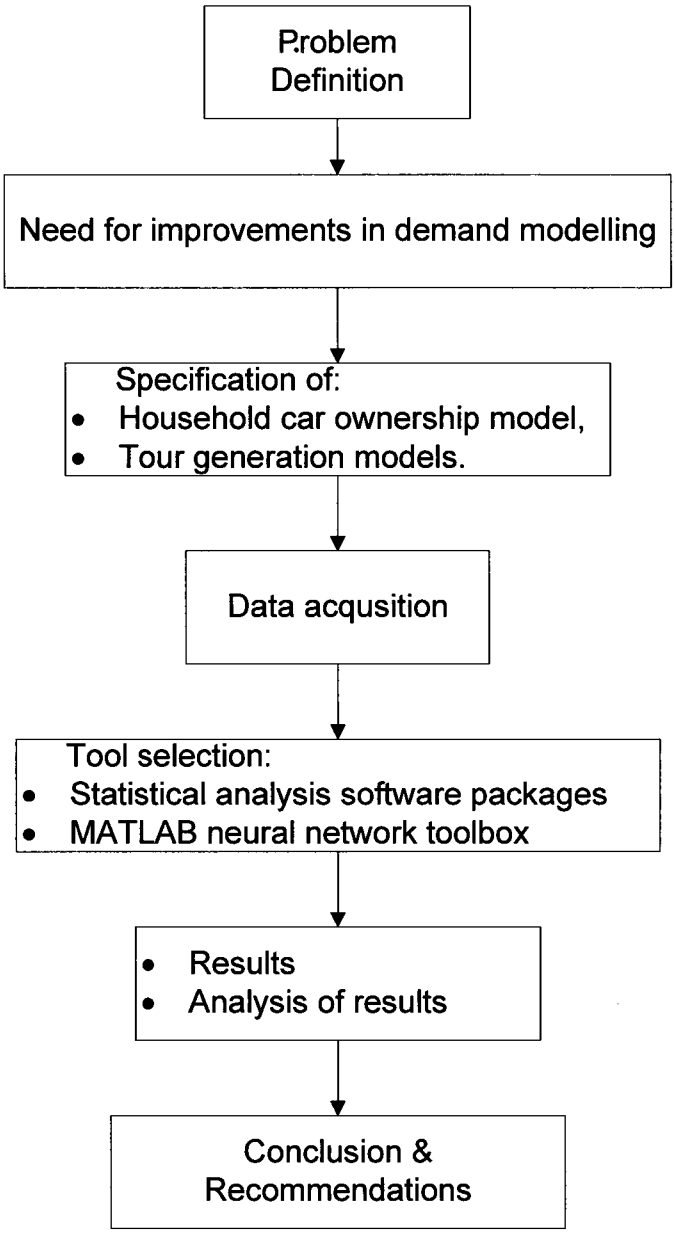


Figure 1-2: Research framework

1.5 Thesis Organization

In this chapter, the difficulties associated with conventional trip-based model and its weaknesses and limitations were introduced and a new tour-based model was advanced.

In addition, the research objectives were presented and the research approach was described. In Chapter 2, background information about tour-based travel demand model is presented. Different analysis methods are also introduced and explained in detail. Chapter 3 describes how the data were collected and prepared and also explains the selection of variables and data sets for different models. Moreover, this chapter explains the tour construction procedure. Chapter 4 covers the methodology for model development, and the scope of various models. The architecture of each model is explained and the calibration and validation process is clarified. In Chapter 5, the results of the validation process for each model are presented and finally are compared to each other. Chapter 6 presents conclusions and provides recommendations for future research.

2 Literature Review

2.1 Introduction

Transportation researchers are trying to improve the travel demand model performance by introducing new methodologies and theories such as activity-based and tour-based. Furthermore, research is required in finding calibration methods that are theoretically sound and are efficient in implementing the new models. Conventional statistical approaches on discrete choice analysis and linear regression have been employed in the past, and they have been investigated exhaustively. Recently, owing to new methods and computational techniques, alternatives to statistical methods have become available. The artificial neural network approach is considered to be a useful alternative for modeling the travel behavior and demand (Celikoglu, 2006).

2.2 Tour-Based Travel Demand Model

Activity-based modeling was introduced to cover the behavioral shortcomings of trip-based models used for many years for forecasting travel demand. As noted previously, activity-based model is a disaggregate model which considers travel decisions as a collection of activities rather than a single isolated trip. Thus, the travel is defined as a change of locations between two successive activities (Bowman & Ben-Akiva, 1996).

The tour-based model was first developed by Bowman and Ben-Akiva based on activity-based concept. Theoretically, it considers the daily activity-travel pattern as a set of tours. A tour is defined as the travel from home to one or more activity locations and back home again (Bowman & Ben-Akiva, 2000). Tours are separated into primary and secondary destinations based on activity priority. Then, activities are ranked regarding to

their purposes. Vovsha, Peterson, & Donnelly (2004) categorized the activity purposes into three main categories:

- Mandatory activities (work, university, or school),
- Maintenance activities (shopping, banking, visiting doctor, etc.), and
- Discretionary activities (social and recreational activities, eating out, etc.).

For this study, mandatory activities are divided into separated work, school and university purposes. The trip inside a tour with the highest priority activity (e.g., work) is considered as the primary destination, and other stops during the tour are designated as secondary tours. Each designated tour should indicate the following characteristics: i) primary destination, ii) time of day, iii) mode of travel for both directions.

In this research, two different analysis methods have been used for implementing the models corresponding to activity-based models. These are conventional statistical methods and artificial neural network. Figure 2-1 represents all of the common statistical methods. However, in this research, only nested logit and regression models were deployed as analysis methods. These methods are described in following sections.

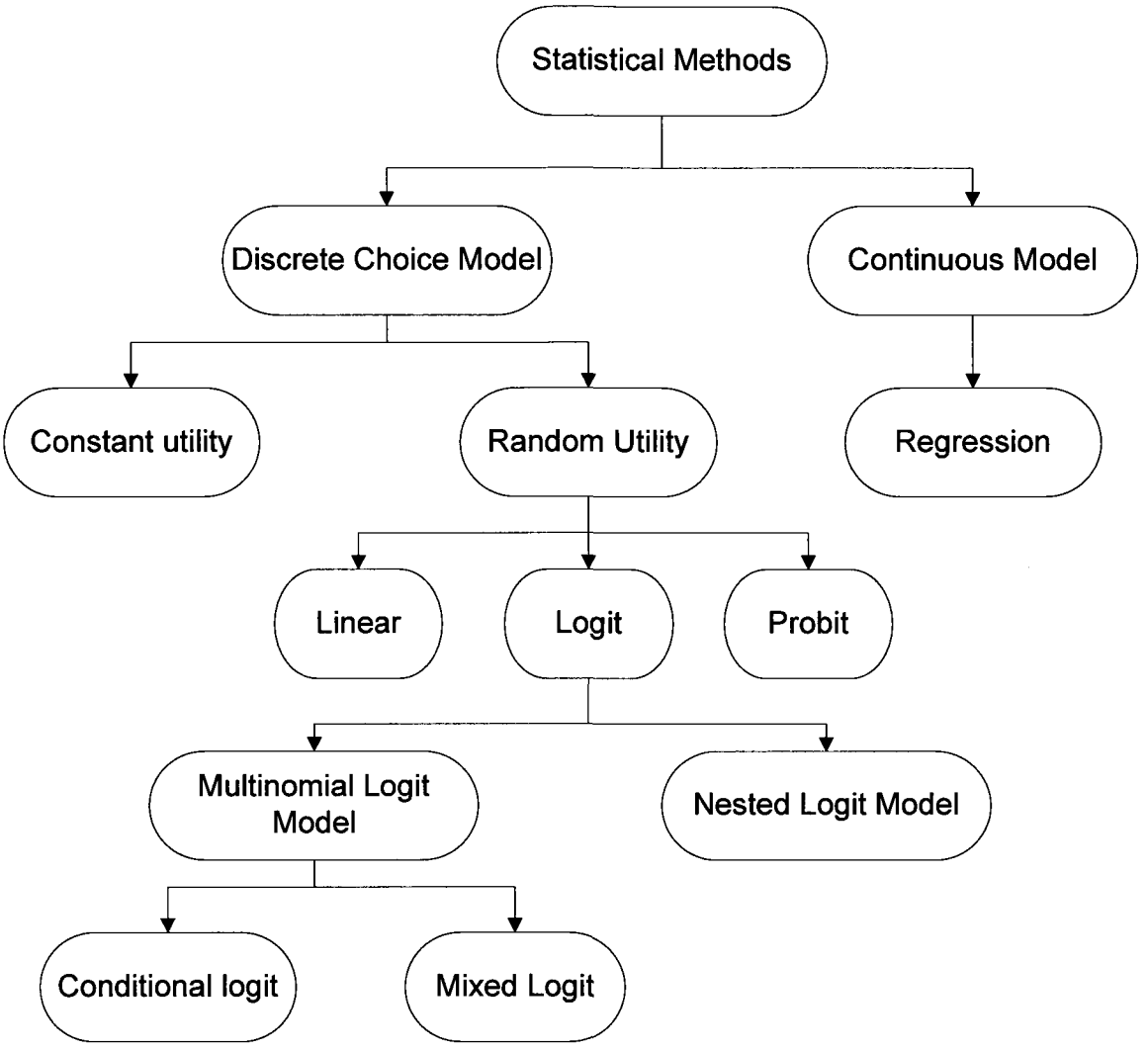


Figure 2-1: Conventional statistical methods

2.3 Artificial Neural Networks

An Artificial Neural Network (ANN) is a mathematical and computational model capable of estimating the complex relationship among observations in order to determine a relationship between dependent variable and independent or explanatory variables, including the situations when the relationship is complicated, or difficult to model. Recently, transportation researchers are exploring the use of artificial neural networks as

a new method and framework to tackle the problems in the field of transportation engineering such as travel demand forecasting, traffic control and operations, transportation planning, construction and maintenance. ANN applications in the transportation field are being used for forecasting/approximation, classification, and image processing (Kartam, Flood, & Garrett, 1997). For example, Mohammadian & Miller (2002), compare the ability of the nested logit model and the multilayer perceptron artificial neural network in terms of their applicability to the household vehicle choice problem.

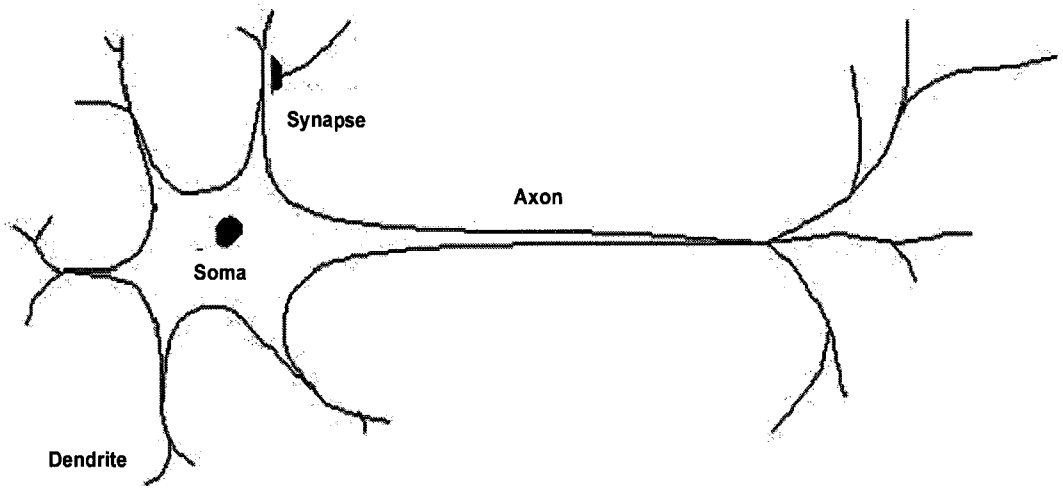
This research is exploring the advantage of using ANN versus conventional statistical methods in forecasting and classification problems.

2.3.1 Basic Components of Artificial Neural Network

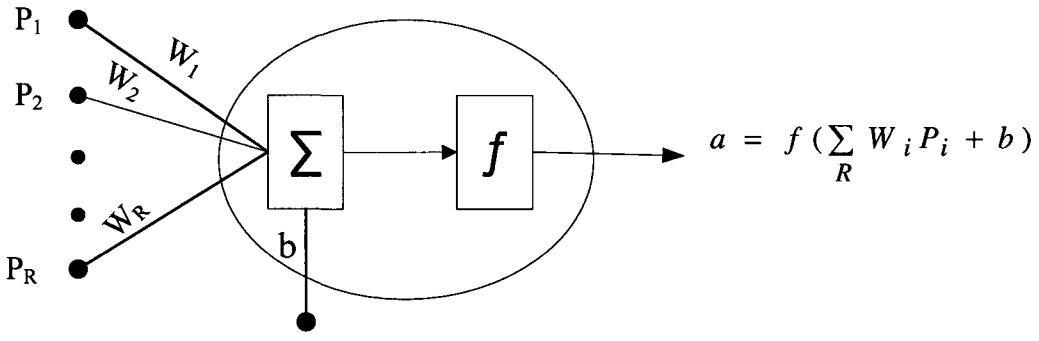
Fundamentally, the neural networks methodology stems from and is inspired by biological nervous systems in the human brain. The brain is composed of large numbers of neurons that are working in parallel and harmony to model and solve various problems.

A single biological neuron is composed of four basic components. These are: cell body (soma), axon, dendrites, and synaptic gap. Dendrites are functioning as a branching input structure and the axon is like a branching output structure. The synaptic gap is a small space between the axon of one neuron and dendrite of another one, where the signals are transmitted through them. Thus, all of the neurons are like simple processing units in which each receives input signals from other sources or neurons, sums the incoming signals, performs a nonlinear operation on the result and then fires the output result if the

total inputs exceed a certain level (Fausett, 1994). Figure 2-2 represents a simplified biological neuron and the relationship between its four components.



(a) Biological neuron



(b) Artificial neuron

Figure 2-2: Schematics of a biological and artificial neuron

Source: http://www.bordalierinstitute.com/target1.html

An artificial neural network aims to model the same process and performance of human brain, for computing and solving a wide variety of problems. In doing so, typical artificial

neural networks consist of some basic features: i) neurons (nodes), ii) connection weights, iii) transfer function iv) layer system or network structure. Figure 2-2 shows the function of a basic artificial neural network and how these elements are connected to each other. These components can be defined as follows:

- **Neurons, nodes, or cells**: A neuron is the processing unit and base element of any type of neural network. A neuron simply receives inputs (p) from other cells and adds them up; then processes them according to allocated transfer function and eventually sends an output to many other cells in other layers or gives the final output of the network.
- **Connection weights**: Each neuron is connected to other neurons by means of directed links to transmit its output. Each connection link has an associated value (w), which can be adjusted during training process. Indeed, this process enables neural networks to be trained and fit a particular input to a specific target output.
- **Transfer function**: It is specified to each cell (layer) and applied to the sum of the inputs to produce an output result. Generally, a nonlinear function such as log-sigmoid or logistic sigmoid and tan-sigmoid or hyperbolic tangent sigmoid noted in equation (2-1) and (2-2) respectively, is used; however, linear or threshold functions can be also employed in combination of nonlinear functions.

$$f(x) = \frac{1}{1+e^{-x}} \tag{2-1}$$

$$f(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{2-2}$$

Where x is the summation of incoming inputs and the allocated weights plus bias value (wp+b). The output value from log-sigmoid function is between 0 and 1, but the output result of tan-sigmoid function is between -1 and 1.

- **Bias**: It is almost like a weight; but with this difference, that it has a constant value of 1. It adds to weighted input (wp) and forms the argument of the transfer function.
- **Layer**: To make the visualization and analysis of neural network easier, neurons can be divided into various layers. Generally, two or more of the neurons can form a layer. It should be noted that the neurons in adjacent layers are connected to each other. All layers of the network typically fall into one of three categories: input, hidden, and output. In any network structure, there is only one input and one output layer. But the number of hidden layers can be one or more than one, for a given network structure. Moreover, the number of neurons in input and output layers is fixed, and equal to the number of independent and dependent variables respectively. The initial guess for the number of neurons in hidden layers can be selected corresponding to the number of neurons in input and output layers, then, it can be changed based on trial and error procedure. The basic structure of a typical feedforward multilayer neural network is shown in Figure 2-3.

2.3.2 Architecture of Neural Networks

Neurons in different layers can be connected in various ways. However, architecture of neural networks, regarding to the connection between neurons, come into two categories:

- Feedforward networks;
- Feedback (recurrent) networks.

In this research, only feedforward neural networks have been used, therefore only this type of architecture is explained at this chapter. Generally, information in feedforward neural network flows only in the forward direction or one way. This flow will be from input layer through intermediate hidden layer to output layer. In other words, the feedforward network is like an acyclic graph in which no path (connection) in the graph can go back to the starting point. Figure 2-3 illustrates the basic structure of feedforward multilayer networks or multilayer perceptrons (MLP).

Feedforward networks are used broadly in various areas, but, in this research, this structure is only used for forecasting purposes. The other important aspect of deployed neural networks in this research is that the transfer function for hidden layer is sigmoid function followed by an output layer of linear neurons. Nonlinear transfer function induces the network to learn nonlinear and linear relationships between input and output layers. The output value of sigmoid neurons is limited between -1 and +1 but the linear transfer function for output layer can take on any value.

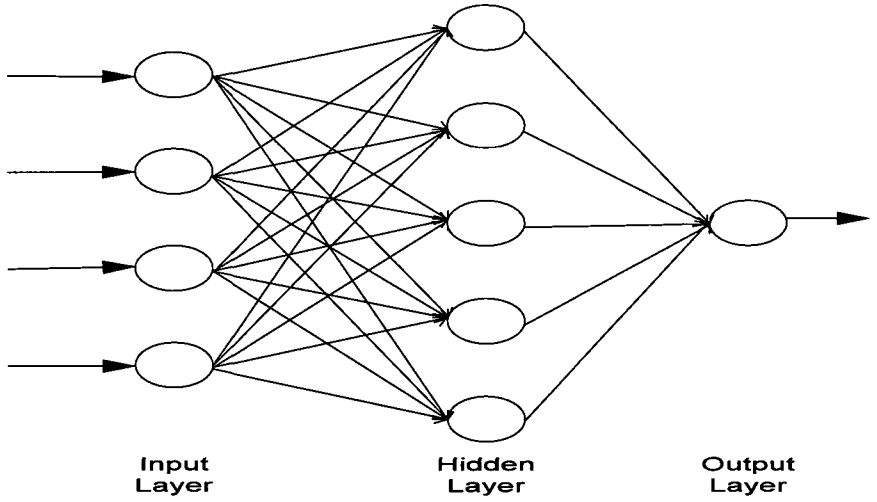


Figure 2-3: Structure of a multilayer neural network

Source: Dreyfus, 2005

2.3.3 The Training of Neural Networks

ANNs acquire their knowledge through learning or training process. In training process, the weights of connections are adjusted adaptively until a particular input leads to a specific output, and neural network fulfills its assigned task as precisely as possible. All learning methods used for neural networks can be classified into two major categories:

- Supervised Training,
- Unsupervised Training.

In supervised training, the finite values of the inputs and of the corresponding values of output (target) are provided, called examples. ANN is trained based on a comparison of the resulted output and the target until error terms between output and target values are minimized. Figure 2-4 indicates a supervised learning process in ANNs.

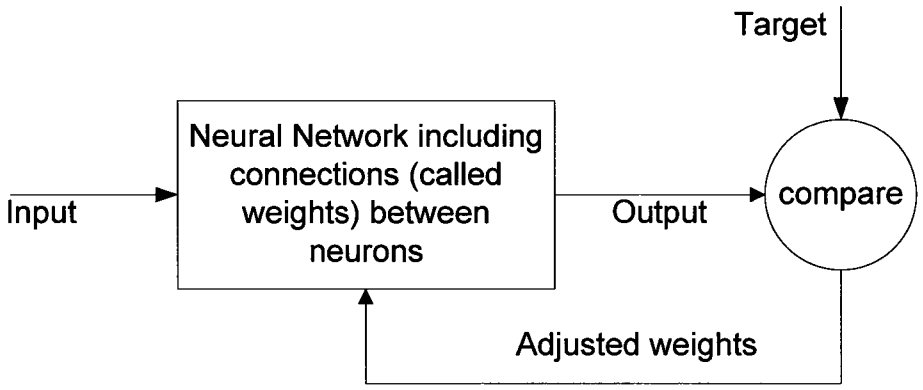


Figure 2-4: NN with supervised training

Source: Demuth & Beale, 2001

On the contrary, in unsupervised learning, the ANN is trained only by using a set of input values without providing the output values. Thus, the ANN adjusts its weights using some built-in criteria based on allocated learning algorithm. The most popular feedforward neural network with unsupervised training is the Kohonen's self-organizing maps (Dreyfus, 2005).

As mentioned earlier, the objective or training process is to minimize the mean square error on training set (output/target). In doing so, various training algorithms have been defined whereas backpropagation is the most popular one and is applied to MLP networks. The backpropagation learning rule is a supervised error-correction law in which the output error, the difference between target and output, is propagated back to the hidden layers and the connection weights are changed by a small amount until the minimum mean square error is reached or a satisfaction criterion is met. Several standard optimization techniques, such as gradient descent, conjugate gradient, and Newton methods, are generally used. More details about backpropagation algorithm can be found in (Rumelhart, Hinton, & Williams, 1986).

2.3.4 Probabilistic Neural Networks (PNN)

The Probabilistic Neural Network (PNN) is trying to define a probability density function for all defined categories, based on training process. Therefore, PNN are generally being used for function approximation. Indeed, PNN is a feedforward network with a single hidden layer that is fully connected to linear output layer (Celikoglu, 2006).

The basic probabilistic neural network architecture is composed of three layers: i) input layer, ii) one hidden layer, iii) output or competitive layer. The transfer function for hidden layer is radial basis function (RBF) which is a nonlinear function as it is shown in equation (2-3). The input variable, x, of RBF should be the distance from other points (i.e., connection weight vector) to the origin point (i.e., input vector); hence, the output of a RBF is determined based on equation (2-4).

$$radbas(x) = \exp(-x^2) \tag{2-3}$$

$$x = \exp\left[-\sum_{i=1}^{n}(p_i - w_i)^2 / 2w_{n+1}^2\right] \tag{2-4}$$

Where n is the number of input/target pairs and w_{n+1} is its standard deviation (Dreyfus, 2005).

Therefore, from equation (2-4), it is clear that whenever the input value of the radial basis function is 0 (i.e., the input value is identical to its connected weight), the output value of RBF is 1. As the distance between input values and connected weight values increases, the output values decrease, but, the minimum output value is close to 0.

Thus, the first layer of PNN computes distances from input vector (p) to the connection weight vector (w) and generates a new vector (w-p) whose element indicates how close the weights are to the input values as an input for second layer. The second layer sums up

these values for each class of inputs and produces the probabilities of each class for each observation. Finally, in the output layer, a complete transfer function picks the maximum of these probabilities and assigns 1 for that class and a 0 for the other classes (Demuth & Beale, 2001). Therefore, the chosen class by decision-maker is specified among the other classes. In spite of feedforward backpropagation networks, the number of neurons in the hidden layer is fixed, and is equal to the number of input/target vector pairs. In the output layer, the amount of neurons is equal to the number of classes of input data.

In this research, the probabilistic neural network is used for classification purpose in predicting the number of cars belonging to each household.

2.3.5 Advantages and Weaknesses of ANN

ANN has been proven successful in a number of fields due to the following reasons (Ritchie & Cheu, 1993):

- The ANN approach is nonparametric. Therefore, there is no need to rely on assumptions that the data are drawn from a given probability distribution, which is easier to use.
- ANN is a data-driven model capable of learning complex relationships from provided examples and information (i.e., as children learn to distinguish apples from oranges based on the examples of apples and oranges).
- The parallel structures of ANNs enable them to be easily implemented and offer extremely fast processing.

- ANN models can be significantly trained with less explanatory variables or the variables that are not strongly correlated to output variables in comparison with statistical methods.

