WORK TRAVEL MODE CHOICE MODELING USING DATA MINING: DECISION TREES AND NEURAL NETWORKS

Chi Xie
Research Assistant
Department of Civil and Environmental Engineering
University of Massachusetts, Amherst
139C Marston Hall
Amherst, MA 01003-5305
Phone: (413) 545-3852
E-mail: cxie@ecs.umass.edu

Jinyang Lu
Research Assistant
Department of Civil and Environmental Engineering
University of Massachusetts, Amherst
139A Marston Hall
Amherst, MA 01003-5305
Phone: (413) 545-3852
E-mail: jilu@ecs.umass.edu

Emily Parkany
Assistant Professor
Department of Civil and Environmental Engineering
Villanova University
800 Lancaster Avenue
Villanova, PA 19085-1685
Phone: (610) 519-4957
Fax: (413) 519-6754
E-mail: emily.parkany@villanova.edu

Submitted for presentation at the 82nd Transportation Research Board Annual Meeting, January 12-16, 2003, Washington, D.C. and publication in the Transportation Research Record
Work Travel Mode Choice Modeling Using Data Mining: Decision Trees and Neural Networks

Chi Xie, Jinyang Lu and Emily Parkany

ABSTRACT

Travel mode choice modeling has received the most attention among discrete choice problems in travel behavior literature. Most traditional mode choice models are based on the principle of random utility maximization derived from econometric theory. Alternatively, mode choice modeling can be regarded as a pattern recognition problem in which multiple human behavior patterns reflected from explanatory variables determine the choices between alternatives or classes. This paper investigates the capability and performance on work travel mode choice modeling of two emerging pattern-recognition data mining methods: decision trees (DT) and neural networks (NN). Models based on these two techniques are specified, estimated, and comparatively evaluated with a traditional multinomial logit (MNL) model. For comparison, the paper presents a unique three-layer formulation of the MNL model, and identifies the similarities and differences of the models’ mechanisms and structures, comparing them in the model specifications and estimations. Two performance measures, individual prediction rate and aggregate prediction rate, respectively representing the prediction accuracies on individual and mode aggregate levels, are applied for the model evaluation. Diary datasets from the San Francisco Bay Area Travel Survey (BATS) 2000 are used in this study. The prediction results show that the two data mining models offer comparable but slightly better performance than the MNL model in terms of the modeling results, while the DT model demonstrates highest estimation efficiency and most explicit interpretability and the NN model gives a superior prediction performance in most cases.

Keywords: travel behavior, mode choice modeling, data mining, decision trees, neural networks, multinomial logit model
INTRODUCTION

Discrete choice modeling of travel behavior can help estimate the proportion of competitive, mutually exclusive alternatives and hence the resulting travel demand and temporal distribution across different travel and activity patterns. The disaggregate choices, e.g., destination, mode, departure time, or route choice, constitute important components in the framework of activity-based travel demand and duration modeling. The choice preference and mechanisms form the foundation of travel and activity patterns in the context of activity generation and scheduling.

Travel mode choice has received the most attention among discrete choice problems in travel behavior literature. Mode choice modeling and prediction relate closely to transportation system policies and travel demand control and congestion mitigation strategies. Most mode choice models are based on the principle of random utility maximization derived from econometric theory. Since the multinomial logit (MNL) model (1) was developed in the 1970s, the parametric model family including different logit models with different structures and components has become the most widely used tool for mode choice analysis. However, many of these models suffer from the property of independence of irrelevant alternatives (IIA), which implies that the effects of attributes of an alternative are compensatory and result in biased estimates and incorrect predictions in cases that violate the IIA property (2), although significant improvements on eliminating the IIA property have been made.

Traditional logit models may not provide a satisfactory degree of behavioral pattern prediction. A main reason for this is that their pre-determined structures may often misestimate or ignore partial relationships between explanatory variables and alternative choices for specific subgroups of a population. The linear property and synergy effects of the utility functions may not adequately model the comprehensive and complex correlations among explanatory variables and between them and dependent variables.

Recently, computational process models have triggered the interests of travel behavior researchers for discrete choice modeling (see (3)). These models estimate choice behavior without making parametric functional form assumptions or with semiparametric assumptions. Bayesian techniques are applied to discrete choice analysis because they provide a principal approach for incorporating non-sample prior information and they avoid asymptotic approximations. Different from classic logit models, Bayesian discrete choice models treat parameters as random variables; thus Bayesian inference conditions depend on the observed data. Consequently, Bayesian methods have been incorporated into conditional and nested logit models (4, 5) and a multinomial probit model (3).

From another perspective, discrete choice modeling acts as a pattern recognition problem in which multiple complex patterns formed by the combination and interaction of explanatory variables determine the choice decisions among alternatives or classes. Many pattern recognition algorithms in the data mining field have been developed to discover the complicated relationships between input variables and output targets, which are difficult to identify by mathematical or statistical methods. Data mining models have more flexible structures to represent the relationship between the attributes of alternatives and choices than traditional logit-based models. As supervised learning systems, they can learn and identify pattern characteristics extracted from sample data and form adaptive structures through the computational process. Thus, they potentially offer insights into the relationships that random utility models cannot recognize. Several recent studies of applying data mining techniques, e.g., decision trees (6-8)
and neural networks (9-11), to discrete choice behavior analysis demonstrate the considerable
effects on prediction performance of discrete choice behavior modeling.