In spite of these advantages that have attracted the attention of transportation researchers, ANN models also have suffered from some shortcomings which should be taken into account. They are briefly described below:

- Neural networks have been considered as a black box. Generally, ANN is a model which tries to build a model of the desired interest from the available variables based only on observations. The "black box" portion of the system contains formulas, calculations, and some interactions that are not clear to the user and it converts directly the input information into output results without providing any explanation. Thus, the correlation between variables or the effect of the model inputs on the output variables is not measured.
- Finding the optimum architecture (i.e., number of layers or neurons), training method, and transfer function may be time-consuming, yet complicated. Often times, they involve trial and error procedures. Moreover, it is difficult to choose the best trained network, because same structure in each training set can produce different results due to changing weights values.
- Neural networks often require a long time to be trained. Some of the training methods are slow for large and practical problems. However, some other algorithms, such as conjugate gradient, Leven-Marquardt, and Quasi-Newton can make the calibration process faster (Demuth & Beale, 2001).

- Another problem which may occur during neural network is called overfitting. It happens when the model fits the training set perfectly, but the model may suffer a large error value for new data set. In the other word, the model is trained well only for provided examples but it has not learned to predict accurately based on a new situation or generalizing property.

2.4 Discrete Choice Models

2.4.1 Overview

Different methods and models are being used in transportation demand analysis; however, two well-known and frequently used models in this context are aggregate models and disaggregate models (i.e., discrete choice models).

Most of the conventional travel demand models are based on the aggregate models in trip distribution and modal split steps. An aggregate model considers the trips or selected transportation modes for a group of households or individuals in a specific area such as traffic zone or district. Aggregate data are mostly acquired by computing average or sum of values of socio-economic or demographic characteristics and travel behavior of each household or individual in each area. Indeed, this method does not consider the differences between individuals' characteristics in the same area. In addition, this model is not significantly sensitive to changes in transportation policies and is not able to predict the accurate effects of this type of changes in travel demand. In the modal split step, aggregate models are costly and hard to develop. Furthermore, in comparison with disaggregate models; they are resulting in serious biases and false estimation due to their reliance on aggregate travel data rather than records of individual trips (Ben-Akiva &

Lerman, 1985). Eventually, it can be concluded that in aggregate models, the base unit of the model is a group of individuals in a specific traffic zone. Thus, it ignores the different characteristics within the group and is taking into account only the characteristics and travel behavior of the group as a whole.

Discrete choice models consider that demand is generated by the various decisions of individuals or collective decision-making units such as households in the defined area. These decisions usually involve a choice made among a finite set of alternatives (Bierlaire, 1997). Disaggregate models are capable of estimating the several possible discrete outcomes or choices, mostly mutually exclusive. An example of possible discrete choices in the context of transportation demand is when an individual wants to leave home, should make some decisions about his/her trips during the day. First, an individual should decide about his primary destination and maybe the other secondary destinations - stops- in her/his way. In addition, she/he should choose among different available modes such as drive a car; take a bus, or biking. Moreover, she/he should choose when to leave home and also, based on chosen mode, the desirable route.

According to Ben-Akiva and Lerman (1985), analyzing the choice of an individual requires the knowledge of what has been chosen and what has not been chosen. The choice procedure from a finite choice set is an internal process used by decision maker among the available information and alternative attributes to estimate attractiveness of each alternative. The choice procedure eventually, based on that process attempts to select his or her choice. To develop a disaggregate model to predict individual choices, the following components should be considered and defined.

1. Decision maker,
2. Alternatives,
3. Attributes of alternatives,
4. Decision rule.

Decision maker is an individual person or a group of persons (e.g., a household) with specific defined characteristics. Any single decision maker, through the choice process, considers a choice set containing possible options and alternatives to make the decision based on the attribute values of each choice. Decision rule is the process that decision maker uses to select his/her actual choice. Different discrete choice models are deploying various sets of decision rules. Most types of decision rules are based on the theory of utility functions which compute the attractiveness of alternatives. Characteristics of alternatives and decision maker determine the alternatives' utilities. Conceptually, the choice process is the selection of the preferred alternative that has highest utility value among the other alternatives. Therefore, the model chooses alternative j if and only if:

$$U_{ij} > U_{il} \quad \forall j \neq l,\ j \,\&\, l \in C_i . \tag{2-5}$$

Where U_{ij} is utility function for individual i and alternative j and C_i is a finite choice set. A utility function can be determined based on various methods such as random utility theory.

2.4.2 Random Utility Models

2.4.2.1 Introduction

Indeed, decision process is complicated and causes inconsistencies in choice behavior results. However, it is assumed that decision maker chooses the alternative with the greatest utility. The utility values are not known to the researcher with certainty. Thus, they are presumed as a random variable and uncertainty must be taken into account (Ben-Akiva & Lerman, 1985). Thus, U_{ij} has two parts, deterministic or systematic portion and random or stochastic portion.

$$U_{ij} = V_{ij} + \varepsilon_{ij} \tag{2-6}$$

Deterministic part, V_{ij}, indicates the relation between the choice and attributes of alternatives and decision maker. Stochastic portion, ε_{ij}, captures the randomness. In other words, ε_{ij}, shows the differences between utility and that part of the utility that the analyst estimates in the deterministic portion.

Random utility models were first described and formalized by Manski (1977). He determined four major sources of uncertainty:

1. Unobserved attributes: the attributes affecting the decision but are not considered in utility variables.
2. Unobserved taste variations: various individuals have different taste and attitude and it is so complicated to develop a model to estimate the decision of individuals.

3. Measurement errors and imperfect information: observed attributes and collected information can be imperfect and imprecise due to many reasons such as human error data collection and generation.
4. Instrumental or proxy variables: use of instrumental variables in the model cause uncertainty.

The joint density random vector for the choice set C_i is indicated as $f(\varepsilon_i) = f(\varepsilon_{i1}, \varepsilon_{i2}, \dots, \varepsilon_{iJ})$ for j alternatives. The analyst can construct probability function, P_{ij}, for decision maker i to choose alternative j as follows:

$$P_{ij} = Prob(U_{ij} > U_{il}\ \forall l \neq j\ )$$

$$= Prob(V_{ij} + \varepsilon_{ij} > V_{il} + \varepsilon_{il}\ \forall l \neq j)$$

$$= Prob(\varepsilon_{il} - \varepsilon_{ij} < V_{ij} - V_{il}\ \forall l \neq j) \quad (2\text{-}7)$$

This probability function has a cumulative distribution and shows the probability that difference in the random terms, $\varepsilon_{il} - \varepsilon_{ij}$, is below the difference of observed quantities $V_{ij} - V_{il}$. Relying on the probability density function for error terms, $f(\varepsilon_i)$, the equation (2-7) can be rewritten as follow (Train, 2003):

$$P_{ij} = \int_{\varepsilon=-\infty}^{\infty} I(\varepsilon_{il} - \varepsilon_{ij} < V_{ij} - V_{il}\ \forall l \neq i) f(\varepsilon_i) d\varepsilon_i \quad (2\text{-}8)$$

Where I(·) is an indicator function, equals to 1 if the expression in parentheses is true or alternative j is chosen and 0 otherwise. This multidimensional integral takes a closed-form for some certain specification of densities and is significantly dependent on the density function of the error terms $f(\varepsilon_i)$ which is also obtained from assumption of

distribution of the unobserved portion of utility. Therefore, varied assumptions about the distribution of error term lead to different families of choice models. In practice, two different families are being mostly used:

1. Probit Model
2. Logit Model

The normal probability unit or Probit model is depending on the assumption that distribution of unobserved portion of utility function is equal to the sum of large number of unobserved but independent quantities. By considering central limit theorem, this distribution can be normal. Although, this assumption is near to realistic situation, the evaluation of the probability function can be so complicated. In addition, it has no closed-form solution due to the underlying multidimensional integrals (Silberhorn, Boztug, & Hildebrandt, 2006). Nonetheless, in spite of its complexity, it has been used by some researchers (see Daganzo, 1979).

Logistic probability unit or logit model is most widely used in practical applications because of its ease in estimation and property of having a closed-form solution for the probability function. It is assumed that error terms are independent and Gumbel distributed.

Table 2-1 compares probit and logit models and also indicates the probability distribution function for binary probit and binary logit model with only two alternatives.

Table 2-1: A comparison between probit and logit models

	Probit	***Logit***
Type of error terms distribution	Normal	Type I Extreme value or Gumbel
Probability Density Function $f(\varepsilon)$	$\frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2\right]$ Where $\sigma \in R, \sigma > 0$	$\mu e^{-\mu(\varepsilon-\eta)}\exp\left[-e^{-\mu(\varepsilon-\eta)}\right]$ Where η is a location parameter and μ is a positive scale parameter.
P_{ij} (Binary model)	$\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{(V_{ij}-V_{il})}\exp\left[-\frac{1}{2}\left(\frac{\varepsilon}{\sigma}\right)^2\right]d\varepsilon$	$\frac{e^{\mu V_{ij}}}{e^{\mu V_{ij}}+e^{\mu V_{il}}}$

As we discussed earlier, utility function consists of two portions: i: deterministic ii: stochastic. The distribution of error terms was explained; however, the deterministic term also should be constructed as a function of the attributes of the alternatives and individuals. These attributes are collected for each individual associating with specific alternatives. Then, all these attributes are recorded in a vector x_{ij}. The following function, when K unknown explanatory variables are considered, is assumed to capture the deterministic portion of individual i for alternative j. Also, it is assumed to be linear in the explanatory variables.

$$V_{ij} = \beta_1 x_{ij1} + \beta_2 x_{ij2} + \beta_3 x_{ij3} + \cdots + \beta_K x_{ijK} = \sum_{k=1}^{K} \beta_k x_{ijk} \quad (2\text{-}9)$$

Where $\beta_1 \ldots \beta_K$ are coefficients for each explanatory variables.

Assume that β' and x_{ij} are vectors that contain coefficients and explanatory variables respectively, $\beta' = [\beta_1,\ldots,\beta_k]$ and $x_{ij} = [x_{ij1},\ldots,x_{ijk}]$, the equation (2-9) can be rewritten

$$V_{ij} = \beta' x_{ij} \quad (2\text{-}10)$$

2.4.2.2 Multinomial Logit Model

Binary choice models, where choice set contains only two alternatives, are the simplest logit models and they were introduced first. In practical applications, the choice set can consist of more than two alternatives, thus resulting in development of multinomial logit model.

The multinomial logit (MNL) model (McFadden D. , 1974) is the most widely used model. It has a simple mathematical structure and ease of estimation based on the property known as independent from irrelevant alternatives (IIA). This property is widely discussed and is controversial between researchers as further discussed in Section 2.4.2.3.

MNL model adopts the assumptions that unobserved random variables are independently and identically distributed (iid) and the probability density function (pdf) is corresponding to type I extreme value or Gumbel distribution with a location parameter $\eta=0$ and a positive scale parameter $\mu>0$. These assumptions tend to simplify the density function indicated in Table 2-1; hence, the probability function indicated in equation (2-8), considering equation (2-10), has the following closed-form.

$$P_{ij} = \frac{e^{\mu V_{ij}}}{\sum_{l \in C_i} e^{\mu V_{il}}} = \frac{e^{\mu \beta' x_{ij}}}{\sum_{l \in C_i} e^{\mu \beta' x_{il}}} \tag{2-11}$$

Where C_i is a choice set for individual i.

Different multinomial models, also, arise owing to various types of attributes of alternatives, x_{ij}. Two different regressors being used in models are:

1. Alternative-varying or choice-specific: it varies across alternatives.
2. Alternative-invariant or chooser-specific: it is constant across alternatives.

For example, in modal split step, some regressors such as cost and travel time, will vary with different choices, therefore, they are called alternative-varying regressors; whereas others, such as income or age, are constant for each mode choice and are called alternative-invariant regressors (Cameron & Trivedi, 2005).

Different multinomial logit models are implemented with different type of regressors. Commonly used methods are:

1- Conditional Logit Model: Alternative-Varying Regressors
2- Multinomial Logit: Alternative-Invariant regressors
3- Mixed Logit: Both types of regressors

Conditional Logit Model: Choice-Specific

Conditional logit models consider only the choice-specific explanatory variables. Thus, in this model, there is only one observation for each alternative. In other words, there is only one coefficient, β, for each explanatory variable for all of the alternatives in the choice set.

Consequently, the probability function of decision maker i choosing alternative j when K explanatory variables are considered can be written as follows:

$$P_{ij} = \frac{\exp\left(\sum_{k=1}^{K} \beta_k x_{ijk}\right)}{\sum_{l=1}^{J} \exp\left(\sum_{k=1}^{K} \beta_k x_{ilk}\right)} \qquad (2\text{-}12)$$

Where x_{ij} should be a choice-specific variable and should vary across the alternatives.

For using the explanatory variables in the model they should be normalized to zero. In doing so, the value for all of the variables for one of the alternatives should be equal to zero as a reference value.

Multinomial Logit Model: Chooser-Specific

Multinomial logit model considers the chooser-specific independent variables and makes the choice probabilities depending on the characteristics of the individuals only. For example in mode choice model, individual socioeconomic characteristics such as income, and age, are constant across the different travel modes.

If x_i denotes the value of the explanatory variables for individual i for all alternatives, then there is a specific coefficient (β) for every independent variable for all of the alternatives, $\beta = [\beta_1 \ldots \beta_J]$. Thus, the deterministic part for each explanatory variable is usually of the form:

$$F_j(x_i, \beta) = F_j(x_i\beta_1 \ldots x_i\beta_J)$$

Therefore, the probability function of decision maker i choosing alternative j with K explanatory variables can be:

$$P_{ij} = \frac{\exp(\sum_{k=1}^{K} \beta_{jk} x_i)}{\sum_{l=1}^{L} \exp(\sum_{k=1}^{K} \beta_{lk} x_i)} \quad (2\text{-}13)$$

Where x_i is a vector of observed variables relating to individual i, x_i is constant for all of the alternatives.

A practical approach to modeling alternative-invariant variables in multinomial logit model is a normalization of parameters to restrict probabilities sum to one. In doing so,

empirical coefficients of all independent variables for first alternative should be restricted to zero, $\beta_{1k}=0$ (Cameron & Trivedi, 2005).

Mixed Logit Model:

Mixed logit (also called random coefficients logit) is a richer model that combines the two preceding models. This model can handle both alternative-varying and alternative-invariant variables. Mixed logit models consider that the choice probability is a mixture of logits with a specified mixing distribution (Revelt & Train, 1998).

Mixed logit model allows that each individual's coefficients, β_i, differ from the population mean, β, for the chooser-specific variables. This difference leads to an additional source of error terms. By capturing individual's coefficients, $\beta_i = [\beta_{1i},\ldots,\beta_{Ki}]$, the model can be formulized as follows (Ben-Akiva & Lerman, 1985):

$$P_{ij} = \frac{e^{\beta_i x_{ij}}}{\sum_{k \in C_i} e^{\beta_i x_{ij}}} \tag{2-14}$$

As β_i is unknown, there is no closed-form solution to the integral in equation (2-8) due to many unknown variables.

Mixed logit models can be implemented as a conditional logit model. In doing so, alternative-invariant explanatory variables can be converted to alternative-varying regressors by using dummy variables.

Assume that x_i is a chooser-specific variable and it is constant across the alternatives and dx_{ijl} is a dummy variable which converts x_i to (J-1) choice-specific variables. dx_{ijl} is equal

to x_i if $j = l$ and equals to zero otherwise. It should be noted that j and l are an element of choice set.

$$dx_{ijl} = [0, \dots, x_i, \dots, 0]$$

2.4.2.3 Independence from Irrelevant Alternative Property (IIA)

An important property of multinomial logit model is the independence of irrelevant alternatives. IIA property implies that for individual i, the ratio of the probabilities of any two alternatives is unaffected, and independent from presence or characteristics of any other alternatives in the choice set (Ben-Akiva & Lerman, 1985; McFadden, 1974). This can be exhibited based on equation (2-11) as follows:

$$\frac{P_{im}}{P_{in}} = \frac{\frac{e^{\mu V_{im}}}{\Sigma_{l \in C_i} e^{\mu V_{il}}}}{\frac{e^{\mu V_{in}}}{\Sigma_{l \in C_i} e^{\mu V_{il}}}} = \frac{e^{V_{im}}}{e^{V_{in}}} \qquad (2\text{-}15)$$

The assumption that unobserved random variables are mutually independent is producing the IIA property. In many practical applications, the alternatives share the unobserved characteristics and the error terms of alternatives are correlated; hence, the IIA property imposes a limitation to the multinomial logit models. Various examples are illustrated to show this limitation which results in biased prediction and incorrect estimate, where alternatives in choice set are significantly correlated, including red buses and blue buses (McFadden, 1974) and path choice (Bierlaire, 1997). In red bus/blue bus example, it is noted that first we suppose the choice probabilities for car and bus are equal.

$$P_i(car) = P_i(bus) = \frac{1}{2}$$

Now suppose that, due to some reasons, buses are divided into two different groups, as two of the available travel mode alternatives. The buses are all sharing the same attributes except that they are painted in blue and red. Therefore, the new choice probabilities acquired by the multinomial logit model are

$$P_i(car) = P_i(red\ bus) = P_i(blue\ bus) = \frac{1}{3}$$

This situation is unrealistic, because, it is clear that commuter is willing to assume that the two bus modes are initially one alternative and the probability of choosing car and bus, for the commuter, is still $\frac{1}{2}$. As a result, the real choice probabilities are

$$P_i(car) = \frac{1}{2}\ \ \&\ \ P_i(red\ bus) = P_i(blue\ bus) = \frac{1}{4}$$

This problem arises in choice situations with significantly correlated alternatives in which their utilities share many unobserved characteristics. For example, in this research, multinomial logit model is not deploying for the purpose of forecasting household car ownership, because, all of the alternatives of this model are significantly correlated to each other.

Other models were developed to overcome this limitation and restrictive assumption (i.e., IIA), and having the solution of allowing alternatives to share unobserved characteristics and correlation between alternatives' attributes. The most widely known relaxation of the limitation of multinomial logit models is the nested logit model.

2.4.3 Nested Logit Model

2.4.3.1 Overview

Nested logit approach is widely used in the field of transportation research but also can be appropriate for other fields (Train, 2003). Nested logit model is an extension of the multinomial logit model and is implemented to allow the correlation between the error terms of utility of alternatives. The basic principle of nested logit model is based on dividing the similar alternatives of the choice set into the same nests or subsets and creating hierarchical structure of the alternatives (Ben-Akiva & Lerman, 1985; Train, 2003). Indeed, the nested logit model consists of some separated multinomial logit models and the IIA property holds within nests but not across nests.

For example, assume that four alternatives are available for urban commuters; drive alone, shared ride, bus rapid transit (BRT), and light rail transit (LRT). It is obvious that first, second, third, and fourth alternatives are correlated to each other, therefore they are violating IIA property and instead of multinomial logit model, nested logit model should be used. In this case, first pair of alternatives should be in the car nest and second pair should be in a transit nest. The two-level nest structure is indicated in Figure 2-5.

Nested logit model first derived by Ben-Akiva (1973) and then was developed by McFadden (1978, 1981) who showed that distribution of the error terms are generalized extreme value (GEM). This is a generalization of the Gumbel distribution that is used in multinomial logit models.

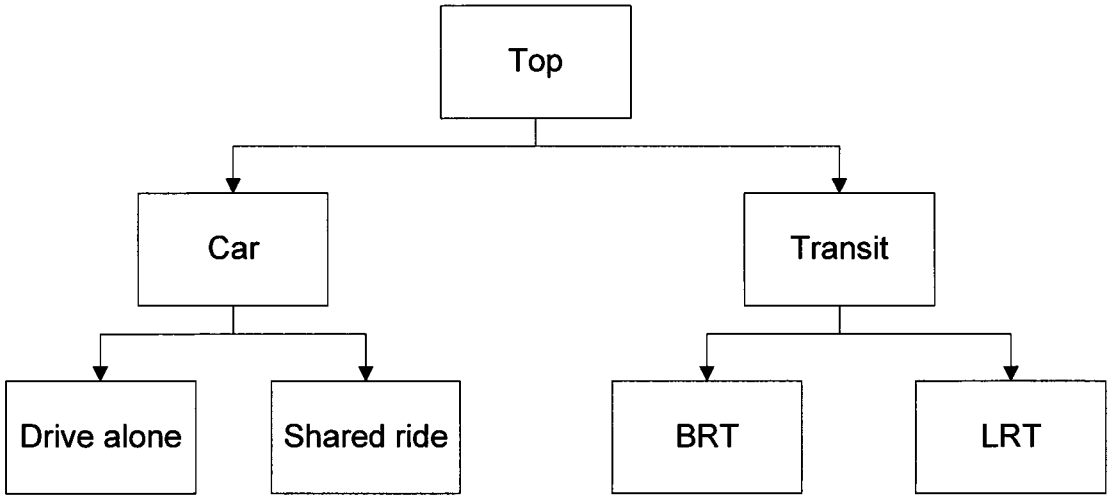


Figure 2-5: Two-level nested choice structure for four travel mode alternative

2.4.3.2 Estimation of Nested Logit Model

Consider a two level nested logit model with K choice-sets, and each choice set has J_k alternatives and the total number of m alternatives, $J_1 + \cdots + J_K = m$. The nest structure can be shown as Figure 2-6. There can also be additional levels, but, for the sake of simplicity, the results for a two-level model, presented in Figure 2-6, indicates that the choice set can split in K different nests consisting of the dependent alternatives. The utility for alternative k in nest j is then

$$U_{jk} = V_{jk} + \varepsilon_{jk} \qquad k = 1,2,\ldots,K, \qquad j=1,2,\ldots,J_k \tag{2-16}$$

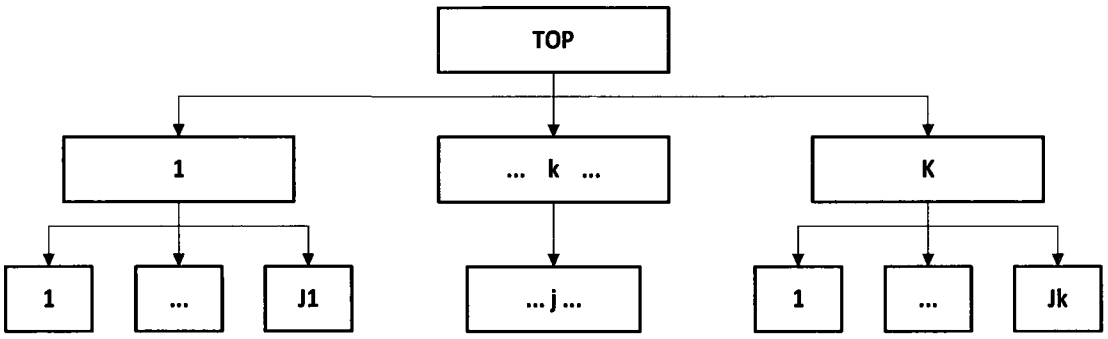


Figure 2-6: Two-level K-choice model

Source: Cameron & Trivedi, 2005

Considering the nest structure shown in Figure 2-6, the utility function of each alternative is composed of two parts, first within the nest in lower level and then across the nests in upper level. Hence, utility function can be rewritten as follows:

$$U_{jk} = U_{C_k} + U_{j|k} = \left(V_{C_k} + V_{j|k}\right) + \left(\varepsilon_{C_k} + \varepsilon_{j|k}\right) \quad (2\text{-}17)$$

Where C_K is kth choice set and j|k shows alternative j within choice set k. In more than two-level nest structures j|k is always an option within nest k in its upper node.

Mc Fadden (1978, 1981) assumed that error terms $\varepsilon_{j|k}$ and ε_{C_k} are independent of each other. Error terms $\varepsilon_{j|k}$, the alternatives in the same nest, are Gumbel distributed, with positive scale parameter σ_k and location parameter equals to zero. ε_{C_k} is also Gumbel distributed with scale parameter μ_k and location parameter equal to zero. It should be taken into account that each nest at each level has a different σ_k and μ_k. Furthermore, the scale parameters μ_k and σ_k are unknown to analysts, and it is impossible to estimate them separately. In fact, only their ratio can be identified from the data as follows (Ben-Akiva & Lerman, 1985):

$$\frac{\mu_k}{\sigma_k} = \sqrt{1 - corr(\varepsilon_{jk}, \varepsilon_{jl})} \quad (2\text{-}18)$$

It is clear that

$$0 \leq \frac{\mu}{\sigma_k} \leq 1 \quad (2\text{-}19)$$

When $\varepsilon_{jk}, \varepsilon_{jl}$ are independent then, $\frac{\mu}{\sigma_k} = 1$ and the nested model converts to a multinomial logit model.

In addition, due to the hierarchal structure of nested models, there is a virtual component in utility function for each nest at each level, called inclusive value or composite utility. Inclusive value of each nest includes the expected value of the maximum utility for the alternatives within that nest. The composite utility for nest C_k is defines as

$$IV'_{C_K} = V_{C_k} + \frac{1}{\sigma_k} ln \sum_{j \in C_k} e^{\sigma_k V_{j|k}} \quad (2\text{-}20)$$

Where V_{C_k} is the portion of the utility for all alternatives in the nest C_k. It should be considered that V_{C_k} is shared among them and for all alternatives at second level is equal to zero. In addition, there is no inclusive value at level 1 (i.e. $IV'_{C_K} = 0$).

The probability of choosing alternative j within the choice set k is dependent on probability of choosing nest k among K choice sets and probability of choosing alternative j within choice set k. Therefore, it can be shown as follows (Cameron & Trivedi, 2005):

$$P_{kj} = P_k \times P_{j|k} \quad (2\text{-}21)$$

P_k and $P_{j/k}$ can be determined according to the following equations:

$$P_k = \frac{e^{\mu_l IV'_{C_l}}}{\sum_{k=1}^{K} e^{\mu_k IV'_{C_k}}} \tag{2-22}$$

Where l and k represent the nests on upper level (in our case, two-level nested logit, it is level 2).

$$P_{j/k} = \frac{e^{\sigma_k V_{n|k}}}{\sum_{j=1}^{J_k} e^{\sigma_k V_{j|k}}} \tag{2-23}$$

Where j and n denote the alternatives on lower level (i.e. level 1).

For solving the equations (2-22) and (2-23) we need to know scale parameters σ_k and μ_k and also the deterministic utility component, $V_{j|k}$. Corresponding to equation (2-10), $V_{j|k}$ is determined by estimating the coefficients of explanatory variables, β_j. The parameters μ and σ_k are unknown to analysts, and it is impossible to estimate them separately. In fact, only their ratio can be identified from the data. A common normalization would be to restrict one of the scale parameters to 1 (Ben-Akiva & Lerman, 1985). If the scale parameter μ is constrained to 1, the model will be normalized from the top and if one of the parameters σ_k is restricted to 1, that model is normalized from the bottom. It is discussed that nested models normalized from the bottom may lead to simplified equation (Bierlaire, 1997).

In this research, the software package **SAS** (SAS 9.1.3) has been used to estimate a nested logit model. There are several points in model estimation to take into account. The critical point is that SAS generates alternative-specific coefficients (Heiss, 2002).

The other important point is how to use alternative-invariant variables in the model. In practice, it is best to convert chooser-specific variables to alternative-specific variables. In such cases, one chooser-specific variable is transformed into the (K-1) alternative-specific variables, where K is the number of alternatives in the choice set and the one reference variable which should be restricted to zero. For example, let x_i be a I×1 vector and a chooser-specific variable. Then, define x_{ij} to be a (I×K)×1 vector with zeros for all of the cells except the kth cell is equal to x_i, where I is the number of individuals i = [1,..., I] and K is the number of alternatives k =[1,...,K]. In addition, one of the x_{ij} should be restricted to zero, for one of the alternatives.

2.4.3.3 Model development

Two approaches are most often used to find coefficients and to develop logit models: i) maximum likelihood approach ii) least square approach.

In the maximum likelihood approach, first the data on the calibration sample is applied into a maximum likelihood estimation package for the logit model. The maximum likelihood approach iteratively tries to find the set of coefficients that produced the best fit to the observed pattern of choices in the input data. The process of iteration of finding coefficients continues until reaching a specified convergence criterion, or the estimation meets the defined maximum number of convergences. Maximum likelihood approach has been discussed in detail in (Ben-Akiva & Lerman, 1985).

Least square approach finds the coefficients that minimize the sum of the squared errors between the observed value and a value given by the model. This method is generally

used in linear regression models. Most of the software packages including SAS, developed logit model based on maximum likelihood approach.

2.4.3.4 Goodness of Fit Measures

Likelihood ratio index is the same as R squared in the linear regression model. This index measures how well a model predicts the dependent variable and measures the goodness of fit of the model. McFadden (1974) suggests the following equation to determine a likelihood ratio index.

$$R^2 = 1 - \frac{\ln L}{\ln L_0} \tag{2-24}$$

Where L is the maximum of the log-likelihood function and L_0 is the maximum of the log-likelihood function when all coefficients, except for an intercept term, are zero. McFadden likelihood ratio index is bounded by 0 and 1.

Estrella (1998) suggests the following requirements for a suitable goodness of fit measure in discrete choice modeling.

- The measure must take values in the [0,1] range, where 0 represents no fit and 1 corresponds to perfect fit.
- The measure should be directly related to the valid test statistic for the significance of all slope coefficients.
- The derivative of the measure with respect to the test statistic should comply with corresponding derivatives in a linear regression.

There are also some other equations for finding the goodness of fit measures that are summarized in Table 2-2.

Table 2-2: Various measures of goodness of fit

Measure	R^2 Equations
McFadden	$1 - \frac{\ln L}{\ln L_0}$
Estrella	$1 - (\frac{\ln L}{\ln L_0})^{-(\frac{2}{N})\ln L_0}$
Adjusted Estrella	$1 - [\frac{(\ln L - K)}{\ln L_0}]^{-(\frac{2}{N})\ln L_0}$
Cragg-Uhler 1	$1 - (\frac{L_0}{L})^{\frac{2}{N}}$
Cragg-Uhler 2	$\frac{1 - (L_0 - L)^{\frac{2}{N}}}{1 - L_0^{\frac{2}{N}}}$
Aldrich-Nelson	$\frac{2(\ln L - \ln L_0)}{2(\ln L - \ln L_0 + N)}$
Veall-Zimmermann	$\frac{2(\ln L - \ln L_0)}{2(\ln L - \ln L_0 + N)} \times \frac{2\ln L_0 - N}{2\ln L_0}$

Where N is the number of observations used, and K represents the number of estimated parameters.

2.4.3.5 Variable coefficients and t-statistics

After calibrating the model, the important point is to choose the significant variables. This can be explored through an examination of their coefficients regarding two major factors: the sign of coefficients and the t-statistic value. The sign of a coefficient should reflect the expected influence of explanatory variables on the dependent variable based on judgment and experience.

Ortuzar and Willumsen (1994) suggest that the current practice is to include a relevant variable with the correct sign even if it fails any significance test. Thus, variables with the correct sign should be included in the model while those with the wrong sign should be neglected as much as possible. The reason that lack of significance of t-test value is not too important is that it can simply be caused by lack of sufficient data in the observed data set.

Generally, the t-test value of a coefficient is determined by dividing the value of a coefficient with the standard error of the estimate. The t-statistic value of a coefficient is a helpful measurement for choosing whether the corresponding variable that explains the dependent variable properly or not.

The significance level of t-statistic is related to the defined confidence interval. A t-statistic of more than 1.65 or less than -1.65 indicates significance of more than 95 percent confidence interval. A test of greater than 2.3 or less than -2.3 shows significance of more than 99 percent confidence interval. The confidence interval in this research is always set to 95 percent.

2.5 Linear Regression Model

2.5.1 Overview

Multiple linear regression (MLR) models have been commonly and widely used in transportation research because of their effectiveness and simplicity. For many years, linear regression was the only method for predicting number of generated trips, either attracted or produced trips.

Multiple regression models aim at predicting dependent variable (Y) on the basis of various independent variables (X_1, X_2, X_3, ..., X_n). In addition, regression equation shows the relationships between dependent variable and explanatory variables. Most of the regression models are being trained until the value of mean squared error (MSE) is minimized. Generally, MSE measures the average of the square of the error. The error is the amount by which the observed dependent variable differs from the obtained dependent variable ($Y - \hat{Y}$). Equation (2-25) shows a simple multiple regression equation.

$$\hat{Y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n \quad \textbf{(2-25)}$$

Where
β_0 = intercept value,
$\beta_1 \ldots, \beta_n$ = coefficients of explanatory variables,
$X_1 \ldots, X_n$ = explanatory variables,
$\hat{Y}$ = predicted value.

The other important issue in improving the efficiency of regression model is choosing the appropriate predictor variables which enable the model to predict the independent variable precisely. In this study, all of the explanatory variables were selected from OD survey dataset in both aggregate and disaggregate levels.

All linear regression models and analysis were implemented and completed in SPSS 16.0, a statistical analysis software package. The output results clearly show the regression equation components such as regression coefficients for each variable and make it easy to find the best variables predicting the daily household tour production rate and daily zonal tour attraction.

2.5.2 Regression Components

The following components of a linear regression are important for interpretation of regression and exploring its efficiency.

Regression coefficients: first in regression equation the coefficients of explanatory variables are unknown. Then, these unknown coefficients (β_n) are determined based on least squares method. The value and magnitude of coefficients show the effects of explanatory variables on dependent variable.

Intercept: it is the predicted value for dependent variable when all of the explanatory variables are zero. In this study, it is assumed that intercept value for all of the regression models is zero.

t-statistic value: the t-statistic value is calculated by SPSS for each explanatory variable. Based on this value, it can be realized that a variable is statistically significant to predict the dependent variable or not.

Probability value: p-value is calculated based on t-test value and it is also used to find that if a variable is statistically significant or not. In this research the confidence interval is assumed to be 95%, therefore a p-value less than 0.05 is considered significant.

2.5.3 Performance Indices

The performance indices are measurements that explain the efficiency of the model and can be used for model validation and comparison. For regression models two measures were considered: Goodness of fit measurement (R^2) and Mean Square Errors (MSE).

Goodness of fit measurement (R^2): the R^2 is a statistical measure of how well the regression line approximates the real data points and indicates the capability of model to predict the dependent variable based on explanatory variables. The values vary from 0 to 1; an R^2 of 1 shows that regression line perfectly fits the data.

Mean Square Errors (MSE): MSE is the root mean square between the observed dependent variable and output results of the model (Zhou, Lu, & Xu, 2007).

$$MSE = \frac{\sum_{n=1}^{N}(Y_n - \hat{Y}_n)^2}{N} \quad \textbf{(2-26)}$$

Where Y_n denotes the observed values and $\hat{Y}_n$ denotes the values predicted from the model by using a data sample with N observations.

3 Data Collection and Compilation

3.1 Surveyed Data Description

The National Capital Area (Canada) Origin-Destination (O-D) survey completed during the fall of 2005 is used as a major source of trip characteristic and trip pattern data for the developed models in this research. Other variables such as travel time between traffic zones, demographic and employment data were collected from the Census of Canada and other sources.

3.1.1 Study Area

The study area is chosen to be the same as survey area, presented in Figure 3-1. This area was split up into 26 urban and rural districts and 556 traffic analysis zones (TAZ). In total 23,868 randomly selected households were interviewed, representing about 5% sample of households in each of 26 districts. The O-D survey captures the information on trips made by persons 11 years of age or older, on average weekday for each of the interviewed households. The collected data was deployed for statistical methods and artificial neural network (ANN) approach, providing a comprehensive and rich data set for each model.

Table 3-1: O-D survey primary data collected (TRANS Committee, December 2006)

Household data	location, size, number of vehicles and dwelling type, etc;
Person data	age, gender, worker/student status, occupation, and place of work/school, etc;
Trip data	origin, destination, purpose, mode of travel, and departure time, etc;

The survey collected three major categories of data in disaggregate-level for each household (Table 3-1). Other requisite variables are computed based on these variables. The survey results were expanded to represent some other variables such as population and employment data in disaggregate and aggregate level. By using the new concept of tour-based models which considers individual's daily commute trips as part of the trip chain and tries to combine a number of shorter trips with different purposes into a tour, trips were combined into closed tours starting from home for modeling purposes (TRANS Committee, December 2006).

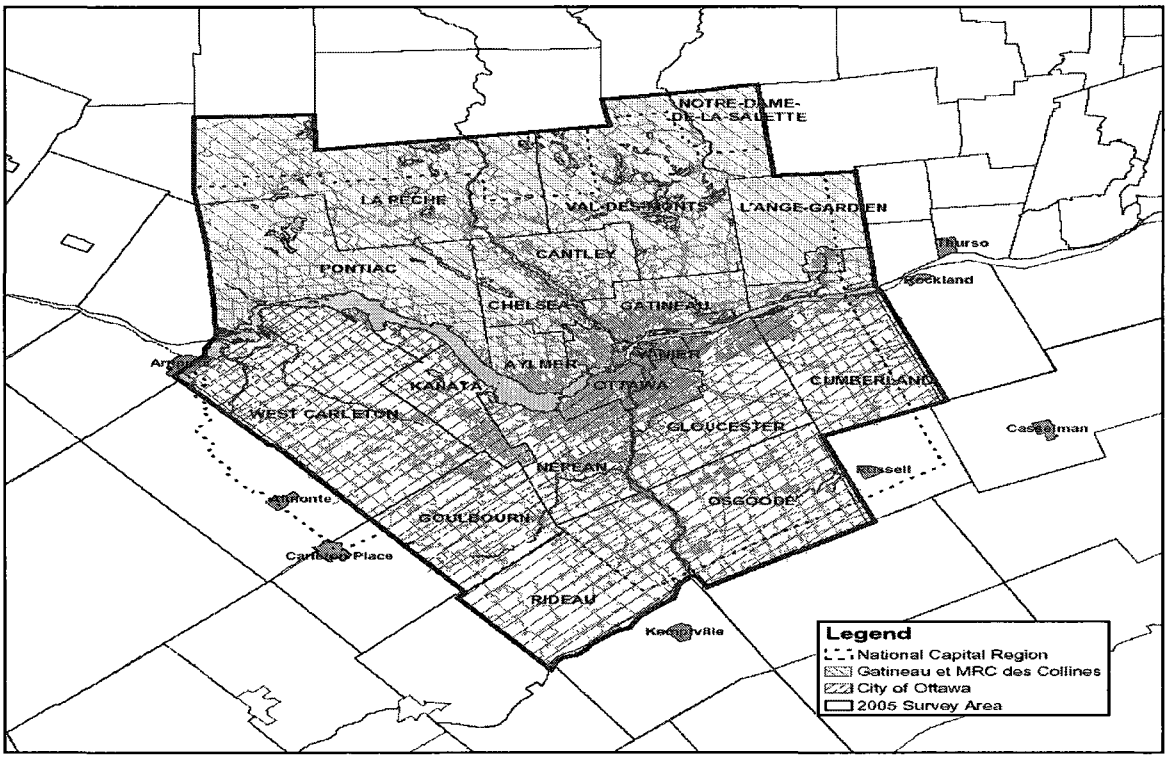


Figure 3-1: Study area

3.2 Data Preparation and Compilation

For achieving the objective of developing a tour-based model, the data should be classified into two levels:

- Disaggregate-level data
- Aggregate-level data

Disaggregate-level data is prepared at the household level, presenting the characteristics and variables of each household in different TAZ. Aggregate-level data characterize various variables for each 556 TAZ. Variables from both levels are required to initiate and calibrate the models based on tour-based context. This makes the models more sensitive to the changes in transportation policies and land use variables and makes the model capable of estimating travel demand and pattern more precisely.

New variables and data are produced by combining primary data shown in Table 3-1 and some variables from other sources at the aggregate-level for traffic zones. These variables were selected and defined regarding each model approach and are common between the statistical and ANN models. These variables are mostly selected from following sources.

Population: population data for each traffic zone was defined by various age groups (0-4 years, 5-14 years, 15-24 years, 25-44 years, 45-64 years, >64 years). The household size was defined into six classes (1, 2, 3, 4, 5, 6+) and two dwelling types were applied, apartments and detached houses and recoded into 0 and 1, respectively.

Household characteristics: The number of workers in each household is reported by O-D survey results. In addition, the number of non-workers in a household can be calculated

on the basis of these two variables. Households are also divided into two dwelling types: apartment and detached houses.

0. Apartment (i.e., row / townhouse and apartment)
1. Detached.(i.e., single and semi detached, cottage)

Employment: Total employment and labor force were determined based on the aggregate results of 2005 O-D survey. Number of workers per household were computed and categorized as follows: no worker, 1 worker, 2 workers, and 3 or more workers and recoded as 0, 1, 2, and 3, respectively. In addition, this variable in combination with non-worker (size minus worker) and dwelling type establish the distribution of household in every TAZ. In addition, employment by place of work is considered to have a significant connection with the trip attraction of each zone and identification of stop in each direction of tour. The following categories of employment type were defined: public and private office, retail, service, education, health, and industrial (TRANS Committee, December 2006).

Land-use and Density: Land-use variables are also significantly related to modeling tour generation. Land-use variables have been calculated at different levels of geography, traffic zone and district. These variables were as follows:

- Share of detached dwelling houses per traffic zone
- Population density (total population per area unit)
- Employment density (total employment per area unit)
- Shopping Gross Leasable Area (GLA)

Income: Household income could be a strong factor in modeling framework of forecasting travel behavior. The percentage of low-income households per zone was used as a practical variable to measure the impact of household income on number of generated trips by each household. The low-income threshold has been set for different household sizes according to Table 3-2 (TRANS Committee, December 2006).

Table 3-2: Household low-income threshold definition

HH Size	Min income	HH Size	Min income
1	$18,371	5	$38,646
2	$22,964	6	$42,719
3	$28,560	7	$46,793
4	$34,572		

Thus, the percentage of low-income households for traffic zones was established based on the definition in Table 3-2.

School and University related variables: trips to both school and university were considered the same but in this modeling study, school trips were split into two categories, school and university, based on the student's age. School purpose was considered for students' tours between 11 to 17 years and university purpose for students older than 18 years.

Accessibility Measurement: Accessibility measure for each zone proved to be an important determinant of travel behavior and was considered to be significant in tour production and attraction models. Accessibility measure indicates the effort of a

commuter to overcome the spatial and timing separation between zones and it can be calculated with the following equation (Allen, D.Liu, & S.Singer, 1993).

$$Acc_i = \frac{1}{n-1}\sum_{\substack{j=1 \\ j\neq i}}^{n} T_{ij} \quad (3\text{-}1)$$

Where Acc_i is the normalized accessibility of zone i; T_{ij} is the free flow travel time between traffic zones i and j; and n is the total number of traffic zones.

It is clear that a smaller value for Acc_i indicates better accessibility. Therefore, households located in the central business district have the best accessibility and those located on the boundary of the study area have the worst accessibility.

3.2.1 Description of Selected Variables and Data Sets

In this study, two models are of interest, household car ownership choice and tour generation. Tour generation has two major categories, tour production and tour attraction where each one of them has 5 separate submodels. Thus, 11 models have been totally implemented and calibrated by statistical methods and ANN.

Each analysis type (i.e., NL vs. PNN or LR vs. ANN) shares the same variables and data set. Moreover, each data set was split into two groups: training and testing. Training data set consists of $\frac{2}{3}$ of original data set and testing (validation) data set accounts for the remaining $\frac{1}{3}$ of the data set with selected variables.

Primarily, training data were deployed to calibrate both the statistical and ANN models for classification and estimation approaches. Afterwards, to determine the accuracy and efficiency of the calibrated model to perform its task, the validation data were used. Thus,

the capability of the models can be explored by comparing the output results and observed data.

It is also important to know how to use aggregate variables in disaggregate models. Household car ownership and daily household tour production rate models are using disaggregate level data for each household, whereas most of the selected variables for those models are at the aggregate level (i.e., zonal or district). Thus, those aggregate variables should be allocated to each household within that traffic zone. For example, zonal population density is an aggregate variable which takes the average value for the traffic zone, and every household in the same traffic zone has the same zone population density. MATLAB program was used to organize and prepare the variables in the input data sets. All of the aggregate and disaggregate variables used in this research are shown in Table 3-3. All models share some of these variables.

Furthermore, for the preparation of input data for nested logit model, it should be noted that chooser-specific variables have to be converted into choice-specific variables due to the inability of the statistical software package (SAS) to recognize alternative-specific variables. More details are provided in 2.4.3.2.

Table 3-3: Variables type

Type	Variable Name
Disaggregate	number of workers in the HH/ number of non-workers in combination with a different number of workers/ type of dwelling
Aggregate	zonal population density/ zonal % of detached dwelling type/ zonal % of low income HH/ shopping GLA/ district population density/ accessibility/ total employment/ total population/ university enrollment/ school enrollment

3.3 Tour Construction

The advantage of developed models is that they are based on activity-based concept. For their development, the first step is to combine a number of individual's daily commute trips with different purposes into one longer trip called tour. By this definition, a tour is always originating at home and goes to other activity locations such as work and eventually ends up at home.

According to Table 3-1, the household-id, origin (home), destination, purpose, departure time, and finally mode were collected for each person trip in the household. All of the trips in one household can produce at least one closed-cycle of trips. In fact, the concept of directed cycle graph was used to produce a tour for households. In this graph, each vertex indicates destination of preceding trip or origin of succeeding trip, and edges indicate trips between two traffic zones. Beginning and finishing points of a cycle are always home, and the other points (destinations) are treated as stops. All of the stops are ranked based on the activity purpose, and finally one of them is selected as the primary destination of the tour. The primary destination breaks the tour into two directions called In-direction and Out-direction. Out-direction shows the path from home to primary destination, and In-direction indicates the path from primary destination to home.

The complete observed tour may be composed of bunch of stops in both directions. In addition, another tour can be started and ended at primary destination called sub-tour. All of these components are presented in Figure 3-2.

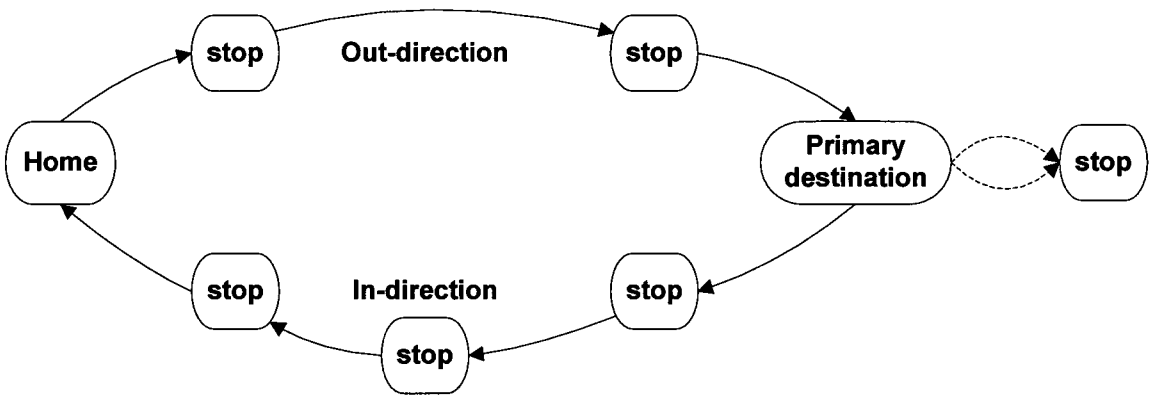


Figure 3-2: Full observed tour

To simplify the tour model (for commuters), it is assumed that commuters have only one stop in each direction, therefore, the extra stops were eliminated based on the stop rankings which is defined based on the activity purpose on that zone. Moreover, the sub-tour is taken out from the model to reduce the number of stops. The defined tour structure is shown in Figure 3-3. Because, the data show the trip pattern of households in only one specific week day, this adjustment allows for better and consistent modeling of household's trips. The insignificant trips which are not taken regularly can be handled separately. This enhances the models for better estimation of future tours. The following data can be determined for each tour regarding the simplified tour structure shown in Figure 3-3:

Household-id, origin, primary destination, purpose, travel time and mode of trip in both directions, and if the commuter had any stops either before reaching to primary destination or on her/his way back home, these are all defined variables for a tour.