This paper investigates the capability and performance on work travel mode choice modeling
of two widely used data mining methods: decision trees (DT) and neural networks (NN). Models
based on these two methods will be specified, estimated and comparatively evaluated with a
MNL model. No published work has compared the three models. For comparison, this paper
presents a unique three-layer formulation of the MNL model, using diary datasets from the San
Francisco Bay Area Travel Survey (BATS) 2000 for the estimation and evaluation of the models.
The remainder of this paper is organized as follows. The next section will briefly describe the
mechanisms and applications of the two data mining techniques. In the third section, the
database extracted from BATS 2000 as well as the candidate explanatory variables as inputs of
the data mining models are presented. The fourth section will focus on the implementation of
the two models, including the introduction of the balance weight, the model specifications and
estimations (or calibrations), and behavioral interpretations of the data mining results. Then the
developed models are tested using the validation dataset and the performances of the two models
are presented and comparatively evaluated with a MNL model. Concluding remarks and
recommended further research are highlighted in the last section.

DATA MINING

Data mining, the exploration and analysis of large quantities of data in order to discover and
establish meaningful patterns and rules (J2), includes many methods or algorithms to reveal and
represent the conditions and mechanisms underlying various interrelated decisions leading to
observed and unobserved complex human behavior patterns. Data mining brings together ideas
and techniques from a variety of fields with very different vocabularies and applications. Two
very common data mining techniques for classification, i.e., tree induction and neural induction,
or the so-called decision trees and neural networks, are introduced below before their
applications, along with their potential benefits for discrete choice modeling.

Both techniques are based on supervised learning, the process of automatically creating a
classification model from a set of observations or cases. The induced models consist of visible
or hidden patterns, essentially generalizations over cases, which are useful for distinguishing the
classes. Once a model is induced, it can help predict the class of other unclassified cases.

Supervised induction techniques offer several advantages over traditional statistics-based
models (e.g., the family of logit models) in discrete choice modeling: 1) no specific model
structure need be specified in advance and no IIA property is assumed, thus reducing the
incompatibility between model structure and explanatory data; 2) they have the capability of
modeling non-linear systems (J2), which represent more complex relationships involved in
human behavior; and 3) the induced patterns can be extracted from a subgroup of observations
with homogeneity while statistical-based models check only for conditions that hold across an
entire population of observations in the training dataset.

Decision Trees

Decision trees are a type of rule-based tool. The attractiveness of decision tree-based models
rests on the fact that decision trees represent intuitive rules. Decision trees are “drawn” with the
root at the top and the leaves at the bottom. An observation enters the tree at the root node,
where a test driven by a trained algorithm determines which branch node the observation will
encounter next. This process repeats until the observation arrives at a leaf node. Different leaves may make the same choice, but each leaf makes that choice for a particular reason. The tests are chosen to best discriminate among target choices. Each path from the root to a leaf represents a decision rule.

The most commonly used decision tree algorithms include $\chi^2$ automatic interaction detection (CHAID) (13), classification and regression trees (CART) (14), and C4.5 algorithm (15). C4.5 algorithm is gaining more popularity. Compared to the first two decision tree algorithms, C4.5 can produce trees with varying numbers of branches at each node (over CART algorithm) and deal with both continuous and discrete variables (over CHAID algorithm). This study uses C4.5 for constructing the decision tree (DT) model.

**Neural Networks**

Artificial neural networks are information processing structures consisting of basic units designed to model the behavior of human neurons. Like the physical architecture of the brain, they are composed of a number of parallel, distributed and interconnected neurons or processing elements (PEs) to produce linear or nonlinear mapping between input and output variables (16, 17). Neural networks have been widely applied to transportation problems including driver behavior modeling, automated driving, vehicle detection, pavement maintenance, traffic pattern recognition, and vehicle scheduling and routing (18-20). For discrete choice modeling, neural networks use pattern association and error correction as the underlying mechanisms to represent a problem or relationship, as compared to the random utility maximization rule used in the logit models.

Neural networks operate as a simple process. A unit or neuron combines its inputs into a single output value, a process called the unit’s activation function. An activation function has two parts: combination function, which merges all the inputs into a single value; and transfer function, which transfers the value of the combination function to the output of the unit. A network can contain units with different transfer functions serving different operations. A common neural network structure used for classification is the topology, or architecture, called multi-layer feedforward neural networks (MLF). The primary advantage of this type of neural network is its ability to solve non-linear multi-dimensional pattern recognition and classification problems (17). We apply an MLF-type neural network (NN) model to the work-trip mode choice problem in this study.

**DATA**

**The Database**

The data used for implementing the data mining models come from the San Francisco Bay Area Travel Survey (BATS) 2000, conducted by MORPACE International, Inc. commissioned by the Bay Area Metropolitan Transportation Commission and Bay Area Rapid Transit (BART). From nine counties in the Bay Area, a total of 15,064 residences containing 34,680 respondents were randomly sampled. Each respondent provided a two-day travel diary in the database, and detailed individual and household socio-demographic data. The whole database consists of four interrelated datasets respectively reflecting household, person, trip, and vehicle characteristics. The first three datasets were merged to form the database used in this study.
This study emphasizes the work trip or trip-to-work mode choice, i.e., the primary trip in a home-to-work activity chain prior to a work activity, in which we assume the longest travel trip in the chain is the work trip. Other activities, such as the mode access trip (e.g., walking to a bus stop or a train station), or mode change activity, (e.g., switching between transit modes), does not belong to the defined work trip and hence is excluded in this study.

Substantial problems exist in the diary records due to incomplete or inconsistent data. One inconsistency comes from the conditional feasibility of specific modes. For example, an unfeasible response includes selecting the single-occupant vehicle (SOV) mode but not having a driver’s license or ability to drive. All observations containing missing, unknown or inconsistent data were removed from the database. After the above extraction and reduction from all types of activities recorded in the two-day travel diaries, the data samples include 4,746 work trips.