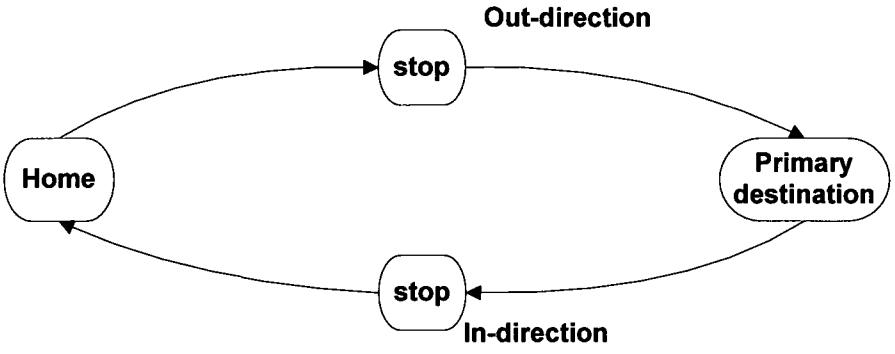


Figure 3-3: Simplified defined tour

The process of tour construction, taking into account the graph context and other assumptions is exhibited in Figure 3-4. This flowchart was implemented in MATLAB programming language to combine the trips into tours in a practical manner.

The following assumptions were made for forming tours:

- Purpose categories used are home, work, university, school, maintenance, and discretionary. If a traffic zone is traveled for different purposes more than once in the same tour, the purpose of all trips belonging to that tour is considered the purpose with higher rank as compared to others.
- The program doesn't consider a tour when all of its trips are within the same zone. In other words, only the trips between different zones are taken into account for tour construction and intra-zonal trips are not considered as a tour. These can be handled separately.

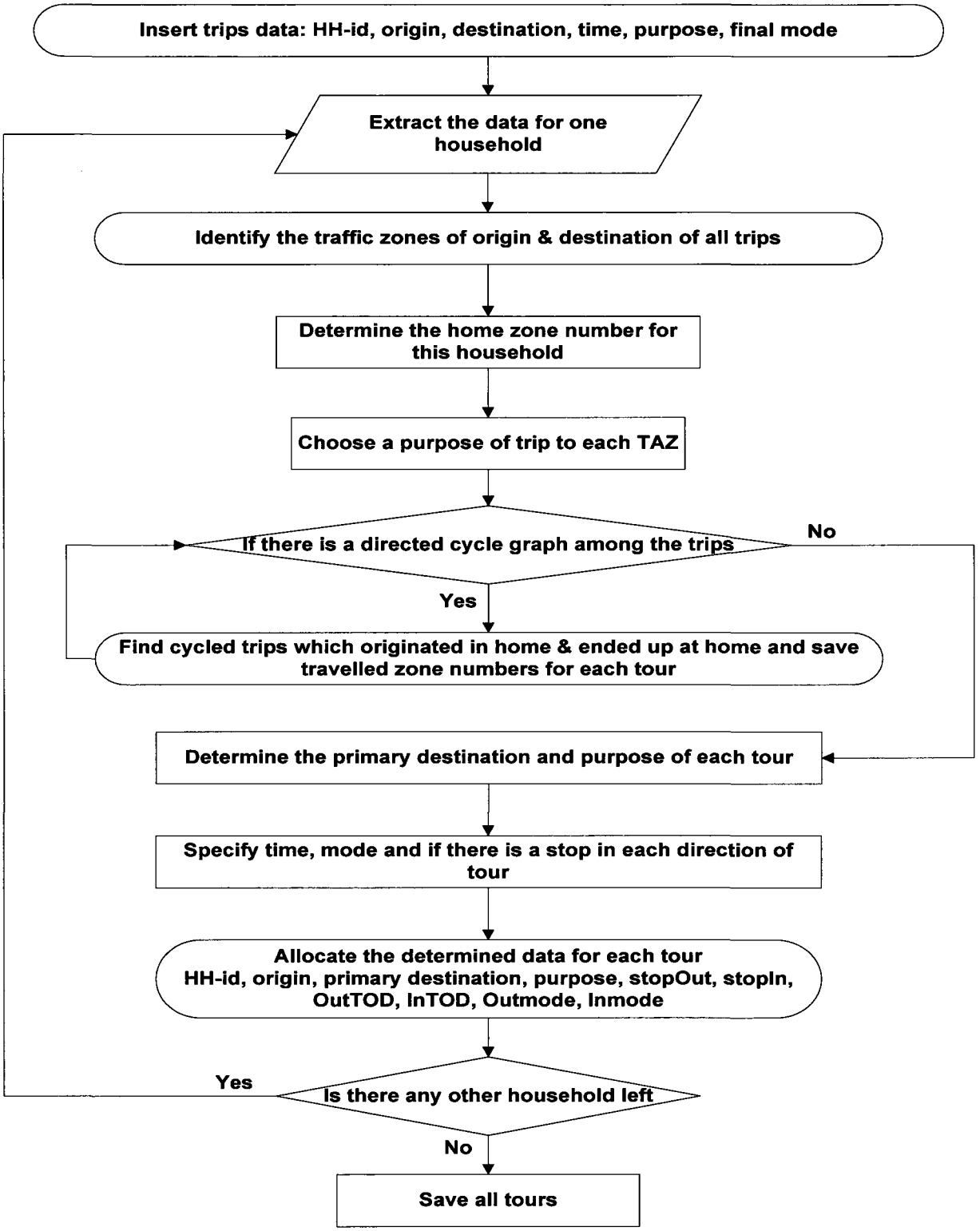


Figure 3-4: Tour construction Flowchart

- The primary destination of the tour is selected on the basis of ranking of the purpose of the trips. The first traffic zone with highest rank among the other destinations is chosen as a primary destination of the tour. Suppose that a person goes from home to bank and then to work and again goes to another location to work, and on her/his way back home finally goes for shopping. The program takes the first work's location as a primary destination.
- The purpose of tour is considered to be the same as the purpose of trip to the primary destination.
- The program only considers one stop in each direction, and eliminates other stops. If there were other stops between home and primary destination, it would be determined that this tour has a stop in its out-direction and that stop would be the first destination with highest rank. The same assumption was made for in-direction as well.
- The time of day period for out-direction is considered the same as the time of last trip which ended at the primary destination. Also, the time of day period for in-direction is considered the same as final trip of a tour that ended at home.
- The same process for selecting time of day period was defined for allocating the mode of travel to both in-direction and out-direction.
- A trip might have more than one travel mode. In such case, the priority is the transit mode, and after that, priority is assigned, in the following order, to auto driver, auto passenger, school bus, and non-motorized mode such as biking and walking.

The program implemented in MATLAB is shown in Appendix A. About 143,970 trips for 23,868 households were inserted as an input data source. This total number of trips was combined into 56,480 tours for about 22,105 households. Other household trips were neglected by the program due to some kind of errors such as missing data. The prepared data and information for each household tour were applied for the development of disaggregate model daily household tour production rate.

The observed frequency of trips and tours by travel purpose is presented in Table 3-4 and Figure 3-5. For tours, the purpose was defined based on the tour's primary activity. For trips, the purpose was defined according to the OD survey results. The comparison of tour and trip distribution indicates more work, school and university related tours than same related trips. This result is realistic and logical. In particular, work tours may include many non-work stops (e.g., for maintenance and discretionary purposes) on the way to and from work. When the tours are broken into trips, these stops produce many non-work trips which results in reducing the number of work trips. However, in reality these trips are a part of work travel and should not be considered as a real trip. Therefore, maintenance and discretionary tours are less than related trips as it is shown in Table 3-4 and Figure 3-5.

Table 3-4: Total observed tours and trips

Purpose	Work	University	School	Maintenance	Discretionary	Total
Trips	28274	2970	5305	27695	20080	84324
Tours	23189	2598	4454	14846	10306	55393
Trips%	34%	4%	6%	33%	24%	100%
Tours%	42%	5%	8%	27%	19%	100%

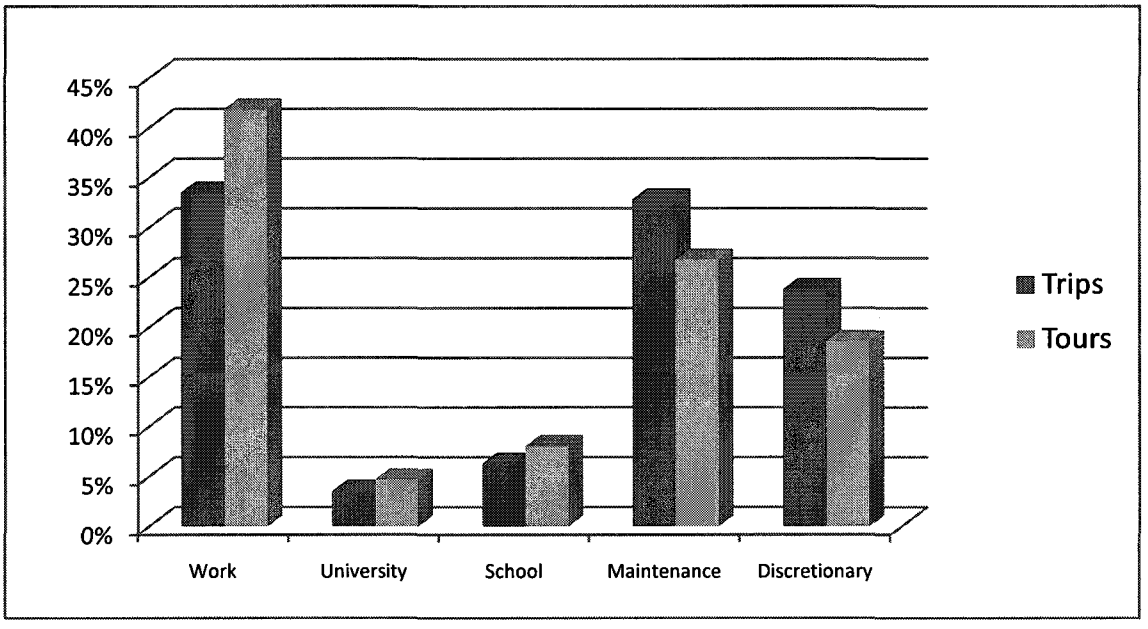


Figure 3-5: Observed frequency of trips and tours by purpose

4 Model Description and Development

4.1 Overall framework

A flowchart representing major components of model development analysis is illustrated in Figure 4-1. It briefly depicts that after extracting required data from 2005 OD survey database, relevant variables for each model should be prepared. Some of these variables, such as number of workers in the HH, or zonal population density, exist in OD survey dataset, whereas some others should be computed from available data (e.g., travel time impedance (accessibility) between the traffic zones can be computed using free flow travel time and equation (3-1)). The next step is to convert all of the chained trips into tours. Chained trips are a sequence of trips that originate at home and go to other locations such as work and finally end at home. The tours split into two directions, in-direction and out-direction. Each direction of a tour is represented by the origin and primary destination zone (tour purpose), time of day, and mode of travel.

The first implemented model is household car ownership choice. This model was calibrated by using nested logit and probabilistic neural network methods for training dataset and tested by validation data set. Then, in order to decrease the variety of the households in each traffic analysis zone (TAZ), households were divided into 4 different segments based on their attributes such as dwelling type and size.

For predicting the daily tour generation for each purpose, first, the daily household tour production rate for each household segment in each traffic zone is estimated. There are five separate models for each purpose. In Addition, five models have been implemented for daily zonal tour attraction. Both models are formulated by using linear regression and

feedforward backpropagation neural network. Again, all models were calibrated and tested by the same training and validating dataset. After the zonal tour production and attraction have been determined, they have to be balanced in order to have the same total tours for each purpose. The final stage of this study is the comparison of capability of ANN and linear regression methods in predicting tour generation based on activity-based concept.

4.2 Main Dimensions of Tour Model and Features

The tour model is segmented into important dimensions. The main dimension is purpose of tour which is segmented into tour attraction and production models. The other dimensions, namely, time of day, mode, and car choice are considered in tour construction procedure and household car ownership model. These dimensions are defined and stratified as follows:

Purpose: A tour is defined as travel between home and primary destination and back to home. The term purpose in this study is used to indicate the activity performed at the primary destination. Five travel purposes were defined based on O-D survey codes:

1. Work,
2. University: school for students of age 19 or older,
3. School: school for students of age under 19,
4. Maintenance: such as shopping and medical visit,
5. Discretionary: such as recreation and visit friends/ family.

Time: 5 time-of-day (TOD) cases were defined based on departure time:

1. Early (4:00 – 6:30),
2. AM (6:30 – 8:59),
3. Midday (9:00 – 15:29),
4. PM (15:30 – 18:29), and
5. Late (18:29 – 3: 59).

Mode: 6 travel modes were defined as follows:

1. Auto Driver,
2. Auto Passenger,
3. Transit,
4. School Bus,
5. Non-Motorized, and
6. Other.

Non-motorized modes are walk, bicycle, and motorcycle.

Car ownership: The number of cars that each household owned is categorized as follows:

1. Zero cars,
2. One car,
3. Two cars, and
4. Three or more cars.

4.2.1 Household segmentation

It is not feasible to estimate the numbers of tours produced by each (separate) household or individual; therefore, households are divided into different types for each traffic zone according to some socio-economic characteristics. Household types can be used daily HH tour production rate models for future research.

The variables are considered to divide households into different types, namely, number of workers versus number of non-workers, and dwelling type. These are defined in the following way:

Worker vs. non-worker: two categories were defined based on number of workers relative to the number of non-workers in the same household.

0. Number of workers in the HH is less than number of non-workers, and
1. Number of workers in the HH is equal or greater than number of non-workers.

Thus, the new variable of worker vs. non-worker is either 1 or 0. The value 0 is acting as a reference variable and 1 is a dummy variable.

Dwelling type: two dwelling types were defined for households.

0. Apartment (i.e., row / townhouse and apartment)
1. Detached.(i.e., single and semi detached, cottage)

Again, dwelling type is a nominal variable with two categories.

Hence, all of the households in each traffic zone can be split up into four types, based on two categories of worker vs. non-worker variable and two categories of dwelling type. Four household types are defined as shown in Table 4-1.

Table 4-1: Household segmentations

Household Type	Worker vs. non-worker	Type of dwelling
1	1	0
2	0	0
3	1	1
4	0	1

The objective of stratifying all of the 22,105 households in 556 TAZ into 4 categories is to simplify the model for future use. In this case, we need to determine the growth factor of each household type in all traffic zones. That is, instead of estimating at the disaggregate-level, data such as number of workers, size, and type of dwelling for each household in the future year, it is acceptable to work with household type.

The disaggregate model, namely daily household tour production rate, is calibrated based on household type concept. Therefore, if a household type and its location is specified, then, future tour production rate for a household can be estimated by using the current trained models.

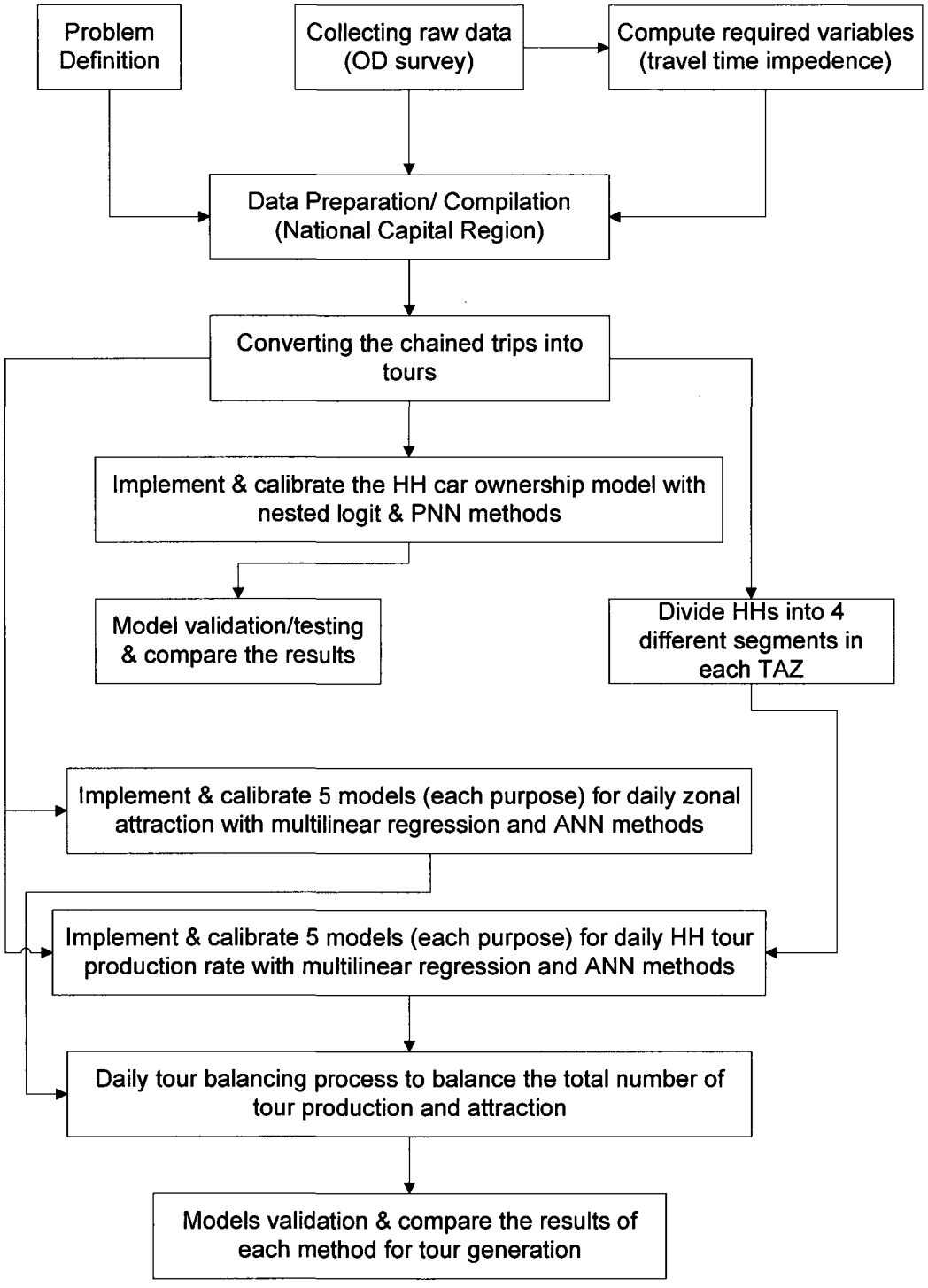


Figure 4-1: Overall Research Framework

4.3 Household Car Ownership Choice Model

The number of cars available for each household is one of the significant parameters for exploring the travel behavior of households. This model aims to predict the choice of each household (decision maker) for the defined car ownership choice set. The choice set has four alternatives: zero cars, 1 car, 2 cars, and 3 and more cars. Due to the small observed distribution of households with 4 or more cars (i.e., 4 cars (1.5%), 5 cars (0.3%), 6+ (0.2%)), all of these choices are combined into one category (3+ cars). The acquired observed frequencies from the OD survey is as follows:

- Zero car: 10.0%
- One car: 43.1%
- Two cars: 38.3%
- Three+ cars: 8.6%

This choice probability is calculated based on two different analytical methods: nested logit model and probabilistic neural network (PNN). These two methods share the same training and testing dataset and explanatory variables.

Explanatory variables deployed in this model are exhibited in Table 4-2. These variables are split up into two categories, household level and zonal level. Both methods are using these variables. However, in the nested logit model analysis, these variables should be reorganized as it is discussed later in Section 4.3.1.2.

Table 4-2: Explanatory variables for car ownership model

Household data (disaggregate)	Zonal data (aggregate)
• Number of workers • Number of non-workers in comparison with number of workers • Dwelling type	• Zonal population density • Percentage of low-income households • Percentage of detached houses • Shopping gross leasable area (GLA) • District population density • Accessibility Measurement

4.3.1 Nested Logit Model

4.3.1.1 Nested Logit Structure

This model is formulated as three-level nested logit model where the choice set consists of four alternatives. It is assumed that a household prefers to have a car(s) or not to have any cars, therefore, in upper level, one or more cars (3 alternatives) can be grouped together and zero-cars choice is in the other nest. Again, it is assumed that the choices of more than one car are more correlated (2 cars or 3+ cars). Those can be grouped together. Finally, in the lower level, all of the alternatives have been set separately. The constructed nest structure for this study is presented in Figure 4-2.

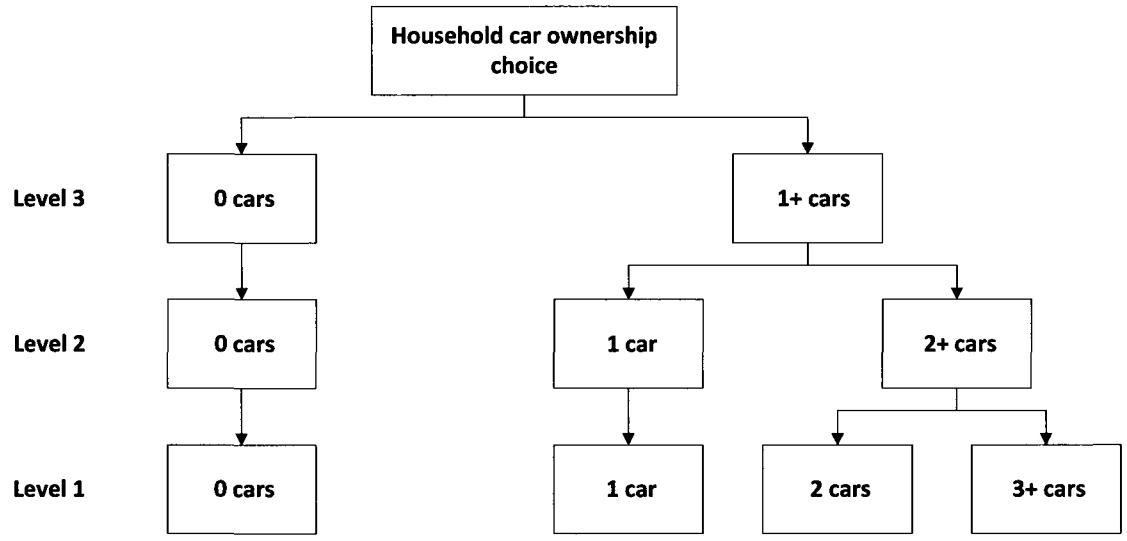
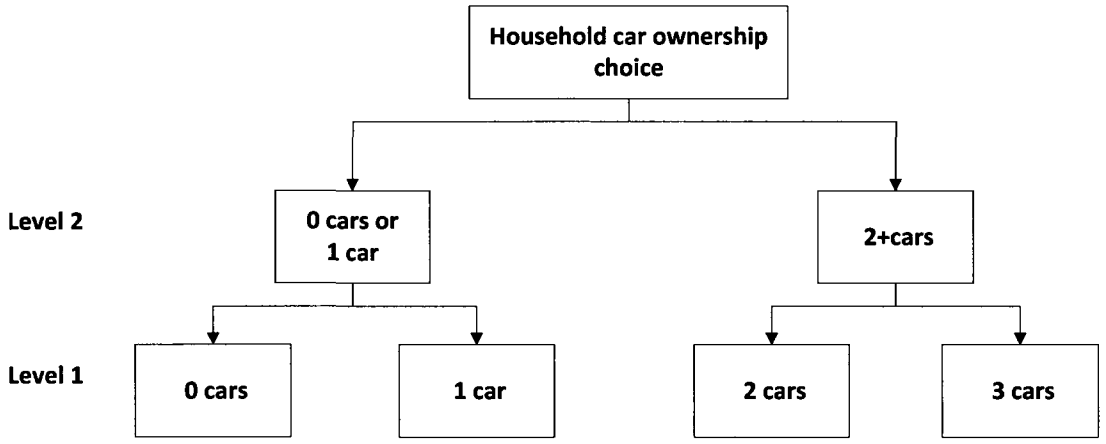


Figure 4-2: 3-level nested structure of car ownership choice

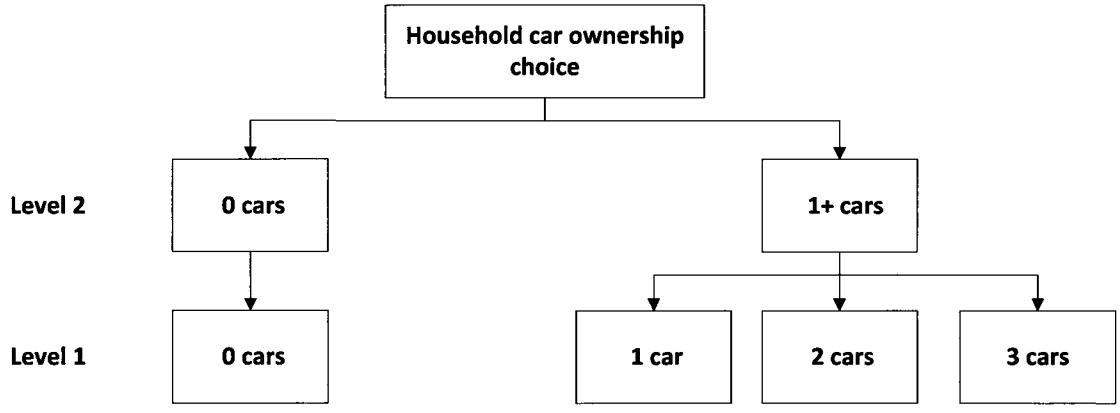
However, some other structures can be considered as well. There are only two other models (i.e., model A & B) that have a rational justification to compare with the main model. These two structures have two-level nested logit structure. Figure 4-3 has exhibited these two structures and shows how the alternatives were grouped, although these structures were not used in this study.

In this research, the analysis of nested logit model was implemented by the software packages SAS (9.1.3). In the multinomial discrete choice (MDC) procedure analysis in SAS, the first calculated output which corresponding to equations (2-6) and (2-9), is the random utility function for all alternatives. The random utility function shows the relative pleasure that the decision maker (households) expects to derive from that choice, considering the other alternatives in the choice set. As shown in equation (2-6), the utilities are not known to the analyst with certainty and contain random components; the individual choice process becomes a probabilistic procedure. Therefore, the probability of

choosing an alternative (car ownership) in a choice set is computed by using SAS. It is assumed that a household chooses the alternative for which the associated probability is highest. It is obvious that the total sum of probabilities of all alternatives for a household should be one.



2-level nested structure (model A)



2-level nested structure (model B)

Figure 4-3: Other possible nested structures

All of the maximum probabilities for each household can be collected in a new vector. This vector indicates the choice of each household in terms of the calibrated model.

4.3.1.2 Development of variables

The prepared data set with defined independent variables shown in Table 4-2 could not be used in the MDC procedure in SAS. Therefore, the following changes were made to the data set.

The dependent variable, called decision, shows the decision of a household for choosing one of the alternatives. It takes the value of 1 when a specific alternative is selected; otherwise it takes the value of 0. Corresponding to SAS manual, each household is allowed to choose one and only one of the possible alternatives. In other words, the variable decision can take the value of 1 only one time for each household.

In doing so, a new variable car choice was defined which goes from 1 to 4 to capture all of the alternatives. Then, each household with the same explanatory variables should be expanded 4 times. For example, there are 23,868 households in this data set. But, in order to generate the data set that meets the above requirements, the final total number of observations should be 95,472 (23,868×4).

The other mandatory variable, which should exist as an input, is the identification variable. This variable specifies the ID number of each household. The variable ID has to begin from one and goes up in sequence from 1 to 23,868 and each number has repeated 4 times.

All of the deployed explanatory variables in the nested logit model are presented in Table 4-2. The important point is that all variables are alternative-invariant variables. As it is discussed earlier in Section 2.4.3.2, SAS can only work with alternative-specific variables. Therefore, according to suggested method in the same section, the chooser-

specific variables should be converted into choice-specific variables. Because, there were four alternatives in the choice set, each variable was transformed into three alternative-specific variables, and the fourth one was considered as a reference variable. Zero cars was assumed to be the reference variable, and its value was zero for the whole data set. Other alternatives were used to construct the other variables. For example, consider the variable zonal population density (zonpopden); it is transferred to following variables: zonpopden-car1, zonpopden-car2, and zonpopden-car3. The value for these variables is equal to zonal population density value for that zone, when the variable car choice for that household is equal to the specified alternative, otherwise it is zero.

4.3.2 Probabilistic Neural Network

4.3.2.1 Overview

PNN seems to be a useful alternative for nested logit model in solving classification problems in travel demand modeling. It has a simple structure that can be easily modeled. This has led to the application of such networks to many practical problems. In this work, it is used for modeling household car ownership. The software package MATLAB (7.5.0) has been used for modeling PNN.

The transform function between hidden layer and output layer is radial basis function (please see equation (2-3) on page 21). This function computes the probability of car choice for each household, and on the last layer, the maximum of these probabilities, is picked and produces a 1 for that car-choice category and a 0 for the other categories. Hence, PNN is a black-box model and its output only indicates the choice of household for number of owned cars. As compared to the nested logit model, there is no information

either about the utility function value or probability function value for each car choice category.

The other important point to note is that each time that the network is trained, a new network is being created. Thus, it is essential to save the network in every step of the training process. Clearly, it is always complicated to find the best trained network. As shown in Figure 2-4, after comparing the output results and target, we can keep the best trained network with more reasonable results. This is very necessary for using the calibrated model for future predictions.

4.3.2.2 Data Preparation

All of the variables exhibited in Table 4-2 have been deployed for input data. Because, "the number of non-workers in comparison with the number of workers" and "dwelling type" are categorical variables, therefore, they should be transformed to dummy variables. The reference variable for "number of non-workers in comparison with number of workers" variable is number of non-workers, if there is no worker in the household and reference variable for dwelling type is apartment. All of the reference variables are coded as 0.

As discussed in 3.2.1, the data set was split into two sub-data sets: training and testing. Both of the models, discrete choice model and PNN, share the same data sets. As a matter of fact, they only use the same number of households and main variables but these variables have been changed due to model requirements. Table 4-4 represents the associated data sets information for each model and their differences.

In training dataset, explanatory variables and dependent variable (number of cars) should be kept separate. The number of neurons in the input layer of PNN is equal to the number of households and the number of neurons in the output layer is equivalent to the choice set, which is four in this model.

In addition, it should be noted that the target value could not be 0 in MATLAB, therefore the dependent variable (car choice) should be recoded as it is shown in Table 4-3.

Table 4-3: Car choice code in PNN model

Variable	Old Code	New Code
carchoice	0	1
	1	2
	2	3
	3	4

Table 4-4: Associated data set in choice and PNN models

Data set	Type	Model	# Observation	# Exp.Variables	Data Size
Carchoice-Train	Training	NL	16,000	39	16000*39
Carchoice-Test	Testing	NL	7,868	39	7868*39
Carchoice-TrainInp	Training	PNN	16,000	12	12*16000
Carchoice-TrainTar	Training	PNN	16,000	1	1*16000
Carchoice-TestInp	Testing	PNN	7,868	12	12*7868

4.4 Tour Generation Sub-Models

Tour generation model is composed of two major models: household daily tour production and zonal tour attraction models. Both models were developed for five travel

purposes. Mainly, tour production model is trying to estimate all of the tours that are produced from all traffic zones by purpose, as tour attraction model determines the attracted tours in observed traffic zones. After zonal tour attraction and production are estimated, it is important that the total number of tours from both models should be equal. If there is a difference between the total number of tours, then they have to be balanced, otherwise either tours will be lost or more tours will be generated.

Thus, all of the produced tour data sets were stratified by five travel purposes based on O-D survey results. Then, each data set was deployed for one model. Thus, there were five models for tour production and 5 models for tour attraction, taking into account the purpose of tour. This division should be done because same variables are not significant in prediction of tour production or attraction for each purpose. Table 5-9 and Table 5-19 show the most significant explanatory variables that explain the household tour generation rate.

4.4.1 Household Daily Tour Production Model

Household daily tour production models were developed for the five travel purposes. This model was estimated as a linear regression and artificial neural network model. Tour production rate is estimated for each household type in all traffic zones, based on the disaggregate level (household attributes) and aggregate level data (zone of residence characteristics). Thus, this is a disaggregate model. In other words, the tour production model does not focus only on zonal variables, but it considers the household attributes and its composition as well.

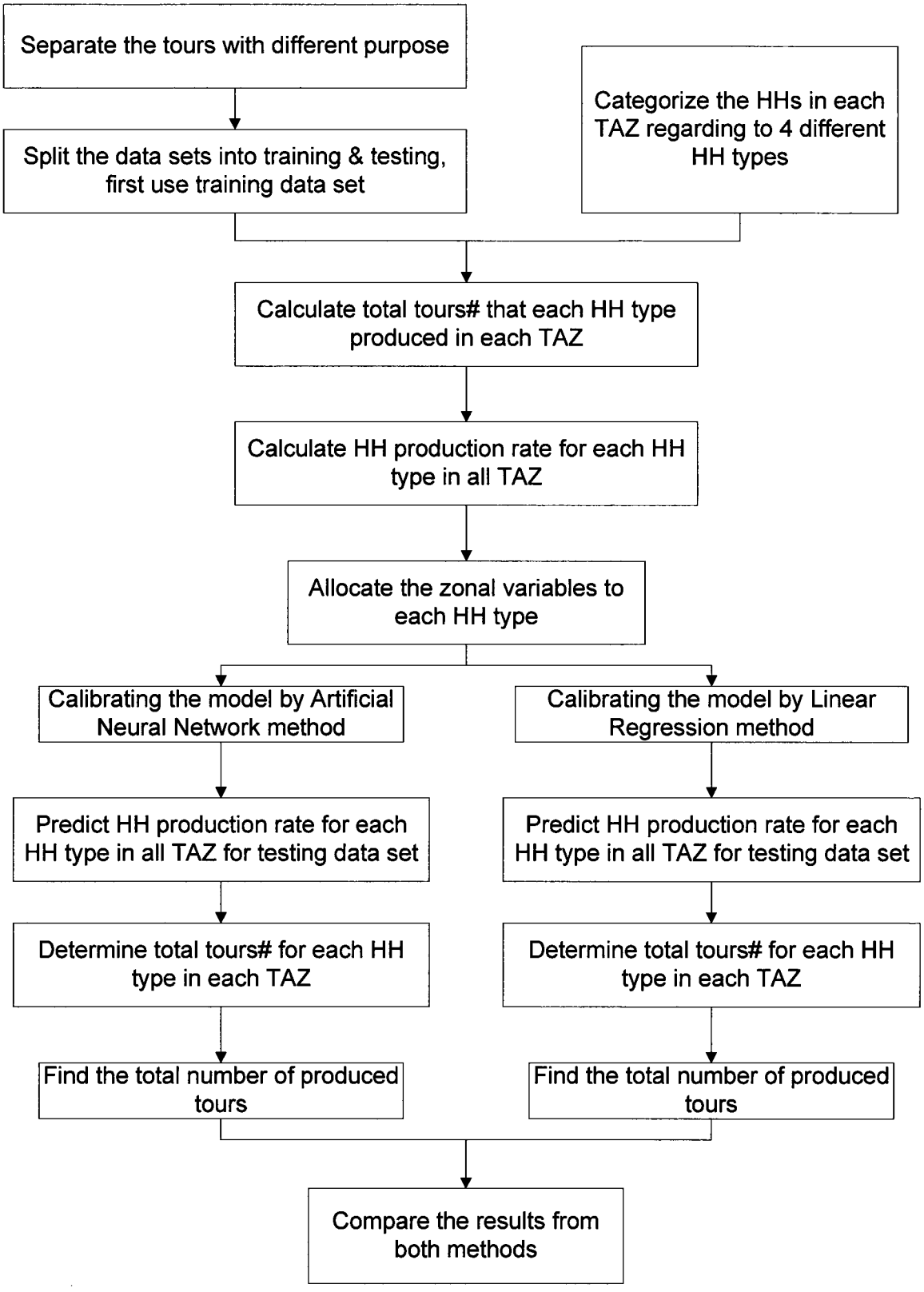


Figure 4-4: Household daily tour production model procedure for regression and ANN.

First of all, the households in each traffic zone were classified according to Table 4-1. Then, the total number of HHs and related produced tours were determined. The independent variable of the model is the household production rate that can be determined, based on equation (4-2). Two different methods, ANN and regression, were deployed to calibrate the models. After the model was calibrated, then the training data set was applied to test the efficiency of the model. The acquired results from both methods can be compared to find relative success of the models. This procedure is shown in Figure 4-4.

All production and attraction models operate with tours and provide daily tour numbers. Therefore, they do not consider the time of day for each tour. Another model can be implemented corresponding to time of day information to consider time of day choice and existence of stop in each leg of tours.

After the model is calibrated, the model results for testing and training data for household segments in each traffic zone are combined into zonal tour production according to equation (4-1).

$$P_{ui} = \sum_{tu} R_{tu} \times H_t \quad \textbf{(4-1)}$$

Where:
i = traffic zone index,
u = tour purpose from 1 through 5,
t = household type from 1 through 4,
P = zonal tour production,
H_t = number of households in the zone by segments,
R = daily household tour production rate.

Then, the sum of all zone tour production indicates the total produced tours in study region.

4.4.1.1 Development of variables

Table 4-5 represents the associated variables in household daily tour production model. The variables "worker greater than non-worker" and dwelling type were defined in Section 4.2.1. Because, both of these have only two categories and coded 0 and 1, therefore, there is no need to create new dummy variables for this model.

Different variables from the list of variables shown in Table 4-5 were selected for each model based on their conceptual effect on predicting the produced tours for different purposes. For example, the total number of employees is not a significant variable for forecasting the produced trips for the purpose of going to school for households. Thus, the variables which were selected for each model have been presented in Table 4-6. Some of these variables were excluded in the final model because their t-statistic value was not significant.

Table 4-5: Explanatory variables in household daily tour production

Household data (disaggregate)	Zonal data (aggregate)
• Worker greater than non-worker • Dwelling type	• Zonal population density • Percentage of low-income households • Percentage of detached houses • District population density • Accessibility Measurement • Total employment

The car ownership as a variable is not included here for two reasons. Firstly, in the definition of household types, different levels of car ownership will add many dimensions to this categorization problem. Secondly, the proxy for car ownership is the percentage of low income households in the zone.

Table 4-6: Selected explanatory variables for tour production models

Variable	Name	Purpose				
		work	University	School	Maintenance	Discretionary
dwelling type (dummy)	Dwell	√	×	√	√	√
worker greater than nonworker (dummy)	Workvsnon	√	√	√	√	√
zonal population density (persons/ha)	zone_pop_den	×	√	√	×	√
zonal percentage low income HH	zone_lowinc	√	×	×	√	√
zonal % detached dwelling units	zone_det	√	√	√	×	√
district population density (persons/ha)	dis_pop_den	√	√	√	√	√
Accessibility index	Access	√	×	×	√	√
total employment	emp_tot	√	×	×	×	×

The independent variable for this model is household production rate. It is calculated by using equation (4-2).

$$R_{tu} = \frac{H_t}{N_{tu}} \quad (4\text{-}2)$$

Where:
N_{tu} = number of produced tours in the zone by household segments.

The output of the model is daily household tour production rate by type and purpose and then the number of produced tours by household segments can also be calculated according to equation (4-1) for testing model and future forecasting.

4.4.1.2 Multiple Linear Regression Method

All 5 statistical models were developed and analyzed as a linear regression model based on version 16.0 of SPSS and the prepared data sets for each purpose were input into SPSS files. The method of least square error was used to find the line that best fits the data.

SPSS outputs provide the overall fit of the model or the value of R and R^2 for the model. In addition, the model parameters are calculated by SPSS. These parameters are coefficients of variables, intercept, t-statistic value, and probability value (p-value). As a general rule, if the p-value of a coefficient is less than the related value for defined confidence interval which in this study is taken as 0.05, then, it is assumed that the variable has a significant effect on household daily tour production. Otherwise, that variable was taken out from the model.

The other important assumption was made is that the regression models were estimated without intercept. This was done to ensure that traffic zones containing no demographic or socio-economic data would not generate data and there is no default household tour generation rate.

Different methods can be used according to SPSS capability for finding the significant predictors. These methods are:

- Hierarchical (Blockwise Entry),
- Forced entry,
- Stepwise,
- Backward.

More detailed explanation can be found in (Field, 2005). In this research, the backward method was selected and used. In backward method, all selected variables (Table 4-6) were placed in the model and then the contribution of each one is explored by looking at the p-value of the t-statistic for each variable. In this study, whenever this significance value was more than 0.05, then the variable was removed from the model and the model was re-estimated for the remaining variables. This process continued until all of the remaining explanatory variables in the models were significant.

After finding the coefficients of variables for the models, these calibrated models can be used to predict the household tour production rate for the test data set. After finding the production rate for each household type in traffic zone, total produced tour can be calculated according to equation (4-1).

4.4.1.3 Artificial Neural Network

Five models were implemented in "Neural Network Toolbox MATLAB 7.5.0" with different structures. The feedforward backpropagation network was only one used for all types of models, since inputs from layers were always passing forward and not moving backward to previous layers. Once these networks were trained to predict the household

production tour rate as precisely as possible, therefore, it was essential to save the optimum network for testing and future predictions.

4.4.1.3.1 Network Architecture

It was found in the literature review that feedforward networks with one hidden layer of sigmoid neurons followed by an output layer of linear neurons could approximate any function with a finite number of input and output neurons, given sufficient neurons in hidden layer. Thus, this structure shown in Figure 4-5 was used for all of the networks. The only difference between them is the number of neurons of hidden layer.

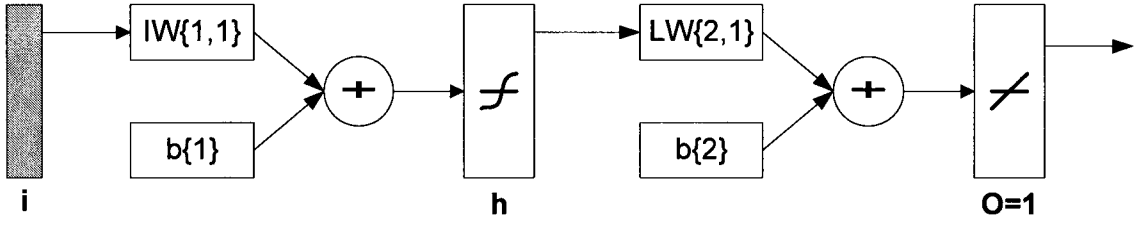


Figure 4-5: Feedforward neural network structure

IW shows associated initial weights of input layer and LW represents weights of hidden layer. The number of neurons of input layer (i) depends on the number of selected explanatory variables for each model. Output layer always has one neuron (o=1), because there is only one value for dependent variable (HH production tour rate). Selecting the number of neurons in hidden layer is an arbitrary process and it can take any finite value. Initial guess for number of neurons can be selected based on number of input data size. In this model, the number of input neurons and number of observed households is shown in Table 4-7, therefore, three different numbers were selected for the number of hidden

neurons for each tour purpose. Then, one of them was selected by comparison between the output results and the target (i.e., observed produced tours). For example, for work, maintenance, and discretionary tours, the number of hidden neurons was assumed to be 150, 200, 250 and eventually the network with 200 hidden neurons produced better results. Same process was used for produced university and school tours and all of the attraction models. The activation function selected was a tan-sigmoid (hyperbolic tangent sigmoid) transfer function for the neurons belonging to the hidden layer while at the output layer a linear activation function, purelin in MATLAB's neural network toolbox, was considered.

The input data is a matrix where the number of rows is equal to explanatory variables and the number of columns is equal to the number of observed households for each tour purposes. Table 4-7 indicates the number of neurons in input layer, the size of input data and selected number of neurons in hidden layer of each model. Total number of tours is only for training the data set.

Table 4-7: Number of neurons in tour production neural network models

Purpose	Number of input neurons	Number of observed HHs	Number of hidden neurons
Work	6	731	200
University	4	250	100
School	5	369	100
Maintenance	5	765	200
Discretionary	7	737	200

4.4.1.3.2 Training of Models

Training adjusts weights and biases to find the minimum performance function. Indeed, performance function in a feedforward network is the mean square error (MSE). Training process for all models was carried out based on MATLAB's algorithm "trainlm" as the network training function. The training parameters for trainlm were as shown below and their initial value is set as presented in Table 4-8.

- epochs: if number of iteration in training exceeds epochs , then training stops,
- goal: if performance function is less than goal, then training stops,
- mem_reduc: it is used to cope with memory problem of central processing unit,
- min_grad: if gradient value is less than min_grad, then training stops,
- mu: value of the learning rate,
- mu_dec: if performance function is reduced by step, then, mu is multiplied by mu_dec,
- mu_inc: if performance function is increased by a step, then, mu is multiplied by mu_inc,
- mu_max: training stops while mu exceeds mu-max value,
- show: demonstrates training status on each epoch,
- time: if training time exceeds this limit (in seconds), then, training stops.

One of the problems which may occur during training process is called overfitting. This happens when the network is well trained for the input data set, but, the network error is large for the new data. . In this study, for minimizing the effect of this problem, the early stopping method, suggested by Demuth & Beale (2001), was used. The train data set split

into three subsets: training, validation, and test set. Training subset consists of $\frac{1}{2}$ data set and validation and test subsets are composed of $\frac{1}{4}$ data set.

Table 4-8: Trainlm algorithm parameters

trainlm parametes	**Value**	**trainlm parametes**	**Value**
Epochs	300	mu_dec	0.1
Goal	0.01	mu-inc	10
max_fail	5	mu_max	$1.0e^{10}$
mem_reduc	1	Show	25
min_grad	1.0e-10	Time	Infinite
Mu	0.001		

To make the neural network training more efficient two preprocessing steps on the network inputs and targets were performed. The objective of these processes is to transform inputs and targets into a better form for the network use. These two preprocessing steps are:

- Normalize the mean and standard deviation of the training data set
- Principal component analysis

It is better to normalize the inputs and target so that they will have zero mean and unity standard deviation. It is useful to perform a principal component analysis and remove those components which account for less than 1% of the variation. These steps can be achieved in neural network toolbox in MATLAB by utilizing the following commands:

- $[p_n , p_{s1}] = \text{mapstd}(p)$

- $[p_{trans}, p_{s2}] = processpca(p_n, 0.01)$

Where p is the original input data, p_n is normalized data, p_{trans} is a transformed input data and the data that are used for training the network.

4.4.2 Zonal Tour Attraction Model for Primary Destination

Daily tour attraction models were implemented for each tour purpose as linear regression and artificial neural network models. Daily zonal tour attraction models were calibrated based on observed daily tours of primary destination and the relevant socio-economic and land–use variables on associated traffic zone. Thus, these models were only based on zonal variables and they were aggregate type of models. The predicted tours from both methods, regression and ANN, should be compared to find the more efficient model.

Figure 4-6, briefly, presents the process used to develop zonal tour attraction model by using two methods.

4.4.2.1 Variable Description

The selected variables for all model types are as follows:

- Zonal population density,
- Percentage of detached houses,
- Shopping gross leasable area (GLA),
- District population density,
- Accessibility measurement,
- Total employment,
- Total population,
- University enrollment,
- School enrollment.