The alternatives for work travel mode choice used in this study include SOV (car, van, motorcycle and moped), carpool (carpooling and vanpooling), transit (bus, train and ferry), bicycle, and walk. These five modes were identified as the five main commute modes in the Bay Area (see RIDES (21)). The travel demand modeling for each mode has a considerable or potentially considerable influence on transportation infrastructure investment decision and travel demand control and management policymaking in this area. Although transit and bicycle capture relatively less market share (approximately 1% each) of the work trips, neither should be ignored. Furthermore, the total database was split randomly into two datasets with the same number of observations: one for the model estimation and another for the subsequent validation test. The actual mode split proportions in the total database as well as the training and test datasets are shown in Table 1.

**Explanatory Variables**

Two sets of explanatory variables, as seen in Table 2, were identified for work travel mode choice modeling: 1) individual/household socio-demographic attributes, which include hhsize, hhbicyc, hhmcycle, tenure, dwelltyp, gender, relation, age, licdrive, empstatus, emptype, hhincome and hhveh; and 2) trip level-of-service attributes, which include samptype, traveltime, peak, and opcost.

For the variables of dwelltyp, relation, empstatus and emptype, a certain value of a variable may be associated with few observations, which are thus ignored. The variable hhincome was coded using the uniform numbers from 1 to 15, indicating low to high income level. The variable startime was initially recorded as minutes elapsed from the starting moment of 3:00 am. Considering trips in the peak and off-peak hours periods have different impacts on mode choice behavior, for convenience and simplicity, startime was transformed to be a dummy variable called peak that equals 1 if the observation is a peak hour trip, and 0 if an off-peak hour trip. The peak hours included: 7:00-9:00 am and 4:00-6:00 pm, according to the Metropolitan Transportation Commission (22), while trips starting in other periods are considered off-peak. The variable traveltime was readily calculated as endtime minus startime. Finally, the variable opcost expresses the sum of all “visible” out-of-pocket costs, including transit fare and vehicle parking fee.
DATA MINING IMPLEMENTATION

This section focuses on the specification and estimation of the decision tree (DT) model and neural network (NN) model. The performance measures and balance weight are explained prior to the model development.

Performance Measures

Mode choice modeling predicts travelers’ mode choice decisions and hence induced travel demand for each mode or demand distribution across modes. Two types of prediction rates or match rates are defined and used to evaluate and compare the mode choice modeling performance of the data mining models. They respectively reflect the modeling performance on individual and aggregate levels.

Individual match rate \((r_i)\), or hit ratio, is the ratio of the number of correctly predicted individual observations for one mode \((N_{pi})\) over the total number of the actual observations choosing this mode \((N_a)\), expressed as,

\[
r_i = \frac{N_{pi}}{N_a}
\]  

Aggregate match rate \((r_a)\) reflects the prediction accuracy on the mode aggregate level, defined as the ratio of the number of predicted observations (including correctly and incorrectly predicted observations) for one mode \((N_{pa})\) over the number of the actual observations choosing the mode \((N_a)\). Its similar functional form is,

\[
r_a = \frac{N_{pa}}{N_a}
\]

In mode choice modeling, we generally are more concerned about the prediction of aggregate choice distribution for each mode, than that of individual choice. The former reflects the modeling performance on the macroscopic level while the latter reflects on the microscopic level. The improvement on the aggregate match rate in travel demand applications may be more meaningful than that of the individual match rate \((6)\), although the latter reflects the actual prediction accuracy. In this study, the individual and aggregate match rates are both used as the performance measures. The individual match rate is always less than 1 while the aggregate match rate may be greater than 1 or less than 1, with the individual match rate always no more than the aggregate match rate. Although the expected best values of both are 1, their implications for the prediction accuracy may not be the same when they are equal to 1.

Balance Weight

In all discrete choice modeling there exists a non-negligible problem in training or calibration caused by the data induction property in a single model structure: the model tends to readily recognize classes with more observations in the training dataset but neglect other classes with less. For example, in this study the percentage of the observations of the SOV mode approaches
80% while that of the transit mode is approximately 1% (see Table 1). The huge discrepancy in the number of observations of the different classes makes the model structure and parameters overwhelmingly adaptive to distinguish the patterns induced by the SOV observations, and therefore the model may fail in the prediction for the transit modes, because the only estimation objective of the data mining models is to maximize the individual match rate (approaching 1). This is reflected by the expected results that the model may have the best prediction performance on the SOV mode but performs worse on the other classes (i.e., erroneously classifies the other modes as the SOV mode). In a recent study using the C4 and CHAID algorithms to predict mode choice the results indicated that the two data mining models underfitted non-SOV modes (see (6)).

In order to balance classification performance among modes (and also balance the optimization of the individual and aggregate match rates), a variable termed balance weight is introduced in the estimation of the models. Applying balance weight improves the aggregate match rate of the models. It has the functional form,

\[
\begin{align*}
W_m &= \left( \frac{N_{\text{max}}}{N_m} \right)^\alpha & \forall m = 1, \ldots, 5 \\
N_{\text{max}} &= \{N_m\} \\
&0 \leq \alpha \leq 1
\end{align*}
\]

where \( w_m \) is the balance weight, applied on mode \( m \) by repeating its observations \( w_m \) times in the training, and \( N_m \) and \( N_{\text{max}} \) represent the number of observations of the mode \( m \) and of the mode with the maximum observations in the training dataset, respectively. The coefficient \( \alpha \) determines the scale of the weight. When \( \alpha = 0 \), then \( w_m = 1 \), and thus the proportion of observations of mode alternatives stays original. On the other end, when \( \alpha = 1 \), \( w_m = N_{\text{max}}/N_m \), and \( w_m \) is the frequency weight. By applying frequency weights, each mode has exactly the same proportion of observations in the training data. By imposing a frequency weight much greater than 1 on infrequently chosen modes, their attributes would be recognized by the model multiple times (i.e., \( w_m \) times) in the training. Consequently, the model may overfit these modes. Therefore, the critical \( \alpha \) value should fall between 0 and 1, calibrated to best match the actual mode split or achieve an optimal “balance” of proportions among the mode alternatives.