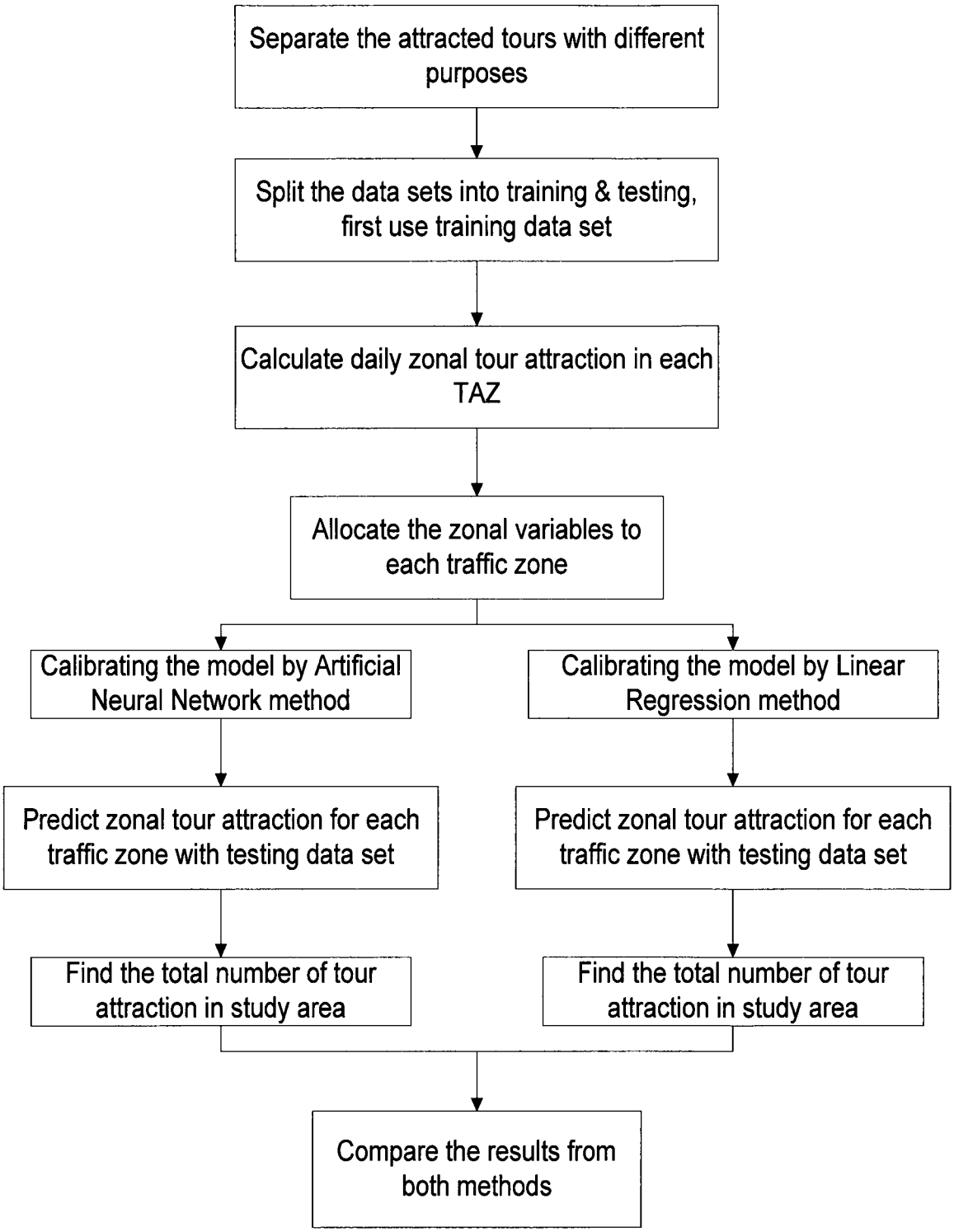


Figure 4-6: Zonal tour attraction model development procedure for regression and ANN methods

All of the mentioned variables were not considered in all of the models but some of them were selected due to their effect on attracting trips (tours) to a traffic zone. For example, university enrollment was a significant variable for estimation the number of attracted tours for going to school in a traffic zone. The selected variables for each model have been shown in Table 4-9. Some of these variables were taken out of the final model because either their p-value was not significant or their coefficient did not have an appropriate sign.

Table 4-9: Selected explanatory variables for the attraction models

Variable	Name	Purpose				
		work	University	School	Maintenance	Discretionary
zonal population density (persons/ha)	zone_pop_den	√	×	×	√	√
zonal % detached dwelling units	zone_det	√	×	×	×	×
shopping gross leasable area (GLA)	GLA	×	×	×	√	√
district population density (persons/ha)	dis_pop_den	√	×	×	√	√
Accessibility index	Access	√	√	√	√	√
total employment	emp_tot	√	×	×	×	×
total population	pop_tot	√	√	√	√	√
university enrollment	uni_enrol	×	√	×	×	×
school enrollment	sch_enrol	×		√	×	×

Required information for the dependent variable, daily zonal tour attraction, was obtained from observed daily tour for primary destination zone for each tour purpose.

4.4.2.2 Linear Regression Method

All of the models for tour attraction were developed in the same way as tour production models. They have the following form of linear regression without an intercept.

$$A_{uj} = \sum_m \beta_m X_{jm} \quad \text{(4-3)}$$

Where:

j	=	traffic zone of the primary tour destination,
A_{uj}	=	zone tour attraction for the primary destination for the specified purpose,
m	=	index of independent variables,
β_m	=	regression coefficient for each explanatory variables,
X_{jm}	=	explanatory variables for each zone.

The models were calibrated by training data, then, were tested by using the testing data set.

4.4.2.3 Artificial Neural Network

All of the tour attraction models developed in ANN, have the same structure as the production models shown in Figure 4-5. The only difference in ANN structure is the size of input matrix (number of neurons in input layer) and selected number of neurons in hidden layer.

Table 4-10 indicates the number of neurons in input layer, the size of input data and selected number of neurons in hidden layer of each model. Total produced tours are determined based on O-D survey results.

Table 4-10: Number of neurons in tour attraction neural network models

Purpose	Number of input neurons	Number of observed HHs	Number of hidden neurons
Work	4	352	100
University	2	150	20
School	2	196	20
Maintenance	3	344	100
Discretionary	5	348	100

4.4.3 Balancing of Total Daily Tour Production and Attraction

After the zonal tour productions and attractions are calculated, they have to be balanced in order to match the total tours in the study area, according to following steps:

Step1: Calculate total production and attraction tours for the study area and for each purpose.

$$P_u = \sum_i P_{ui} \quad \& \quad A_u = \sum_j A_{uj} \tag{4-4}$$

Step 2: Calculate a balanced total tour for each purpose based on the total tour attraction and tour production.

$$B_u = \sqrt{P_u \times A_u} \tag{4-5}$$

Step 3: Find a scale to convert the tour production and attraction of each traffic zone in such a way, that the total original production and attraction tours match the total tours in study area (B_u).

$$S_{pu} = \frac{B_u}{\sum_i P_{ui}} \quad \& \quad S_{au} = \frac{B_u}{\sum_j A_{uj}} \tag{4-6}$$

Where
S_{pu} = production balance factor,
S_{au} = attraction balance factor.

5 Model Estimation Results

5.1 Household Car Ownership Model

5.1.1 Nested Logit Model

The car ownership model was estimated as a 3 level nested logit model based on disaggregate household data according to Section 4.3.1. Table 5-1 provides information about model fit such as number of observations and number of iterations.

Table 5-1: Model fit summary

Dependent Variable	Decision
Number of Observation	23868
Number of Cases	95472
Log Likelihood	-23067
Maximum Absolute Gradient	4856
Number of Iterations	242
Optimization Method	Dual Quasi-Newton
Schwarz Criterion	46577

As reflected in Table 5-1, the number of observations shows the number of households and number of cases is equal to number of households multiplied by number of car choice(s). Decision (dependent variable) is one for each household case which is equal to its observed owned cars, otherwise it is zero.

5.1.1.1 Model Calibration and Validation Results

The utility coefficients for each alternative were estimated and the results were summarized in Table 5-2. Detailed estimation results and statistics were provided in the Appendix B. Zero-car alternative is considered as the reference value with all coefficients set equal to zero.

Table 5-2: Estimation results for car ownership model

Variable Description	Coefficients by alternative			
	0 Cars	**1 Car**	**2 cars**	**3+cars**
0 workers in HH		-0.4088 (0.116)	-2.0485 (0.001)	-4.4996 (0.000)
1 worker in HH		0.5604 (0.000)	-0.5480 (0.041)	-2.2846 (0.000)
2 workers in HH		1.2740 (0.000)	0.8307 (0.004)	-1.1095 (0.000)
3+ workers in HH		0.8599 (0.000)	0.3938 (0.078)	-0.5108 (0.000)
# non-workers if workers=0		0.4572 (0.000)	0.7725 (0.000)	0.9471 (0.000)
# non-workers if workers=1		0.2811 (0.001)	0.4880 (0.000)	0.5057 (0.000)
# non-workers if workers>=2		0.4742 (0.001)	0.5631 (0.000)	0.7627 (0.000)
Detached home		-0.0249 (0.526)	-0.0120 (0.770)	-0.0146 (0.803)
Zonal population density		-0.001887 (0.019)	-0.004320 (0.002)	-0.0140 (0.000)
Zonal % low income HH		-1.3870 (0.000)	-1.5937 (0.000)	-1.4196 (0.000)
% Detached dwelling units		0.008331 (0.001)	0.0127 (0.000)	0.0121 (0.002)
Variable Description	**Coefficients by alternative**			
	0 Cars	**1 Car**	**2 cars**	**3+cars**
District population density		-0.007736 (0.000)	-0.009626 (0.000)	-0.0151 (0.000)
Gross Leasable Area (GLA)		-8.876e-7 (0.047)	-1.413e-6 (0.009)	-2.278e-6 (0.0084)
Accessibility index		0.0422 (0.005)	0.0556 (0.001)	0.0684 (0.000)
Inclusive value (level 3)	0.59	0.59		
Inclusive value (level 2)		2.11	2.11	

Note: The numbers in parenthesis show the probability value for each parameter.

The magnitude of "Inclusive value" is the same for all of the nests at the same level; therefore, there are only two inclusive values for each level.

Three different measures can be applied to test how well the logit model fits into the observed data. These are:

- Examination of coefficients and t-statistics
- Goodness of fit measure (R^2)
- Percent correct estimation

Examination of coefficients and t-statistics:

In Table 5-2, the value of coefficients shows the effect of explanatory variables on dependent variable. Therefore, the estimation results indicate that the number of workers in the household is the most significant variable and the existence of shopping gross leasable area in the traffic zone is less significant for predicting the number of cars available to the household. Other variables that can help predicting the number of cars available to a household are: the remaining number of non-workers of the household, whether the home is detached or not, percentage of low-income households in the zone and accessibility measure between traffic zones.

The significance of the individual variables may be investigated through the p-value and the sign of the coefficient. Table 5-2 represents the p-value of all coefficients. As shown, the p-value of the variable "detached home" is insignificant according to the t-test, but this variable was not removed from the model, because, lack of significance may simply be caused by lack of sufficient data.

The sign of a variable should also be considered as an important factor and should be logical in regard to the effect of that variable on dependent variable. Thus, the coefficients of explanatory variables included in the model can be interpreted in terms of sign and magnitude.

- Households with 1 worker are most likely to have 1 car, less likely to have 2 cars or no car, and much less likely to have 3 cars.
- Households with 3 or more workers in the household show a higher preference to have 3 or more cars in comparison to other households with less than 3 workers.
- As the percentage of detached dwelling units in a traffic zone increases, households are more likely to have 2 cars or more.
- The predicted number of cars available is substantially correlated (in a negative way) to the percentage of low-income households in the zone. That is, this variable has a strong negative impact on car ownership categories, and this is logical. Households with lower income are more likely to have zero car and less likely to have one car and much less likely to have more than 2 cars.

Goodness of fit measure (R^2):

The other measure that can show the goodness of fit of the model is the likelihood ratio index (R^2). As it was discussed in 2.4.3.4, this index was determined by several equations and the results were presented in Table 5-3. The value of R^2 must take a value between 0 and 1.

Table 5-3: Likelihood ratio index values

Measure	R^2 Value
McFadden	0.3029
Estrella	0.6322
Adjusted Estrella	0.6304
Cragg-Uhler 1	0.5682
Cragg-Uhler 2	0.606
Aldrich-Nelson	0.4564
Veall-Zimmermann	0.6211

This table shows that the model predicts the number of cars available to a household significantly, since the R^2 values are less than 1.

Calculated Results:

After estimation of utility equations of each alternative of a household, the probability of choosing each choice set was calculated according to equations (2-21) to (2-23) in Section 2.4.3.2. Then, the alternative with the highest value was selected as the number of cars available for the households. In order to investigate the accuracy of the model, it was essential to find out how many of car choices of households were predicted correctly, instead of comparing the total predicted choice of one category to total observed frequencies from the OD survey. For example, the total number of households with one car from the model should not be compared with total observed households with one car from the survey.

Table 5-4 indicates the observed frequencies calculated from the OD survey for four alternatives and predicted frequencies determined from the model.

Table 5-4: Comparing NL model results with observed OD survey car choices.

Alternatives	NL Model	OD Survey	% Correct
Zero Cars	327	2387	13.70%
1 Car	7048	10277	68.58%
2 Cars	6129	9133	67.11%
3+ Cars	66	2071	3.19%
Total	13570	23868	56.85%

NL: Nested Logit Model

Table 5-4 shows that about 60 percent of predicted choice decisions for all households in comparison with OD survey results are correct for this model. This model can predict the households with one or 2 cars almost precisely but the model is not predicting the household choice for zero cars and more than 3 cars. We may explain this result in the light of the sample sizes. In the case of zero cars and 3+ cars, the sample size was much smaller than that for the category of 1 car and 2 cars.

Table 5-5: Sample size for all four categories for household car ownership model

Alternatives	Size	Percentage
0 Cars	2387	10.00%
1 Cars	10277	43.06%
2 Cars	9133	38.26%
3+ Cars	2071	8.68%

5.1.2 Probabilistic Neural Network

Probabilistic neural network models with radial basis function were implemented by using MATLAB neural network toolbox. As discussed in Section 2.3.5, the ANN model was like a black box model. It converted directly the input information into output results

without providing any explanation. Thus, the correlation between variables or the relative effect of the model inputs on the output variables was not measured.

The only indicator of model's quality is the output results. Therefore, the model efficiency in terms of how well this model fits into the observed data can be determined only by comparing the percentage of correct output results. The other measures used for nested logit method, namely coefficients and related t-statistics value and goodness of fit measure (R^2) are not available for the PNN model.

First, the network was trained with training data set and the trained network was saved. Then, the testing data set was used to verify the saved trained network. The final results were compared with OD survey results (Table 5-6).

Table 5-6: Comparing PNN model results with observed OD survey car choices.

Alternatives	PNN Model	OD Survey	% Correct
Zero Cars	1180	2387	49.43%
1 Car	6854	10277	66.69%
2 Cars	5262	9133	57.62%
3+ Cars	484	2071	23.37%
Total	13780	23868	57.73%

PNN: Probabilistic Neural Network

For the examination of results, the households with correct choice are only selected. This model is capable of predicting correctly the decision of all of the households for choosing the number of cars at about 60% level. In addition, the percentage of households with zero cars, 1 car, 2 cars, and 3 cars which have been correctly estimated is about 50%, 70%, 60%, and 25%, respectively.

5.1.3 Comparative Results

The results of the nested logit model and the probabilistic neural network model can be compared in terms of percent of correct estimation by each model. For this purpose, Table 5-4 and Table 5-6 have been utilized, and were extracted for respective categories and shown in Table 5-7 in order to compare the final results of both models.

This table indicates the percent difference of correctly predicted households for each category from both methods against the observed frequencies calculated from the OD Survey. It shows how accurately both models can forecast the number of cars available for households.

Table 5-7: Comparison of final results of NL and PNN models

Alternative	**OD Survey**	**NL**		**PNN**	
		Correct #	% Difference	Correct #	% Difference
Zero Cars	2387	327	***13.70%***	1180	***49.43%***
1 Car	10277	7048	***68.58%***	6854	***66.69%***
2 Cars	9133	6129	***67.11%***	5262	***57.62%***
3+ Cars	2071	66	***3.19%***	484	***23.37%***
Total	23868	13570	***56.85%***	13780	***57.73%***

The results show that the PNN model predicts the number of available cars to each household more accurately than the nested logit model, although the total percent of corrected predicted households are almost the same.

5.2 Household Daily Tour Production

5.2.1 Linear Regression Methods

The household daily tour production rate model has been estimated for 5 travel purposes as a linear regression model based on disaggregate household data from OD Survey and according to the specifications and variables mentioned in Section 4.4.1.1. The regression models were estimated without intercept because no default household tour production rate was assumed.

First, the regression model was run with all of the selected variables noted in Table 4-6 then, insignificant variables (p-value > 0.005) were taken out from each model. The eliminated variables are shown in Table 5-8.

Table 5-8: Eliminated variables in final regression models for tour production

	Variable(s)
Work	Zonal population density (p = 0.284)– Total employment (p = 0.379)
University	Detached house (p = 0.518)
School	None
Maintenance	Zonal population density (p = 0.889)
Discretionary	None

p-values are in parenthesis

Final regression models with significant variables were run in SPSS. The model estimation results (i.e., coefficients) for 5 travel purposes are summarized in Table 5-9. The detailed estimated results and statistics for these models are included in Appendix B.

Table 5-9: Regression coefficients by travel purpose for HH daily tour production

Variables	Regression coefficients by travel purpose				
	work	University	School	Maintenance	Discretionary
detached house	-0.062 (0.000)	0.025 (0.518)	-0.048 (0.109)	-0.088 (0.000)	-0.122 (0.000)
worker greater than nonworker	-0.066 (0.000)	0.108 (0.002)	0.176 (0.000)	0.072 (0.000)	0.042 (0.005)
zonal population density		0.004 (0.000)	0.003 (0.000)		0.001 (0.007)
zonal percentage low income HH	0.261 (0.000)			0.249 (0.000)	0.268 (0.000)
zonal % detached dwelling units	0.166 (0.000)	0.788 (0.000)	0.672 (0.000)		0.232 (0.000)
district population density	0.008 (0.000)	0.005 (0.000)	0.006 (0.000)	0.007 (0.000)	0.005 (0.000)
Accessibility	0.020 (0.000)			0.023 (0.000)	0.019 (0.000)
total employment					

The insignificant and unselected variables are shown by gray and empty cells.
Note: The numbers in parenthesis show the probability value for each parameter.

5.2.1.1 Interpretation of Coefficients

Work:

All of the explanatory variables are significant in terms of their p-values. The zonal percentage low income households is a very strong explanatory variable for the number of work-related tours. As the number of low-income households increases, the number of work-related tours also grows. The average range of produced work-tours for detached dwelling units is less than apartment/row house units. District population density and accessibility variables have little effect on dependent variable.

University:

University tours are best predicted by zonal percentage of detached dwelling units. Traffic zones with higher percentage of detached dwelling units have more university

tours. Detached dwelling units produced more university tours in comparison with apartment/row house units. The detached house variable has a p-value greater than 0.005, therefore, it is an insignificant variable in terms of t-statistics, but it is kept in the final model due to its logical sign and effect on number of university tours.

School:

School tours are also well explained by zonal percentage of detached dwelling units. Existence of more detached houses than apartment/row house units in a traffic zone causes the production of more school tours. Households living in an apartment/row house produce more school tours than those who are living in detached units. Zonal and district population density are not a strong explanatory variable for the number of school tours.

Maintenance and Discretionary:

The zonal percentage of low income households is a very strong explanatory variable for the number of maintenance and discretionary tours. Detached houses produce fewer tours than apartment/row house units for both purposes. Again, zonal and district population density has a small effect on produced tours.

5.2.1.2 Model Validation

The validation of regression model can be explored through the following parameters.

Goodness of fit (R^2):

R^2 is a statistical measure of how well the regression line approximates the real data points and indicates the capability of model to predict the dependent variable based on explanatory variables. Table 5-10 presents the goodness of fit of all models.

Table 5-10: R^2 values for different travel purposes for daily tour production models

	Travel Purpose				
	Work	**University**	**School**	**Maintenance**	**Discretionary**
R^2	0.934	0.911	0.92	0.933	0.932

As it can be observed, all models are capable of predicting household daily tour production rate accurately.

Mean Square Errors (MSE):

MSE was calculated for each model based on equation (2-26) and results were shown in the following table. This value can be used to compare models developed by using different methods.

Table 5-11: MSE values from regression method for tour production models

	Travel Purposes				
	Work	**University**	**School**	**Maintenance**	**Discretionary**
MSE	0.0398	0.0664	0.0660	0.0388	0.0446

Multicollinearity:

Multicollinearity exists when there is a strong correlation between two or more explanatory variables in a regression model. Menard (1995) suggests that the tolerance below 0.1 indicates a serious problem, and below 0.2 indicates a potential problem in terms of multicollinearity problem. Table 5-12 indicates the tolerance value for explanatory variables in all tour production models. The detailed results for checking multicollinearity are provided in the Appendix B.

Table 5-12: Assessing the assumption of no multicollinearity

Variables	Tolerance				
	work	University	School	Maintenance	Discretionary
detached house	0.341	0.300	0.211	0.344	0.316
worker greater than nonworker	0.443	0.529	0.520	0.518	0.508
zonal population density		0.279	0.257		0.241
zonal percentage low income HH	0.331			0.267	0.266
zonal % detached dwelling units	0.103	0.270	0.211		0.096
district population density	0.388	0.244	0.241	0.344	0.220
accessibility	0.084			0.231	0.077
total employment					

As can be seen, the tolerance value indicates serious problem regarding multicollinearity for two explanatory variables, "accessibility" from work tour model and "zonal % detached dwelling units" from discretionary model. These two variables must be eliminated from final models for related travel purpose.

5.2.1.3 Zonal Tour Production Results

The final results of calibrated models show daily household tour production rate for each household type in different traffic zones. These production rates are converted to total production tours for the study area corresponding to equations (4-1) and (4-4) for all 5 travel purposes. The final production tours stimulated from the models are compared with observed production tours from OD Survey results (Table 5-13).

Table 5-13: Tour production comparison between OD survey and regression results

	Travel purpose					
	work	**University**	**School**	**Maintenace**	**Discretionary**	**Total**
Survey	23189	2598	4454	14845	10306	55392
Regression Model	22630	2578	4351	14705	9980	54244
%Diff.	-2%	-1%	-2%	-1%	-3%	-2%

The results show that the regression models are capable of predicting the produced tours precisely. For example, the percentage difference between observed produced tours and predicted tours by regression model for discretionary purpose is only 3% and the difference between total produced tours for all travel purposes is about 2%.

5.2.2 FeedForward Backpropagation Method

Daily household tour production ANN models for all 5 travel purposes were developed by using the same training and testing data sets as used for regression models. The ANN toolbox in MATLAB was used. The strategy and network architecture for all models were discussed in Section 2.3.3 and 4.4.1.3. Although the feedforward backpropagation network is like a black box, it is capable to provide "R" and performance function value (i.e., equivalent to MSE in the regression method).

These values are computed only for initial network in the training process, and it is not given when trained network is used for testing data therefore the MSE value for ANN models have been determined based on equation (2-26) and output results. Results are shown as follows:

Table 5-14: MSE values from ANN methods for tour production models

	Travel purpose				
	work	**University**	**School**	**Maintenance**	**Discretionary**
MSE	0.0331	0.0525	0.0468	0.0423	0.0371

The Appendix shows the output results of all ANN models. The final output results of each model indicate household daily tour rate for each household type. These results were converted, as discussed in Section 5.2.1.3, into total tour production for all traffic zones within the study area and eventually compared with OD survey tour production results (Table 5.15). As shown, all ANN models predict the produced tours very well.

Table 5-15: Tour production comparison between OD survey and ANN results

	Travel purpose					
	work	**University**	**School**	**Maintenance**	**Discretionary**	**Total**
OD survey	23189	2598	4454	14845	10306	55392
ANN	23030	2609	4457	14755	9917	54768
%Diff.	-1%	0.4%	0%	-1%	-4%	-1%

5.2.3 Comparative Results for Tour Production

The results from linear regression model and artificial neural network can be compared by examing the percent of correct estimation and also the value of mean square error for both models.

Table 5-16: Comparison of MSE values for regression and ANN methods in tour production

		Travel Purpose				
	Method	**work**	**University**	**School**	**Maintenance**	**Discretionary**
MSE	**Regression**	0.040	0.066	0.066	0.039	0.045
	ANN	0.0331	0.0525	0.0468	0.0423	0.0371

Table 5-17: Comparison of final estimation of tour production based on regression and ANN methods

Travel purpose	OD Survey	Regression		ANN	
		Correct #	% Difference	Correct #	% Difference
Work	23189	22630	***-2%***	23030	***-1%***
university	2598	2578	***-1%***	2609	***0%***
School	4454	4351	***-2%***	4457	***0%***
maintenance	14845	14705	***-1%***	14755	***-1%***
discretionary	10306	9980	***-3%***	9917	***-4%***
Total	55392	54244	***-2%***	54768	***-1%***

The MSE values for all 5 ANN models are smaller than the calibrated models by regression methods. Thus, this shows that the ANN models are more efficient as estimators than regression models.

In terms of accuracy of predicting the number of tour production, both methods provide good results. But, in relative terms, the ANN models are slightly better.

5.3 Daily Zonal Tour Attraction

5.3.1 Linear Regression Model

The zonal daily tour attraction model was implemented as a linear regression model based on the OD survey aggregate data (i.e., aggregated by traffic zone) for all 5 travel purposes. The regression models were estimated without intercept because no default zonal tour attraction was considered. Regression models were run first with all selected explanatory variables noted in Table 4-9; then, insignificant variables were cancelled from final model. Table 5-18 shows the omitted variables from regression models.

Table 5-18: Omitted variables from daily zonal tour attraction models

	Variable(s)
Work	Total population (p = 0.620)
University	Total population (p = 0.835)
School	Total population (p = 0.203)
Maintenance	Zonal population density (p = 0.199), Total population (p = 0.884)
Discretionary	None

p-values are in parenthesis

Regression models were run in SPSS and model estimation results for 5 travel purposes were summarized in Table 5-19. The detailed results and SPSS output tables were included in Appendix B.

Table 5-19: Regression coefficients by travel purpose for daily zonal tour attraction

Variables	Regression coefficients by travel purpose				
	Work	**University**	**School**	**Maintenance**	**Discretionary**
Zonal population density	-0.730 (0.000)				-0.081 (0.038)
%Detached Dwelling Units	-0.320 (0.002)				
Zonal Shopping GLA				0.343 (0.000)	0.071 (0.000)
District Population Density (persons/ha)	2.199 (0.000)			0.408 (0.000)	0.496 (0.000)
Accessibility index	1.367 (0.000)	0.079 (0.017)	0.211 (0.000)	0.372 (0.000)	0.298 (0.000)
Total employment	0.001 (0.816)				
Total population					0.001 (0.021)
University enrollment		0.030 (0.000)			
school enrollment			0.046 (0.000)		

The insignificant and unselected variables are shown by gray and empty cells.
Note: The numbers in parenthesis show the probability value for each parameter.

Interpretation of coefficients:

Work:

Work tours (attraction) are well explained by the district population density and accessibility. Traffic zones with larger population density attract less work tours, whereas districts with larger population density attract more work trips. More work tours end in traffic zones with a lower percentage of detached dwelling units. Since "total employment" has a great effect on attracted work trips in each traffic zone and has a correct sign, thus, it is kept in the final model.

University and School:

Accessibility and school or university enrollment are significant explanatory variables for the attracted school or university tours.

Maintenance and Discretionary:

Leasable area of stores, accessibility and district population density are strong variables for predicting attracted tours for maintenance or discretionary purposes. People prefer to travel to closer traffic zones for maintenance and discretionary purpose.

5.3.1.1 Model validation

Goodness of fit (R^2) measurement and mean square error (MSE) for all attraction models are presented in Table 5-20 and Table 5-21, respectively.

Table 5-20: R^2 values for tour attraction regression models

	Travel Purposes				
	Work	**University**	**School**	**Maintenance**	**Discretionary**
R^2	0.443	0.987	0.710	0.775	0.634

R^2 values are almost satisfactory for all regression models.

Table 5-21: MSE values for tour attraction regression models

	Travel Purposes				
	Work	**University**	**School**	**Maintenance**	**Discretionary**
MSE	3203.49	106.16	235.51	883.11	267.78

Multicollinearity:

The tolerance value for explanatory variables for all regression models are presented in Table 5-22. Detailed results are shown in the Appendix B.

Table 5-22: Checking Multicollinearity

Variables	**Tolerance**				
	work	**University**	**School**	**Maintenance**	**Discretionary**
zonal population density	0.291				0.328
%detached dwelling units	0.19				
zonal shopping GLA				0.948	0.910
district population density	0.291			0.602	0.334
accessibility	0.181	0.983	0.830	0.594	0.314
total employment	0.256				
total population					0.315
university enrollment		0.983			
school enrollment			0.830		

Because, all tolerance values are larger than 0.1 then there is no violation of Multicollinearity assumption. Therefore, all of the variables can be kept in the final tour attraction models.

5.3.1.2 Zonal Tour Attraction Results

The final results of regression models generate daily zonal tour attraction for each traffic zone. Then, attracted tours are added up together to determine the zonal tour attraction in the study area for all travel purposes.

Table 5-23: Tour attraction comparison between OD Survey and regression results

	Travel purpose					
	work	**University**	**School**	**Maintenance**	**Discretionary**	**Total**
OD survey	23189	2598	4454	14845	10306	55392
Regression Model	21190	2497	4693	14507	10092	52974
Diff.	-9%	-4%	5%	-2%	-2%	-5%

The results indicate that the models are capable of predicting tour attractions. The maximum difference (9%) between observed (attracted) tours and predicted (attracted) tours for the regression model is for work tours.

5.3.2 Artificial Neural Network Method

ANN daily zonal tour attraction models for 5 travel purposes were developed by using the same training and testing data sets as used for the regression models. The ANN toolbox in MATLAB was employed for this purpose. The architecture of the networks was explained in Section 2.3.3 and 4.4.2.3. As discussed earlier, the final output (i.e., results of attraction tours for each traffic zone) is the only information that is obtainable from the ANN method. The MSE value can also be determined based on equation (2-26) and simulation results. These are presented in Table 5-24.

Table 5-24: MSE values from ANN method for attraction tours

	Travel purpose				
	work	**University**	**School**	**Maintenance**	**Discretionary**
MSE	3012.93	1834.28	322.08	1086.55	351.02

The Appendix B shows the detailed output results of ANN models. The results were summarized in Table 5-25 and compared with OD survey tour attraction data.

Table 5-25: Tour attraction comparison between OD survey and ANN results

	Travel purpose					
	work	**University**	**School**	**Maintenance**	**Discretionary**	**Total**
OD survey	23189	2598	4454	14845	10306	55392
ANN	22304	2607	4267	14719	10299	54196
Diff.	-4%	0%	-4%	0%	0%	-2%

The predicted results of ANN models are very close to OD survey results.

5.3.3 Comparative Results for Tour Attraction

The predicted results of tour attraction from the linear regression models and the ANN models can be compared by evaluating the percent correct estimation and the MSE values from both methods.

Table 5-26: Comparison of MSE values from regression and ANN methods in tour attraction

	Method	Travel Purpose				
		Work	**University**	**School**	**Maintenance**	**Discretionary**
MSE	**Regression**	3203.49	106.16	235.51	883.11	267.78
	ANN	3012.93	1834.28	322.08	1086.55	351.02

The MSE values determined by regression approach are smaller than those calculated by ANN models, except for work tours.

Table 5-27: Comparison of final estimation of tour attraction for ANN and regression methods

Travel purpose	**OD Survey**	**Regression**		**ANN**	
		Correct #	% Difference	Correct #	% Difference
work	23189	21190	***-9%***	22304	***-4%***
university	2598	2497	***-4%***	2607	***0%***
school	4454	4693	***5%***	4267	***-4%***
maintenance	14845	14507	***-2%***	14719	***0%***
discretionary	10306	10092	***-2%***	10299	***0%***
total	55392	52974	***-5%***	54196	***-2%***

Table 5-27 presents a comparison of results produced by the models against OD data. For all five proposes, the "percent correct" estimates provided by the ANN models are much better than obtained from regression models.

5.3.4 Balancing of Total Daily Tour Production

The last process of tour generation step is balancing of total daily tour productions and attractions. The total tours are balanced corresponding to equations (4-4), (4-5), and (4-6). Balanced tours and balance factor for each purpose are presented in Table 5-28 and Table 5-29.

Table 5-28: Balanced tours estimated by regression method

Regression Model	Production	22630	2578	4351	14705	9980
	Attraction	21190	2497	4693	14507	10092
Balance total tour	$\sqrt{P \times A}$	**21898**	**2537**	**4519**	**14606**	**10036**
Production balance factor	B/P	0.97	0.98	1.04	0.99	1.01
Attraction balance factor	B/A	1.03	1.02	0.96	1.01	0.99

Table 5-29: Balanced tours estimated by ANN method

ANN	Production	22630	2578	4351	14705	9980
	Attraction	22304	2607	4267	14719	10299
Balance total tour	$\sqrt{P \times A}$	**22466**	**2592**	**4309**	**14712**	**10138**
Production balance factor	B/P	0.99	1.01	0.99	1.00	1.02
Attraction balance factor	B/A	1.01	0.99	1.01	1.00	0.98

5.4 Summary

The comparative results for household car ownership model produced by two methods are presented in Table 5-7.

The final results on tour generation (i.e., daily household tour production and daily zonal tour attraction), for both methods are summarized and compared in the following tables.

Table 5-30: Summary results of tour generation from regression method

		Travel Purpose					
		work	**Univ**	**School**	**Maintenance**	**Discretionary**	**Total**
survey	Production	23189	2598	4454	14845	10306	55392
	Attraction	23189	2598	4454	14845	10306	55392
Regression Model	Production	22630	2578	4351	14705	9980	54244
	Attraction	21190	2497	4693	14507	10092	52974
%Diff.	Production	**-2%**	**-1%**	**-2%**	**-1%**	**-3%**	**-2%**
	Attraction	**-9%**	**-4%**	**5%**	**-2%**	**-2%**	**-5%**
MSE	Production	0.0398	0.0664	0.0660	0.0388	0.0446	
	Attraction	3203.49	106.16	235.51	883.11	267.78	
Balance total tour	$\sqrt{P \times A}$	21898	2537	4519	14606	10036	
Production balance factor	B/P	0.9675	0.9842	1.0386	0.9932	1.0056	
Attraction balance factor	B/A	1.0335	1.0161	0.9629	1.0068	0.9944	

Table 5-31: Summary results of tour generation from ANN method

	Type	Travel purpose					
		work	**Univ**	**School**	**Maint**	**Disc**	**Total**
survey	Production	23189	2598	4454	14845	10306	55392
	Attraction	23189	2598	4454	14705	10306	55252
NN	Production	22630	2578	4351	14705	9980	54244
	Attraction	22304	2607	4267	14719	10299	54196
%Diff.	Production	**-2%**	**-1%**	**-2%**	**-1%**	**-3%**	**-2%**
	Attraction	**-4%**	**0%**	**-4%**	**0%**	**0%**	**-2%**
MSE	Production	0.0331	0.0525	0.0468	0.0423	0.0371	
	Attraction	3012.9	1834.28	322.08	1086.55	351.02	
Balance total tour	$\sqrt{P \times A}$	22466	2592	4309	14712	10138	
Productio n balance factor	B/P	0.9928	1.0056	0.9903	1.0005	1.0159	
Attraction balance factor	B/A	1.0073	0.9944	1.0098	0.9995	0.9844	

6. Conclusions and Recommendations

6.1 Conclusions

This study aimed at developing a tour activity-based model to overcome the weaknesses of current trip-based conventional models. In doing so, first chained trips traveled from home to one or more activity locations and back home again, were all combined into a tour. One of the activity locations was considered as a primary or most important activity of the day and the other stops were considered as secondary stops.

The tours have two directions, from home to the primary location and vice versa. Time of day, mode of travel, existence of stop, and destination of each direction of tours are also specified for each household in all traffic zones. The defined tours enable the use of disaggregate models in combination with aggregate models in tour generation process. Tour production is a disaggregate model and it considers household attributes including socio-economic and land-use variables. On the other hand, tour attraction is an aggregate model, given that attributes of destinations are not surveyed at the same level of detail as the location of production. Another feature of tour-based models is that it eliminates the unnecessary stops in both direction and just considers the main purpose of chained trips in a tour.

Another objective of this research was to explore the relative effectiveness of candidate methodologies. The probabilistic neural network model was compared against the nested logit model in solving the classification problem in household car ownership model and feedforward backpropagation neural network model was studied as a potential replacement for the linear regression model in developing the tour generation model.

Major conclusions of this study have been summarized below. They were separated into two sections: i) conclusions about the household car ownership model ii) conclusions regarding to tour generation models.

6.1.1 Conclusions Regarding Household Car Ownership Model

- Final results presented in Table 5-7 show that the PNN model produces slightly better predictions of the number of cars available to household than the nested logit model.
- The PNN model predicts the alternative having a small sample population more accurately than the nested logit model. In this research, PNN predicts correctly about 50% and 25% of zero car and 3+ cars alternatives, respectively. On the other hand, the nested logit model only predicts 15% and 5% of these alternatives correctly. It should be noted that the zero car and 3+ cars alternatives comprise only 10% and 8.6% of total cars owned, respectively.
- It is a challenge to choose the best structure (i.e., number of levels and how to categorize the alternatives in each level) for nested logit model, whereas, the model structure of PNN (i.e., number of neurons in layers, transfer functions, etc) can be found by following the guidelines described in the methodology of this research. Development of input variables for nested logit model in SAS is complicated. To begin with, all of the observations (i.e., households) should be expanded 4 times (i.e., equal to the number of choice set) for each alternative. Then the dependent variable (e.g., decision) would be defined so that it takes the value of 1 when the household chooses that alternative, otherwise it is zero. The other mandatory variable is the identification variable for each household. It has to start from one and goes up in sequence. It also should remain the same for each household. However, these

variables are not required in the PNN model and the dependent variable is the same as the number of available cars to each household.

- In nested logit model, the designation of the type of variable in terms of chooser-specific or choice-specific is important. The SAS software is not able to distinguish chooser-specific variables, thus these variables should be converted into alternative-specific variables according to the process discussed in Section 4.3.1.2. Nevertheless, different types of variables are not important in the PNN models.
- Finally, the SAS output results of nested logit model just demonstrate the probability of choosing an alternative in a choice set. Then, the alternative with highest probability should be considered as a selected alternative. But, PNN outputs directly indicate the number of available cars to each household.

6.1.2 Conclusions Regarding Tour Generation Models

- In tour production models, the regression and ANN methods produced approximately the same results (Table 5-17). But, mean square error (MSE) values in ANN models were smaller than regression models, except for maintenance tours.
- In tour attraction models, the ANN models produce better results in comparison with regression models in terms of percent correct estimation for all travel purposes (Table 5-27). None of the tour attraction models are superior in terms MSE values.
- Tour production models statistically fit much better than tour attraction. Tour production model was estimated based on 22,105 households while the attraction model was based on 556 traffic zones. In here, the point which needs stressing is the tour production model is a disaggregate model whereas tour attraction is an aggregate model.

- Finding the optimum architecture for ANN models is challenging. The parameters such as number of neurons in hidden layer(s), number of hidden layers and type of transfer function for each layer may be fluctuating, whereas, linear regression models do not have such problems.

6.2 Recommendations

This research advances two parts of the overall field of travel demand modeling, namely car ownership, and tour generation. Further research on demand prediction techniques is required. This section outlines the recommendations for further research, and suggestions for improving the entire travel demand modeling process. In further research, the choice of the most suitable modeling technique (i.e., statistical methods vs. ANN models) should be investigated.

- **Improving household segmentation**: in this study, household are divided into different types according to two variables: i) number of workers versus number of non-workers ii) dwelling type. In addition to these variables, however, other variables such as size of household, and number of cars or other dimensions can also be considered.
- **Adding time of day choice model**: the time of day information for each direction of a tour is defined in the database, thus another model for choosing the time of day period can be implemented for both tour attraction and production. Therefore, combination of time of day period of all produced and attracted tours can be determined.

- **Tour mode choice model**: mode of travel in each direction for all tours are defined, therefore, another PNN model can be implemented to predict the tour mode choice in each direction.
- **PNN method**: is an excellent and efficient tool in travel demand modeling. A comprehensive PNN model can be developed to predict simultaneously the choice of time of day, destination and mode of travel. The number of alternatives in the choice set will be large but still it seems to be practical.

References

Allen, W., D.Liu, & S.Singer. (1993). Accessibility Measures of U.S.Metropolitan Areas. *Transportation Reasearch* , Vol. 27B, No. 6, pp. 439-449.

Ben-Akiva, M. (1973). *Structure of Passenger Travel Demand Models.* MIT, Cambridge, Mass.: Ph.D. Dissertation. Department of Civil Engineering.

Ben-Akiva, M., & Lerman, S. R. (1985). *Discrete Choice Analysis: Theory and Application to Travel Demand.* Cambridge, Massachusetts: MIT Press.

Bierlaire, M. (1997). Operations research and decision aid methodologies in traffic and transportation management. *Discrete Choice Models, NATO Advanced Studies Institute.* Balatonfured, Hungary.

Bowman, J. (1995). *Activity Based Travel Demand Model System with Daily activity Schedules.* M.S. Thesis. Massachusetts Institute of Technology.

Bowman, J. L., & Ben-Akiva, M. (1996). Activity-Based Travel Forecasting. *Summary, Recommendations and Compendium of Papers from the Activity-Based Travel Forecasting Conference, New Orleans* , 3-36.

Bowman, J., & Ben-Akiva, M. (2000). Activity-Based Disaggregate Travel Demand Model System with Activity schedules. *Transportation Research Part A: Policy and Practice , 35* (1), 1-28.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometritcs: Methods and Application.* New York: Cambridge University Press.

Celikoglu, H. B. (2006). Application of radial basis function and generalized regression neural networks in non-linear utility function specification for travel mode choice modelling. *Mathematical and Computer Modelling* , 640-658.

Daganzo, C. F. (1979). *Multinomial Probit: The theory and its application to demand forecasting.* New York: Academic Press.

Demuth, H., & Beale, M. (2001). *Neural Network Toolbox: For Use with MATLAB.* Massachusetts: Math Work Inc.

Dickey, J. W. (1983). *Metropolitan Transportation Planning, 2nd Edition.* Washington D.C.: Hemisphere.

Domencich, T. A., & McFadden, D. (1975). *Urban travel demand: a behavioral analysis.* Amsterdam: North-Holland Publishing.

Dreyfus, G. (2005). *Neural Networks: Methodology and Applications.* Paris: Springer.

Estrella, A. (1998). A New Measure of Fit for Equation with Dichotomous dependent Variables. *Business and Economic Statistics* , 198-205.

Fausett, L. (1994). *Fundamentals of Neural Netwroks: architectures, algorithms, and applications.* New Jersey: A Paramount Communications Company Englewood Cliffs.

Field, A. (2005). *Discovering Statistics Using SPSS.* London: SAGE Publications Ltd.

Heiss, F. (2002). Structural choice analysis with nested logit models. *The Stata Journal* , 227-252.

Kartam, N., Flood, I., & Garrett, J. H. (1997). *Artificial Neural Networks for Civil Engineers: Fundamentals and Applications.* New York: ASCE.

Kim, Y. S. (2008). Comparison of the decision tree, artificial neural network, and linear regression methods based on the number and types of independent variables and sample size. *Expert Systems with Applications* , 34(2), 1227-1234.

Koppelman, F. S., & Wen, C.-h. (1998). Alternative Nested Logit Model: Structure, Properties and Estimation. *Elsevier Science* , 289-298.

Manski, C. F. (1973). *The Analysis of Qualitative Choice.* United States: Ph.D. dissertation, Massachusetts Institute of Technology.

Manski, C. F. (1977). The Structure of Random Utility Models. *Theory and Decision 8* , 229-254.

McCormic Rankin Corporation/ Parsons Brinckerhoff. (April 2008). *TRANS Model Redevelopment.* http://www.ncr-trans-rcn.ca/uploadedFiles/resources/TRANS_ModRedReptFINAL.pdf.

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka, *Frontiers in econometrics* (pp. 105-142). New York: Academic Press.

McFadden, D. (1981). Econometric Models of Probabilistic Choice. In C. F. Manski, & D. Mc Fadden, *Structural Analysis of Discrete Data with Econometric Applications.* Cambridge: MIT Press.

McFadden, D. (1978). Modelling the Choice of Residntial Location. In A. Karlqvist, L. Lundqvist, F. Snickars, & J. W. Weibull, *Spatial Interaction Theory and Residential Location* (pp. 75-96). Amesterdam: North Holland.

McNally, M. G. (2000). *Handbook of Transport Modelling Edited by D.A. Hensher and K.J Button.* Elsevier Science Ltd.

Menard, S. (1995). *Applied logistic regression analysis.* Thousand Oaks, CA: Sage.

Meyer, M., & Miller, E. (2001). *Urban Transportation Planning: A decision-Oriented Approach.* New York: McGraw-Hill.

Mohammadian, A., & Miller, E. J. (2002). Nested Logit Models and Artificial Neural Netwroks for Predicting Household Automobile Choices. *Transportation Research Record* , 92-100.

Ortuzar, J. D., & Willumsen, L. G. (1994). *Modeling Transport.* West Sussex, England: John Wiley & Sons Ltd.

Pas, E. (1996). Recent Advances in Activity-Based Travel Demand Modeling. *Activity-Based travel Forecasting.* New Orleans.

Revelt, D., & Train, K. (1998). Mixed Logit with Repeated Choice of Appliance Efficiency. *Review of Economics and Statistics* , 647-657.

Ritchie, S. G., & Cheu, R. L. (1993). Simulation of Freeway Incident Detection Using Artificial Neural Network. *Transportation Research: Part C,* , 1(3), 203-217.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning Representation by Back-Propagation Errors. *Nature* , 533-536.

Shiftan, Y. (1998). Practical Approach to Model Trip Chaining. *Transportation Research Record , 1645*, 17-23.

Silberhorn, N., Boztug, Y., & Hildebrandt, L. (2006). Estimation with the Nested Logit Model: Specifications and Software Particularities. *SFB 649* .

Train, K. (1980). A Structural Logit Model of Auto Ownership and Mode Choice. *Review of Economics Studies* , 357-370.

Train, K. (2003). *Discrete Choice Methods with Simulation.* Cambridge University Press.

TRANS Committee. (December 2006). *2005 Origin-Destination Survey - Summary of Results.* Ottawa: www.ODSurvey.ca.

Vovsha, P., Peterson, E., & Donnelly, R. (2004). Model for Allocation of Maintenance Activities to Household Members. *Transportation Research Record: Journal of the Transportation Rsearch Board, No. 1894, TRB, National Research Council, Washington, D.C.* , 170-179.

Zhou, Q., Lu, H.-P., & Xu, W. (2007). New Travel Demand Models with Backpropagation Network. *Third International Conference on Neural Computation (ICNC), IEEE* .

Appendix A

MATLAB Codes:

Note: Adding all MATLAB codes for all models and data preparation process required many pages. Therefore, similar codes for travel purposes were not denoted.