The balance weight can be determined when the estimation objective of the aggregate match rate is empirically realized as,

\[
\min \left[ \sum_{m=1}^{5} \frac{N_{am} - N_{pm}}{N_{am}} \right] & \forall m = 1, \ldots, 5
\]

where \( N_{am} \) and \( N_{pm} \) denote the number of actual and predicted observations of mode \( m \), respectively. Such a balance weight is applied to the estimation of the DT, NN, and MNL models compared here.
The determination of balance weight is based on the assumption that each observation under different modes has the same “importance” or the involved “cost” of an observation under different modes is equal. However, this may be not true in the real world. The resulting average social and operational cost, apparently, differs for any individual choosing different travel modes for the same trip (i.e., the trip with the same route, destination and departure time). For example, the cost for one to choose a SOV mode should be much higher than that for a transit mode because the former causes much more infrastructure investment and maintenance activities, vehicle cost, consumed fuel, emitted pollution, congestion, participation degree of the traveler and others. Consequently, the “cost” of one traveler switching from one mode to another also differs. When we consider using balance weight to adjust the proportions of the observations in the training data, it is reasonable and necessary to incorporate a “cost” coefficient for each mode prior to the estimation of the balance weight so as to reasonably model the variations from the actual mode splits. It is a great challenge to estimate the average individual social and operational cost on a reasonable and acceptable level for each mode and incorporate it into the mode choice modeling.

**Decision Tree Model**

The DT model applied for mode choice modeling in this study is based on the C4.5 algorithm (15). As a supervised learning algorithm, C4.5 uses recursive partitioning to form a tree structure with *if*-then rules (each of which is applied with an explanatory variable) as splitting criteria. Each branch on different levels of the tree represents a subgroup of observations with homogeneity of different degrees. Homogeneity increases from top to bottom where the bottom leaves contain the cases with the same mode choice while the top branches offer the roughest split. Each branch from the top node to a bottom leaf node can be described as an *if*-then rule sequence or ruleset.

The C4.5 algorithm generates a DT model in two phases: construction and pruning. The construction of a DT model follows these principles. From top to bottom of a decision tree, a training data group is divided at each stage of subdivision (i.e., node) according to an explanatory variable selected based on the splitting criterion. The division continues until all observations in a subgroup have the same mode choice at the bottom. The splitting criterion used in C4.5 is the so-called “information gain ratio” based on information theory. The detailed procedure to generate the gain ratio can be referred to in (15) and (6-8).

The generated C4.5 decision tree often becomes too complex, overfitting the training data when another hypothesis that performs less well on the training data actually performs better on the test data (6). A pruning strategy helps avoid overfitting based on the expected error rate. C4.5 prunes the decision tree using a pessimistic pruning method: the error rate at each leaf is examined based on the assumption that the true error rate will be substantially worse. For a given confidence level, C4.5 gives the confidence interval, which is the range of expected error rates. C4.5 assumes that the observed error rate on the training data marks the low end of this range and substitutes the high end to get a leaf’s predicted error rate on the test data. Out towards the few observation leaves, substituting this pessimistic error rate for the observed one often causes the error rate of a whole subtree to be higher than that of one of the nodes above it, in which case it gets pruned.

In the pruning process, a question arises: how to determine a confidence level (or confidence interval) for the pessimistic pruning, or, what is an acceptable tree size of the trained DT model
with sufficient prediction accuracy after pruning? We examine the relationship of prediction error rate and decision tree size through empirical analysis. We obtain little improvement on prediction accuracy when the tree size increases to a certain value (50 leaves in this study), making this threshold an appropriate place to prune. It represents a critical point for the balance between simple and readable tree structure and high prediction accuracy.

According to the empirically determined “optimal” tree size, we estimate the DT model for mode choice. The “best” DT model has 50 leaves and the total classification error rate (i.e., total individual erroneous match rate) is 21.0%. The detailed tree structure appears in Figure 1. Each leaf with its parameters represents one if-then decision rule. On each leaf the number of observations \((n)\) mapped by the model and the number of observations erroneously classified \((m)\) are presented as \(n/m\). The primary leaves with relatively high individual match rate \((n-m)/n\) for each mode alternative are underlined and followed by their match rate. Clearly, a number of decision rules are generated for the choice of the SOV mode and they involve a variety of attributes; moreover, these rules provide quite high prediction rates ranging from 80% to 100%. Two primary if-then rules for the carpool mode are identified, which imply these conditions: 1) a traveler with a bicycle and no license or ability to drive, with 3 vehicles or less in the household, would take a zero-out-of-pocket-cost trip in peak hours with travel time of more than 7 minutes (as an example, this branch is highlighted with bold-italic text in Figure 1); and 2) a traveler without a license but with one or more bicycles, older than 32 years, with 3 vehicles or fewer in the household, also with a zero-out-of-pocket-cost trip with travel time of more than 7 minutes. There is only one decision rule for taking the transit mode: one would take a bus or train if the charged out-of-pocket money is no more than $4. This rule makes sense to us because only transit users and parkers report non-zero-out-of-pocket costs. Since most parking fees exceed $4, this is a good threshold for identifying transit users. For the bicycle mode, no successful rule with an acceptable match rate appears in the generated decision tree. The DT model cannot identify underlying patterns for the bicycle mode, possibly due to the very limited observations of the bicycle mode in the training dataset. Finally, four primary leaves or rules with significant match rate for the walk mode also appear. The involved explanatory variables identified for the walk mode include the number of vehicles and bicycles in household, household income, age, possession of a driver’s license, travel time and out-of-pocket cost. Please refer to the tree structure in Figure 1 for details.

In the DT model tree structure, variables lying on different hierarchies indicate different effective scopes of impact on mode choice, hence different impacted numbers of mode choice decisions. For example, the variable out-of-pocket cost occupies the tree top, thus impacting all mode choices. A variable existing only on the bottom, such as gender, has relatively limited influence on travel mode decisions. Some explanatory variables, such as dwelling type, residential tenure type, household size, employment type and sampling type (i.e., urban or suburban sampling), are winnowed in the tree structure because the model recognizes their weak impacts on travelers’ mode choice decisions. These variables would likely be insignificant in traditional econometric mode choice models like logit models. As a result, the most important variables identified by the DT model for the work travel mode choice include trip out-of-pocket cost, household vehicle numbers, household income, possession of a driver’s license, traveler’s age, and travel time.
Neural Network Model

Neural network modeling does not require an algorithm to compute a specific output for a preset functional form (18). Nonlinear transfer functions and the autonomous identification of information contained in input patterns of a neural network enable mode choice estimation. The NN model with a MLF (multi-layer feedforward) topology or architecture used in this study consists of PEs (processing elements) arranged in three layers: the input layer, the hidden layer, and the output layer, as shown in Figure 2. Interconnected PEs in adjacent layers have connection weights \( w_{i,j} \) and \( v_{j,k} \).

When an input vector (or an observation) is presented to the network, each input element (or an explanatory variable) is multiplied by an appropriate weight and connected to all the PEs on the hidden layer. This procedure is realized through a transfer function \( f(x) \) and produces an output value \( h_j \) on the hidden layer, expressed mathematically as,

\[
h_j = f\left(\sum_{i=1}^{n} w_{i,j} o_i + b_j^h\right) \quad \forall j = 1, \ldots, p
\]

(5)

where \( o_i \) represents the \( i \)th explanatory variable, the bias value of the \( j \)th PE in the hidden layer, \( b_j^h \), adjusts the magnitude of the output, and \( f(x) \) is a sigmoid (or logistic) function that is familiar to transportation researchers from trip demand analysis and other travel behavior studies (23). The sigmoid function,

\[
f(x) = \frac{1}{1 + e^{-x}}
\]

(6)

is repeated between the hidden layer and output layer, which can be expressed in similar functional form as,

\[
p_m = f\left(\sum_{j=1}^{p} v_{j,m} h_j + b_m^o\right) \quad \forall m = 1, \ldots, 5
\]

(7)

where \( p_m \) and \( b_m^o \) indicate the \( k \)th output element and bias value on the output layer.

The input vector \( o \) comprises 17 explanatory variables (\( n \)) (23 if dummy variables are included) as input elements. Each input vector represents an observation. The number of PEs (\( p \)) were decided on the hidden layer, in terms of the best classification result from a series of experiments with \( p \) on this layer ranging from 6 to 45; the output layer gives numerical mode choice results where the output element that produces the maximum value is transformed to 1 and others to 0. Each output vector fits into one of the five mode choices, where the output vector \( [1 \ 0 \ 0 \ 0 \ 0]^T \) corresponds to the SOV mode, and \( [0 \ 1 \ 0 \ 0 \ 0]^T \), \( [0 \ 0 \ 1 \ 0 \ 0]^T \), \( [0 \ 0 \ 0 \ 1 \ 0]^T \) and \( [0 \ 0 \ 0 \ 0 \ 1]^T \) denote the other four modes, respectively.
The NN model may be trained through presenting a set of input-target pairs to the network and updating the network parameters in terms of the commonly used backpropagation algorithm (16, 17), which seeks to implement a gradient descent of an error function of the network’s output (see (24) for the detailed procedure). For each input vector, the output produced by the network is compared with the target vector, to adjust the weights (i.e., $w_{i,j}$ and $v_{j,m}$) and biases (i.e., $b_j^h$ and $b_m^o$) in the gradient descent direction of the output error surface until the magnitude of output error becomes acceptable. During the training, an adaptive learning rate and a momentum rate (0.7 in this study) are employed for faster training while keeping learning stable. The initial learning rate was set as 0.02.

A total of 640 training sessions (for 40 different numbers of PEs, i.e., 6-45 PEs, on the hidden layer and 4 different transfer functions on the hidden and output layers) used the same training dataset to seek the optimal configuration of the NN model structure and parameters. The training dataset was grouped into two parts: 80% for parameter adjustment and 20% for cross validation. The training session lasted for a maximum of 5,000 epochs, with each epoch equivalent to one presentation of all of the input vectors in random order. At the end of each epoch, the trained network is tested on the cross validation dataset to determine whether the updated parameters are saved or discarded. The training results indicate that, as aforementioned, 33 PEs are identified on the hidden layer and the sigmoid transfer function is applied on both the hidden and output layer. Table 3 summarizes the features of the trained NN model.

The prediction results from applying the estimated NN model on the same dataset show that the NN model can provide 78.3% individual match rate (i.e., 1,587 out of 2,373 are correct hits) for the total of five mode choices. High individual match rate occurs not only for the SOV mode, but also for the modes of transit (100.0% match rate), bicycle (97.5%) and walk (92.6%). For the carpool mode, however, a relatively low individual match rate (59.9%) is realized. The NN model provides aggregate match rates of 90.9%, 102.7%, 234.8%, 277.5%, and 151.6% for the SOV, carpool, transit, bicycle, and walk modes respectively, which shows a comparable mode choice modeling capability to the DT model on the aggregate level. Overall, it can be suggested that the NN model has a more robust performance on travel mode choice prediction compared to the DT model in the model estimation stage.

Sensitivity analysis enables neural networks to explain which inputs are more important than others. This analysis can be performed inside the network, using the errors generated from backpropagation, or externally, by poking the network with specific inputs (12). In this study, the latter method examines the importance of the explanatory variables using the predetermined subgroups of observations, one of which contains only observations with the single value of one variable (here, the continuous variables are discretized into the ordered categorical variables) and one type of travel mode. A total of 340 subgroups are produced (apparently, these subgroups mutually overlap). The examination results prove that the most significant variables across the modes include household size, household vehicle number, gender, possession of a driver’s license, travel time, time-of-day, and out-of-pocket cost.