Data Preparation

```
% This code allocates the TAZ %detached dwelling units to related HH based
% on the carchoice alternative. No car is detemined as the reference case
% so detached% dwelling type for no car is zero for all HHs.
E = load('carchoice.txt');
F = load('detach.txt');
for i=1:556
  for j=1:95472
    if F(i,1)==E(j,1) && E(j,7)==2
      E(j,8)=F(i,2);
    elseif F(i,1)==E(j,1) && E(j,7)==3
      E(j,9)=F(i,2);
    elseif F(i,1)==E(j,1) && E(j,7)==4
      E(j,10)=F(i,2);
    end
  end
end
N = E(:,[1 2 8 9 10]);
save ('detachedHH2.txt','N','-ascii');
A=load('carchoice.txt');
B=load('access.txt');
for i=1:556
  for j=1:95472
    if B(i,1)==A(j,1)&& A(j,7)==2
      A(j,8)=B(i,2);
    elseif B(i,1)==A(j,1)&& A(j,7)==3
      A(j,9)=B(i,2);
    elseif B(i,1)==A(j,1)&& A(j,7)==4
```

```
        A(j,10)=B(i,2);
      end
    end
end
N = A(:,[1 2 8 9 10]);
save ('accessibility.txt','N','-ascii');
**********************************************************************
% This code allocates the TAZ population density to related HH based
% on the carchoice alternative. No car is determined as the reference case
% so it is zero for all HH with no car.
A=load('carchoice.txt');
B=load('pop-den.txt');
for i=1:556
   for j=1:95472
      if B(i,1)==A(j,1)&& A(j,7)==2
         A(j,8)=B(i,2);
      elseif B(i,1)==A(j,1)&& A(j,7)==3
         A(j,9)=B(i,2);
      elseif B(i,1)==A(j,1)&& A(j,7)==4
         A(j,10)=B(i,2);
      end
   end
end
N = A(:,[1 2 8 9 10]);
save ('zonepopden2.txt','N','-ascii');
```

Tour Construction

```
% Seed matrix H composed of 6 columns HH-id, origin, destination, Purpose,
% TOD, and mode in retrospect.
H = load('AllTrips2.txt');
TourCon = [];
```

```
Hid = unique(H(:,1));
for i = 1:size(H,1)
   H(i,7) = find(HHid==H(i,1));
end
id = 1;
for n = 1:size(H,1)
   m=1;
   for j = 1:size(H,1)
      if H(j,7) == id
      items(m,(1:6)) = H(j,(1:6));
      m = m + 1;
      end
   end
uniqItems = unique(items(:,3));
uniqItems(:,2:4) = 10;
modH = items;
for i = 1:size(items,1)
   for j = 2:3
      modH(i,j) = find(uniqItems==items(i,j));
   end
end
matrix = sparse(modH(:,2),modH(:,3),ones(1,size(modH,1)));
for i = 1:size(uniqItems,1)
   for j = 1:size(items,1)
      if uniqItems(i,1)==items(j,3) && items(j,4)==6
         uniqItems(i,(2:4)) = items(j,(4:6)) ; break, end
      if uniqItems(i,1)==items(j,3) && uniqItems(i,2)>=items(j,4) && uniqItems(i,2)~=6
         uniqItems(i,(2:4)) = items(j,[4 5 6]);
      if uniqItems(i,2)==6, break, end
      end
```

```
    end
end
 [home,c] = find(uniqItems(:,2)==6);
counter1 = 0;
while(~graphisdag(matrix))
    counter1 = counter1 + 1;
    [i,j,w] = find(matrix);
    a = find(i==home);
    if isempty(a), break, end
    finalList(counter1,1) = home;
    pos = home;
    check = false;
    counter2 = 1;
    while(pos~=home || check==false)
        counter2 = counter2 + 1;
        check = true;
        ind = find(i==pos,1);
        if(isempty(ind))
        %finalList(counter1,counter2) = home;
            finalList(counter1,:) = 0;
            counter1 = counter1 - 1; break
        end
        lastpos = pos;
        pos = j(ind);
        if(counter2 > 2 && pos == finalList(counter1,counter2-2))
            tempInd = find(i==lastpos);
            if(size(tempInd,1) > 1)
                ind = tempInd(2);
                pos = j(ind);
            end
```

```
      end
      if (w(ind) > 0)
         w(ind) = w(ind)- 1;
      end
      matrix = sparse(i,j,w);
      finalList(counter1,counter2) = pos;
      if (finalList(counter1,counter2)==finalList(counter1,counter2-1)) &&
(finalList(counter1,counter2)==home)
         finalList(counter1,:) = [];
         counter1 = counter1 - 1; break
      end
      end
   end
   matrix = sparse(i,j,w);
end
finalList;
finalTour = finalList;
for i= 1:size(finalList,1)
   for j=1:size(finalList,2)
      if finalList(i,j) ~= 0
      finalTour(i,j) = uniqItems(finalList(i,j));
      end
   end
end
finalTour;
tours = zeros(size(finalTour,1),10);
for i = 1:size(finalTour,1)
   for j = 1:size(uniqItems,1)
   tours(i,1) = items(1);
      if uniqItems(j,2)==6
```

```
      tours(i,2) = uniqItems(j,1);
      end
    end
end
H2 = items;
for i = 1:size(finalTour,1)
  [r c v] = find(finalTour(i,:));
  for k = 2:size(v,2)-1
    if tours(i,3)~=0, break, end
      for j = 1:size(uniqItems,1)
        if v(k)==uniqItems(j,1) && uniqItems(j,2)==1
        tours(i,3) = uniqItems(j,1); break
        end
      end
  end
  for k = 2:size(v,2)-1
    if tours(i,3)~=0, break, end
      for j = 1:size(uniqItems,1)
        if v(k)==uniqItems(j,1) && uniqItems(j,2)==2
        tours(i,3) = uniqItems(j,1); break
        end
      end
  end
  for k = 2:size(v,2)-1
    if tours(i,3)~=0, break, end
      for j = 1:size(uniqItems,1)
        if v(k)==uniqItems(j,1) && uniqItems(j,2)==3
        tours(i,3) = uniqItems(j,1); break
        end
      end
```

```
end
for k = 2:size(v,2)-1
  if tours(i,3)~=0, break, end
    for j = 1:size(uniqItems,1)
      if v(k)==uniqItems(j,1) && uniqItems(j,2)==4
      tours(i,3) = uniqItems(j,1); break
      end
    end
end
for k = 2:size(v,2)-1
  if tours(i,3)~=0, break, end
    for j = 1:size(uniqItems,1)
      if v(k)==uniqItems(j,1) && uniqItems(j,2)==5
      tours(i,3) = uniqItems(j,1); break
      end
    end
end
DesInd = find(finalTour(i,:)==tours(i,3),1,'first');
homInd = find(finalTour(i,:)==tours(i,2),1,'last');
if DesInd > 2
  tours(i,5) = 1;
else
  tours(i,5) = 0;
end
if homInd - DesInd > 1
  tours(i,6) = 1;
else
  tours(i,6) = 0;
end
```

```
   ind1 = find(v==tours(i,3),1,'first');
   ind2 = find(v==tours(i,2),1,'last');
   for m = 1:size(H2,1)
      if H2(m,3)==v(ind1) && H2(m,2)==v(ind1-1)
         tours(i,7) = H2(m,5);
         tours(i,9) = H2(m,6);
         tours(i,4) = H2(m,4);
         H2(m,:) = 0;
      end
      if H2(m,3)==v(ind2) && H2(m,2)==v(ind2-1)
         tours(i,8) = H2(m,5);
         tours(i,10) = H2(m,6);
         H2(m,:) = 0;
      end
      if tours(i,7:10)~=0, break, end
   end
end
TourCon = [TourCon;tours];
id = id + 1
if id~=H(:,7), break, end
clear items H2 uniqItems modH matrix finalList finalTour tours i j w
end
save ('AllTours.txt','TourCon','-ascii');
```

****

PNN model for HH car ownership

```
P = load('CarchoiceInput2.txt');
P = P';
Tc = load('CarchoiceTarget2.txt');
Tc = Tc';
T = ind2vec(Tc);
```

```
net2 = newpnn(P,T);

Y = sim(net2,P);

save net2

Yc = vec2ind(Y);

Yc = Yc';

save('TrainCarchoice2.txt','Yc','-ascii');

P2 = load('TestInp12.txt');

P2 = P2';

Y = sim(net2,P2);

Yc = vec2ind(Y);

Yc = Yc';

save('TestOut12.txt','Yc','-ascii');

P3 = load('TestInp2.txt');

P3 = P3';

Y = sim(net2,P3);

Yc = vec2ind(Y);

Yc = Yc';

save('TestOut22.txt','Yc','-ascii');
```

****

ANN Tour Production (work)

```
p = load('worktrain-inp.txt');

t = load('worktrain-target.txt');

p = p'; t = t';

 [pn,pp1] = mapstd(p);

[tn,tp] = mapstd(t);

[ptrans,pp2] = processpca(pn,0.001);

[R,Q] = size(ptrans);

iitst = 2:4:Q;

iival = 4:4:Q;

iitr = [1:4:Q 3:4:Q];
```

```
validation.P = ptrans(:,iival);

validation.T = tn(:,iival);

testing.P = ptrans(:,iitst);

testing.T = tn(:,iitst);

ptr = ptrans(:,iitr);

ttr = tn(:,iitr);

worknet = newff(minmax(ptr),[200 1],{'tansig' 'purelin'},'trainlm');

worknet.trainParam.goal = .01;

worknet.trainParam.epochs = 300;      % Show intermediate results every five iterations.

 [worknet,tr]=train(worknet,ptr,ttr,[],[],validation,testing);

save worknet;
```

Appendix B

SPSS outputs

Tour Production

SPSS Output Tables

Regression: Work

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	.967[a]	.934	.934	.1984578

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	405.709	6	67.618	1716.828	.000[a]
	Residual	28.555	725	.039		
	Total	434.263[b]	731			

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	detached house (dummy)	**-.062**	.016	-.065	-3.968	.000
	worker# greater or equal than non-worker in HH	**-.066**	.015	-.065	-4.555	.000
	zonal percentage low income HH	**.261**	.033	.133	8.006	.000
	zonal % detached dwelling units	**.166**	.031	.161	5.426	.000
	district population density	**.008**	.000	.246	16.108	.000
	accessibility between zones	**.020**	.001	.639	19.420	.000

Regression: University

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	.954[a]	**.911**	.909	.2674108

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	178.931	5	35.786	500.446	.000[a]
	Residual	17.520	245	.072		
	Total	196.451[b]	250			

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	detached house (dummy)	**.025**	.038	.023	.647	.518
	worker# greater or equal than non-worker in HH	**.106**	.034	.082	3.142	.002
	zonal population density	**.004**	.001	.257	7.121	.000
	zonal % detached dwelling units	**.766**	.049	.570	15.514	.000
	district population density	**.005**	.001	.174	4.501	.000

Regression: School

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	.959[a]	**.920**	.919	.2287352

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	218.701	5	43.740	836.018	.000[a]
	Residual	19.044	364	.052		
	Total	237.746[b]	369			

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	detached house (dummy)	**-.048**	.030	-.052	-1.609	.109
	worker# greater or equal than non-worker in HH	**.176**	.023	.155	7.515	.000
	zonal population density	**.003**	.000	.187	6.380	.000
	zonal % detached dwelling units	**.672**	.034	.638	19.745	.000
	district population density	**.006**	.001	.178	5.881	.000

Regression: Maintenance

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	.966[a]	**.933**	.932	.1993501

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	417.468	5	83.494	2100.974	.000[a]
	Residual	30.203	760	.040		
	Total	447.671[b]	765			

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	detached house (dummy)	**-.088**	.015	-.092	-5.719	.000
	worker# greater or equal than non-worker in HH	**.072**	.014	.066	5.061	.000
	zonal percentage low income HH	**.270**	.040	.124	6.801	.000
	district population density	**.007**	.000	.225	14.011	.000
	accessibility between zones	**.023**	.001	.730	37.265	.000

Regression: Discretionary

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	.965[a]	**.932**	.931	.2068898

ANOVA[c,d]

Model		Sum of Squares	Df	Mean Square	F	Sig.
1	Regression	425.947	7	60.850	1421.606	.000[a]
	Residual	31.246	730	.043		
	Total	457.194[b]	737			

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	detached house (dummy)	**-.122**	.017	-.125	-7.279	.000
	worker# greater or equal than non-worker in HH	**.042**	.015	.038	2.802	.005
	zonal population density	**.001**	.000	.053	2.716	.007
	zonal percentage low income HH	**.268**	.040	.125	6.650	.000
	zonal % detached dwelling units	**.232**	.034	.216	6.935	.000
	district population density	**.005**	.001	.169	8.202	.000
	accessibility between zones	**.019**	.001	.580	16.612	.000

Checking Multicollinearity

Work:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	detached house (dummy)	-.062	.016	-.065	-3.968	.000	**.341**	2.932
	worker# greater or equal than non-worker in HH	-.066	.015	-.065	-4.555	.000	**.443**	2.255
	zonal percentage low income HH	.261	.033	.133	8.006	.000	**.331**	3.021
	zonal % detached dwelling units	.166	.031	.161	5.426	.000	**.103**	9.704
	district population density	.008	.000	.246	16.108	.000	**.388**	2.574
	accessibility between zones	.020	.001	.639	19.420	.000	**.084**	11.928

University:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	detached house (dummy)	.025	.038	.023	.647	.518	**.300**	3.332
	worker# greater or equal than non-worker in HH	.106	.034	.082	3.142	.002	**.529**	1.890
	zonal population density	.004	.001	.257	7.121	.000	**.279**	3.584
	zonal % detached dwelling units	.766	.049	.570	15.514	.000	**.270**	3.708
	district population density	.005	.001	.174	4.501	.000	**.244**	4.098

School:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF

1	detached house (dummy)	-.048	.030	-.052	-1.609	.109	**.211**	4.740
	worker# greater or equal than non-worker in HH	.176	.023	.155	7.515	.000	**.520**	1.922
	zonal population density	.003	.000	.187	6.380	.000	**.257**	3.893
	zonal % detached dwelling units	.672	.034	.638	19.745	.000	**.211**	4.738
	district population density	.006	.001	.178	5.881	.000	**.241**	4.143

Maintenance:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	detached house (dummy)	-.088	.015	-.092	-5.719	.000	**.344**	2.905
	worker# greater or equal than non-worker in HH	.072	.014	.066	5.061	.000	**.518**	1.929
	zonal percentage low income HH	.270	.040	.124	6.801	.000	**.267**	3.751
	district population density	.007	.000	.225	14.011	.000	**.344**	2.910
	accessibility between zones	.023	.001	.730	37.265	.000	**.231**	4.328

Discretionary:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	detached house (dummy)	-.122	.017	-.125	-7.279	.000	**.316**	3.162
	worker# greater or equal than non-worker in HH	.042	.015	.038	2.802	.005	**.508**	1.969
	zonal population density	.001	.000	.053	2.716	.007	**.241**	4.141
	zonal percentage low income HH	.268	.040	.125	6.650	.000	**.266**	3.756
	zonal % detached dwelling units	.232	.034	.216	6.935	.000	**.096**	10.390
	district population density	.005	.001	.169	8.202	.000	**.220**	4.538

Tour Attraction

SPSS Output Tables

Regression: Work

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	.665[a]	.443	.435	57.561

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	912920.210	5	182584.042	55.106	.000[a]
	Residual	1149723.790	347	3313.325		
	Total	2.063E6	352			

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	zonal population density	**-.730**	.152	-.358	-4.812	.000
	zonal % detached dwelling units	**-.320**	.104	-.283	-3.077	.002
	district population density	**2.199**	.238	.687	9.251	.000
	Accessibility between zones	**1.367**	.293	.440	4.666	.000
	total employment	**.001**	.004	.018	.233	.816

Regression: University

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	.993[a]	**.987**	.987	8.727

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	838337.810	2	419168.905	5503.544	.000[a]
	Residual	11272.190	148	76.163		
	Total	849610.000[b]	150			

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Accessibility between zones	**.079**	.033	.023	2.414	.017
	university enrollment	**.030**	.000	.990	103.661	.000

Regression: School

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	.843[a]	.710	.707	17.227

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	141016.824	2	70508.412	237.575	.000[a]
	Residual	57576.176	194	296.784		
	Total	198593.000[b]	196			

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	Accessibility between zones	**.211**	.055	.163	3.841	.000
	school enrollment	**.046**	.003	.762	17.957	.000

Regression: Maintenance

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	.880[a]	**.775**	.773	26.736

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	839629.607	3	279876.536	391.543	.000[a]
	Residual	243748.393	341	714.805		
	Total	1.083E6	344			

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	zonal shopping GLA	.343	.012	.759	28.748	.000
	district population density	.408	.076	.178	5.364	.000
	Accessibility between zones	.372	.076	.163	4.895	.000

Regression: Discretionary

Model Summary

Model	R	R Square[b]	Adjusted R Square	Std. Error of the Estimate
1	.796[a]	**.634**	.628	16.396

ANOVA[c,d]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	159599.379	5	31919.876	118.739	.000[a]
	Residual	92206.621	343	268.824		
	Total	251806.000[b]	348			

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	zonal population density	**-.081**	.039	-.119	-2.083	.038
	zonal shopping GLA	**.071**	.008	.291	8.490	.000
	district population density	**.496**	.061	.458	8.097	.000
	Accessibility between zones	**.298**	.064	.269	4.622	.000
	total population	**.001**	.001	.135	2.323	.021

Checking Multicollinearity

Work:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	zonal population density	-.730	.152	-.358	-4.812	.000	**.291**	3.439
	zonal % detached dwelling units	-.320	.104	-.283	-3.077	.002	**.190**	5.273
	district population density	2.199	.238	.687	9.251	.000	**.291**	3.438
	Accessibility between zones	1.367	.293	.440	4.666	.000	**.181**	5.533
	total employment	.001	.004	.018	.233	.816	**.256**	3.913

University:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Accessibility between zones	.079	.033	.023	2.414	.017	**.983**	1.018
	university enrollment	.030	.000	.990	103.661	.000	**.983**	1.018

School:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	Accessibility between zones	.211	.055	.163	3.841	.000	**.830**	1.205
	school enrollment	.046	.003	.762	17.957	.000	**.830**	1.205

Maintenance:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	zonal shopping GLA	.343	.012	.759	28.748	.000	.948	1.055
	district population density	.408	.076	.178	5.364	.000	.602	1.661

Discretionary:

Coefficients[a,b]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Collinearity Statistics	
		B	Std. Error	Beta			Tolerance	VIF
1	zonal population density	-.081	.039	-.119	-2.083	.038	**.328**	3.045
	zonal shopping GLA	.071	.008	.291	8.490	.000	**.910**	1.099
	district population density	.496	.061	.458	8.097	.000	**.334**	2.992
	Accessibility between zones	.298	.064	.269	4.622	.000	**.314**	3.181
	total population	.001	.001	.135	2.323	.021	**.315**	3.177

Appendix C

SAS Code and Results

SAS Code

```
proc mdc data=carchoice maxit=300 outest=a;
model decision= car1w0 car2w0 car3w0 car1w1 car2w1
car3w1 car1w2 car2w2 car3w2 car1w3 car2w3 car3w3 nww0c1
nww0c2 nww0c3 nww1c1 nww1c2 nww1c3 nww2c1 nww2c2 nww2c3
det_home_c1 det_home_c2 det_home_c3 zone_pop_den_c1 zone_pop_den_c2
zone_pop_den_c3 lowinc_c1 lowinc_c2 lowinc_c3 zone_det_home_c1
zone_det_home_c2 zone_det_home_c3 dist_pop_den_c1
dist_pop_den_c2 dist_pop_den_c3 gla_c1 gla_c2 gla_c3/ type=nlogit samescale
choice=(carchoice 1 2 3 4);
id ID;
output out=probdata pred=p;
utility u(1,)= car1w0 car2w0 car3w0 car1w1 car2w1
car3w1 car1w2 car2w2 car3w2 car1w3 car2w3 car3w3 nww0c1
nww0c2 nww0c3 nww1c1 nww1c2 nww1c3 nww2c1 nww2c2 nww2c3
det_home_c1 det_home_c2 det_home_c3 zone_pop_den_c1 zone_pop_den_c2
zone_pop_den_c3 lowinc_c1 lowinc_c2 lowinc_c3 zone_det_home_c1
zone_det_home_c2 zone_det_home_c3 dist_pop_den_c1
dist_pop_den_c2 dist_pop_den_c3 gla_c1 gla_c2 gla_c3;
nest level(1)=(1@1,2@2,3 4@3),
        level(2)=(1@1,2 3@2),
        level(3)=(1 2@1);
run;
```

SAS output

The SAS System

The MDC Procedure

Nested Logit Estimates

Algorithm converged.

Model Fit Summary

Dependent Variable	decision
Number of Observations	23868
Number of Cases	95472
Log Likelihood	-23067
Maximum Absolute Gradient	4856
Number of Iterations	242
Optimization Method	Dual Quasi-Newton
AIC	46222
Schwarz Criterion	46577

Discrete Response Profile

Index	carchoice	Frequency	Percent
0	1	2387	10.00
1	2	10277	43.06
2	3	9133	38.26
3	4	2071	8.68

Goodness-of-Fit Measures

Measure	Value	Formula
Likelihood Ratio (R)	20042	2 * (LogL - LogL0)
Upper Bound of R (U)	66176	- 2 * LogL0
Aldrich-Nelson	0.4564	R / (R+N)
Cragg-Uhler 1	0.5682	1 - exp(-R/N)
Cragg-Uhler 2	0.606	(1-exp(-R/N)) / (1-exp(-U/N))
Estrella	0.6322	1 - (1-R/U)^(U/N)
Adjusted Estrella	0.6304	1 - ((LogL-K)/LogL0)^(-2/N*LogL0)
McFadden's LRI	0.3029	R / U
Veall-Zimmermann	0.6211	(R * (U+N)) / (U * (R+N))

N = # of observations, K = # of regressors

The SAS System
The MDC Procedure
Nested Logit Estimates

Parameter Estimates

Parameter	DF	Estimate	Standard Error	t Value	Approx Pr > |t|
car1w0_L1	1	-0.4088	0.2597	-1.57	0.1156
car2w0_L1	1	-2.0485	0.5318	-3.85	0.0001
car3w0_L1	1	-4.4996	0.6355	-7.08	<.0001
car1w1_L1	1	0.5604	0.1594	3.52	0.0004
car2w1_L1	1	-0.5480	0.2675	-2.05	0.0405
car3w1_L1	1	-2.2846	0.2862	-7.98	<.0001
car1w2_L1	1	1.2740	0.2946	4.32	<.0001
car2w2_L1	1	0.8307	0.2929	2.84	0.0046
car3w2_L1	1	-1.1095	0.1687	-6.58	<.0001
car1w3_L1	1	0.8599	0.2035	4.23	<.0001
car2w3_L1	1	0.3938	0.2236	1.76	0.0782
car3w3_L1	1	-0.5108	0.0993	-5.14	<.0001
nww0c1_L1	1	0.4572	0.1295	3.53	0.0004
nww0c2_L1	1	0.7725	0.1798	4.30	<.0001
nww0c3_L1	1	0.9471	0.1924	4.92	<.0001
nww1c1_L1	1	0.2811	0.0857	3.28	0.0010
nww1c2_L1	1	0.4880	0.1146	4.26	<.0001
nww1c3_L1	1	0.5057	0.1188	4.26	<.0001
nww2c1_L1	1	0.4742	0.1427	3.32	0.0009
nww2c2_L1	1	0.5631	0.1558	3.61	0.0003
nww2c3_L1	1	0.7627	0.1598	4.77	<.0001
det_home_c1_L1	1	-0.0249	0.0393	-0.63	0.5261
det_home_c2_L1	1	-0.0120	0.0412	-0.29	0.7701
det_home_c3_L1	1	-0.0146	0.0585	-0.25	0.8031
zone_pop_den_c1_L1	1	-0.001887	0.000806	-2.34	0.0192
zone_pop_den_c2_L1	1	-0.004320	0.001395	-3.10	0.0020
zone_pop_den_c3_L1	1	-0.0140	0.002154	-6.49	<.0001
lowinc_c1_L1	1	-1.3870	0.3464	-4.00	<.0001
lowinc_c2_L1	1	-1.5937	0.3779	-4.22	<.0001
lowinc_c3_L1	1	-1.4196	0.3609	-3.93	<.0001
zone_det_home_c1_L1	1	0.008331	0.002470	3.37	0.0007
zone_det_home_c2_L1	1	0.0127	0.003019	4.20	<.0001
zone_det_home_c3_L1	1	0.0121	0.003885	3.12	0.0018
dist_pop_den_c1_L1	1	-0.007736	0.002169	-3.57	0.0004
dist_pop_den_c2_L1	1	-0.009626	0.002602	-3.70	0.0002
dist_pop_den_c3_L1	1	-0.0151	0.003399	-4.44	<.0001
gla_c1_L1	1	-8.876E-7	4.4608E-7	-1.99	0.0466
gla_c2_L1	1	-1.413E-6	5.3848E-7	-2.62	0.0087
gla_c3_L1	1	-2.278E-6	8.6394E-7	-2.64	0.0084
Access_c1_L1	1	0.0422	0.0149	2.83	0.0046
Access_c2_L1	1	0.0556	0.0168	3.31	0.0009
Access_c3_L1	1	0.0684	0.0179	3.82	0.0001
INC_L2G2	1	2.1152	0.5551	3.81	0.0001
INC_L3G1	1	0.5900	0.1522	3.88	0.0001