**EVALUATION**

The same test dataset derived from our database is applied to examine the performance of the two developed data mining models and compare them with a traditionally used multinomial logit (MNL) model.
Result Analysis

Confusion (or misclassification) matrices measure the effectiveness of the discrete mode choice models. Tables 4a and 4b present confusion matrices induced by the DT and NN models for both the training and test datasets. In a confusion matrix, each row represents the actual observations of each mode while each column denotes the predicted observations. The sum on each row or column represents the actual or predicted number of observations for each mode. The row head shares the same order of mode alternatives with the column head. Thus, the diagonal cells give the match number between reality and prediction and non-diagonals provide the erroneous classification. The match rate for each mode appears in the table as the index of prediction performance.

Overall, the NN model outperforms the DT model on both the training and test datasets when comparing of the individual prediction rates (i.e., the DT model has 79.0% and 76.8% match rates in the training and test while the NN model has 83.6% and 78.2% respectively). The misclassification results reflect that the two models present the same prediction characteristics. For example, they both offer high prediction accuracy for the SOV mode in either the training or the test. Neither easily distinguishes the SOV and carpool modes in that many observations under these two modes are mutually misclassified. The phenomenon indicates that the SOV and carpool modes, which share physical, technical, and socio-economical attributes, exhibit more homogeneity within the explanatory variables than the other modes, which inevitably leads to the classification difficulty between them. The prediction results for the transit mode are different. Both models yield the high individual match rate for the transit mode under a “large” number of conditions (i.e., 3 out of 4) where most of the observations choosing the transit mode are not misclassified as the other modes. However, the NN model makes a relatively worse prediction (i.e., the individual prediction rate of 30.8%) in the test dataset.

In the confusion matrices of the DT model, the number of observations of the transit mode choice is estimated very well but the bicycle mode is underestimated heavily, although they both have fewer observations in the whole database. A large part of the misclassified observations of the bicycle mode go to the carpool mode, which may imply some unobserved similar preferences between carpooling passengers and bicycle users. From the same perspective, the transit mode is unique, which sharing few attributes or travelers’ preferences with other modes.

Compared to the DT model, the NN model shows worse transferability in that the match rates for both the total or single mode decrease considerably with varying degrees from the training to the test, especially to the transit, bicycle, and walk modes. The NN model may overfit the training data on these modes, or may not identify complete data patterns. Underestimation may be caused by insufficient observations under these modes.

Through analyzing the aggregate prediction rates, the discrepancy between actual and predicted mode choice patterns can be identified. We conclude that both data mining models made acceptable predictions of the mode choice distribution on the aggregate level. Moreover, both models show similar aggregate prediction patterns, providing a relatively accurate aggregate prediction rate for the SOV and carpool modes, but both overestimate the aggregate numbers of observations of the transit, bicycle and walk mode.
Comparison with a Multinomial Logit (MNL) Model

A MNL model developed as a benchmark to evaluate the performance of the two developed data mining models, can be expressed using a similar structure with different parameters and operational functions compared to the MLF-type NN model, as illustrated in Figure 3. The diagram and formulation below are unique to this paper. The MNL model has three layers: the input layer, the utility layer and the output layer. The input layer’s input elements are the explanatory variables (including the dummy variables). The utility layer (comparable to the hidden layer of the MLF-type NN model) contains 5 nodes corresponding to the utilities of the 5 travel modes. The transfer function on the utility layer has the following functional form as,

\[ h_m = f \left( \sum_{i=1}^{n} w_{i,m} o_i + b_m \right) \]  

(8)

where \( o_i \) represents the \( i \)th explanatory variable, \( w_{i,m} \) the coefficient of the \( i \)th explanatory variable in the utility function of the mode \( m \) (comparable to the connection weight \( w_{i,j} \)), \( b_m \) the mode specific constant (comparable to the bias value \( b_j \)), and \( f(x) \) an exponential operation (different from the sigmoid function) to the mode utility \( x \), \( f(x) = e^x \). Then, the likelihood function on the output nodes has the mathematical expression,

\[ p_m = \frac{h_m}{\sum_{m=1}^{s} h_m} \]  

(9)

as distinct from the sigmoid transfer function on the output modes in the NN model. The connection weights and biases between the utility layer and the output layer are 1 and 0, respectively. The output layer gives the probabilities of each mode choice where the node with the maximum probability is assigned as 1 (selected) and others nodes are assigned as 0 (discarded), same as the NN model.

Furthermore, two primary discrepancies exist between the NN and MNL models: 1) estimation method; and 2) interpretability. The NN model is estimated using the standard backpropagation method while the MNL model takes the maximum likelihood method for estimation. In the MNL model, the magnitude and sign of the coefficients indicate the importance and impact of the corresponding variables on mode choices and the values of Z-statistic or t-statistic test indicate their confidence level. Comparatively, the NN model shows weak interpretability, although the importance of explanatory variables can be identified through sensitivity analysis.

We used the same training dataset to estimate the MNL model, expanding each observation to accommodate the dummy variables for the incorporation of all values of the categorical variables. The SOV mode is arbitrarily used as the base alternative. The base utilities of the other modes relative to this base mode are represented by alternative specific constants. From the estimation results, the most significant variables to influence a traveler’s mode choice decision identified by the MNL model include: household size, household income, household...
vehicle number, license possession, travel time, trip time-of-day and out-of-pocket cost. These variables approximately match the important variables induced by the DT or NN models.

An overall performance comparison was conducted based on the prediction results of the three models tested on the validation dataset. Figure 4 shows the correctly predicted individual and aggregate observations of the three models by each travel mode, in which the actual numbers of observations for each mode are also labeled. The three models show comparable prediction performances. None of them can give the best prediction rate for each mode on individual level or aggregate level.

On the individual prediction level (see Figure 4a), the NN model (88.0%) shows a best overall performance over the other two models (86.0% and 86.7% for the DT and MNL model) in the prediction of the SOV mode. The MNL model performs worst in the prediction for the carpool mode, underestimating most of the carpool observations. Hence, the MNL model (72.9%) is worse than the DT and NN models (76.8% and 78.2%) on the overall individual prediction performance. The two data mining models show the very close prediction results for each mode (the difference is less than 3%).

On the aggregate prediction level (see Figure 4b), the three models all demonstrated satisfactory accuracy and powerful practicability. The MNL model shows the best prediction capability on the SOV mode (101.4%), but performs worst on the aggregate prediction for the carpool mode. The NN model outperforms on the prediction for carpool, bicycle and walk modes and shows close performances to the best prediction results for the SOV and transit modes. In most cases, the NN model proves to be the best or nearly best model on the aggregate mode choice prediction level.

CONCLUSIONS

The performance of mode choice modeling rests on two aspects: data preparation and model development and estimation. Data preparation includes data collection and sampling, dependent and explanatory variable identification, data extraction and deduction, and data discretization and transformation. This paper focuses on the second aspects: investigating the feasibility of applying the emerging data mining techniques to travel mode choice behavior modeling. Two representative data mining models, the decision tree (DT) model and neural network (NN) model, were developed and comparatively evaluated with a widely used multinomial logit (MNL) model. Among the work-trip observations, five mode alternatives are identified: SOV, carpool, transit, bicycle and walk, for a comprehensive analysis on the mode choice prediction and mode share pattern.

Both data mining models show high flexibility and adaptability of the model structure or parameters to the training data due to their data induction property. No IIA property needs to be assumed so that the compatibility between the model structure and the observations is enhanced in the model estimation and hence the prediction performance can be improved compared to the MNL model. Both models, however, have difficulty estimating modes with insufficient observations in the database, for the models are induced by the data to recognize the mode with the most observations. Balance weight is introduced in the model estimation to empirically maintain a balance of individual and aggregate prediction accuracies among the modes. Two performance measures, individual prediction rate and aggregate prediction rate, respectively representing the prediction accuracies on individual and mode aggregate levels, are used for the
model evaluation. Comparative evaluation shows that the two data mining models have comparable but slightly better prediction capability than the MNL model on work travel mode choice modeling. The three models provide comparable and sufficiently accurate prediction results of mode choice (except for the carpool mode) and mode share pattern. However, the prediction results based on the separate test dataset show, on both individual and aggregate levels, that the NN model outperforms the other two models. The most significant explanatory variables identified by all three models include two socio-demographic attributes: household vehicle number and license possession, and two level-of-service attributes: travel time and out-of-pocket cost.

The DT model performs with its structure flexibility to adapt to the training data, and has the capability to produce explainable if-then rules and identify the significance of explanatory variables for each mode. The DT model induces an if-then ruleset using a sequence of explanatory attributes; however, it cannot capture the correlations among attributes or among rulesets. Decision trees have problems with processing continuous data; data have to first be grouped into ranges manually or automatically by the software tool. The selection of the ranges may unwittingly hide useful patterns. The DT model may suffer from its estimation algorithm: during the estimation, once the model makes a decision about a variable on which to split the node, the decision cannot be revised or improved, due to the absence of a backtracking technique, for which the NN model makes provision (25). Another data mining model, the NN model with a MLF topology, has a similar three-layer structure compared to the MNL model with its new structure interpretation. However, the discrepancies on the number of middle layer nodes, layer connection weights and biases, and transfer functions yield different performance and interpretability. The NN model, though saddled with the interpretation problem, can identify the importance of explanatory variables through external sensitivity analysis. However, sensitivity analysis is time-consuming and other types of models can substitute simple and feasible methods for the same purpose. Considering its superior prediction performance, therefore, we believe that it has practice value rather than explainability. In some applications, the NN model performs well with its expected high prediction rate when the travel demand amount of each mode is the only concern. In other applications, the DT model is preferred when the ability of a model to interpret the reason for a choice and to identify important variables for policymaking is crucial.

The primary features and capabilities of the DT, NN and MNL models in this study are summarized in comparative form in Table 5.

ACKNOWLEDGEMENTS

The authors would like to acknowledge Dr. Kenneth Vaughn and Mr. Chuck Purvis at the San Francisco-area Metropolitan Transportation Commission (MTC) for providing the BATS 2000 datasets and constructive suggestions for building the work-trip database. Dr. Ross Quinlan at University of New South Wales, Australia, helped clarify the See5/C5.0 program. The authors also benefited from the discussion on selection and construction of the neural network model in e-mail exchanges with Mr. Fang Yuan at University of Tennessee. However, the contents of this paper reflect only the views of the authors, who are solely responsible for the facts and the accuracy of the data presented herein.
REFERENCES


LIST OF TABLES

TABLE 1 Summary of the Mode Splits in the Datasets
TABLE 2 Explanatory Variables
TABLE 3 The Features of the Developed NN Model
TABLE 4 Confusion Matrices Generated by the DT and NN Models
TABLE 5 Summary of the Features and Capabilities of the DT, NN and MNL Models
<table>
<thead>
<tr>
<th>Mode</th>
<th>Total Database</th>
<th></th>
<th>Training Dataset</th>
<th></th>
<th>Test Dataset</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number</td>
<td>Per (%)</td>
<td>Number</td>
<td>Per (%)</td>
<td>Number</td>
<td>Per (%)</td>
</tr>
<tr>
<td>SOV</td>
<td>3,789</td>
<td>79.84</td>
<td>1,894</td>
<td>79.81</td>
<td>1,895</td>
<td>79.86</td>
</tr>
<tr>
<td>Carpool</td>
<td>622</td>
<td>13.11</td>
<td>294</td>
<td>12.39</td>
<td>328</td>
<td>13.82</td>
</tr>
<tr>
<td>Transit</td>
<td>36</td>
<td>0.76</td>
<td>23</td>
<td>0.97</td>
<td>13</td>
<td>0.55</td>
</tr>
<tr>
<td>Bicycle</td>
<td>68</td>
<td>1.43</td>
<td>40</td>
<td>1.69</td>
<td>28</td>
<td>1.18</td>
</tr>
<tr>
<td>Walk</td>
<td>231</td>
<td>4.87</td>
<td>122</td>
<td>5.14</td>
<td>109</td>
<td>4.59</td>
</tr>
<tr>
<td>Sum</td>
<td>4,746</td>
<td>100.00</td>
<td>2,373</td>
<td>100.00</td>
<td>2,373</td>
<td>100.00</td>
</tr>
</tbody>
</table>
### TABLE 2  Explanatory Variables

<table>
<thead>
<tr>
<th>Variable/Label$^1$</th>
<th>Definition</th>
<th>Values</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Socio-demographic attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1 DWELLTYP</td>
<td>Dwelling type</td>
<td>A single-family detached house; Duplex or duet; Apartment; Condominium or townhouse</td>
</tr>
<tr>
<td>2 TENURE</td>
<td>Residential tenure type</td>
<td>Rent; Own</td>
</tr>
<tr>
<td>3 HHSIZE</td>
<td>Household size</td>
<td>Continuous</td>
</tr>
<tr>
<td>4 HHINCOME</td>
<td>Household income</td>
<td>1: Below $10,000; 2: $10,000-15,000; 3: $15,000-20,000; 4: $20,000-25,000; 5: $25,000-30,000; 6: $30,000-35,000; 7: $35,000-40,000; 8: $40,000-45,000; 9: $45,000-50,000; 10: $50,000-60,000; 11: $60,000-75,000; 12: $75,000-100,000; 13: $100,000-125,000; 14: $125,000-150,000; 15: Above $150,000</td>
</tr>
<tr>
<td>5 HHVEH</td>
<td>No. of vehicles in household</td>
<td>Continuous</td>
</tr>
<tr>
<td>6 HHMCYCLE</td>
<td>No. of motorcycles in household</td>
<td>Continuous</td>
</tr>
<tr>
<td>7 HHBICYC</td>
<td>No. of bicycles in household</td>
<td>Continuous</td>
</tr>
<tr>
<td>8 AGE</td>
<td>Age of traveler in years</td>
<td>Continuous</td>
</tr>
<tr>
<td>9 GENDER</td>
<td>Gender of traveler</td>
<td>Male; Female</td>
</tr>
<tr>
<td>10 RELATION</td>
<td>Relation in household</td>
<td>Husband/wife/partner; Unrelated adult/partner; Son/daughter; Father/mother/father-in-law/mother-in-law</td>
</tr>
<tr>
<td>11 EMPSTATUS</td>
<td>Work status</td>
<td>Full-time, Part-time</td>
</tr>
<tr>
<td>12 EMPTYPE</td>
<td>Type of employment</td>
<td>Private, for-profit company; Private, not-for-profit company; Governmental agency; Self-employed</td>
</tr>
<tr>
<td>13 LICDRIVE</td>
<td>Licensed or capable of driving</td>
<td>1: Yes; 0: No</td>
</tr>
<tr>
<td><strong>Level-of-service attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14 SAMPTYPE</td>
<td>Sampling type</td>
<td>Urban; Suburban</td>
</tr>
<tr>
<td>15 TRAVELTIME</td>
<td>Travel time in minutes</td>
<td>Continuous</td>
</tr>
<tr>
<td>16 PEAK</td>
<td>Activity travel time</td>
<td>1: Peak hour; 0: Off-peak hour$^2$</td>
</tr>
<tr>
<td>17 OPCOST$^3$</td>
<td>Out-of-pocket travel cost in $</td>
<td>Continuous</td>
</tr>
</tbody>
</table>

$^1$ Explanatory variables with multiple options such as DWELLTYP or RELATION are converted to dummy variables for the MNL model used in the model comparison.

$^2$ Peak hours are between 7:00 and 9:00 am and between 4:00 and 6:00 pm. Please refer to Metropolitan Transportation Commission (21).

$^3$ OPCOST (i.e., out-of-pocket cost) includes TRFARE (i.e., fare of public transportation for a trip) and PARKCOST (i.e., parking cost for a trip).
### TABLE 3 The Features of the Developed NN Model

<table>
<thead>
<tr>
<th>NN Model Features</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Input features</td>
<td>17 explanatory variables (23 if dummy variables included)</td>
</tr>
<tr>
<td>Output features</td>
<td>5 travel mode alternatives</td>
</tr>
<tr>
<td>Learning algorithm</td>
<td>Backpropagation training algorithm with gradient descent momentum and adaptive learning rate</td>
</tr>
<tr>
<td>Transfer functions</td>
<td>Sigmoid function on both the hidden and output layer</td>
</tr>
<tr>
<td>Maximum training cycles</td>
<td>5,000 epochs</td>
</tr>
<tr>
<td>Classification error criterion</td>
<td>Mean squared error (MSE)</td>
</tr>
<tr>
<td>Stop criteria</td>
<td>MSE increase on cross validation test of over 100 epochs</td>
</tr>
</tbody>
</table>
TABLE 4 Confusion Matrices Generated by the DT, NN and MNL Models

<table>
<thead>
<tr>
<th>Actual Mode Choice</th>
<th>Predicted Mode Choice</th>
<th>Individual Match Rate (%)</th>
<th>Aggregate Match Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Training Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOV (1,765)</td>
<td>Carpool (287)</td>
<td>Transit (54)</td>
</tr>
<tr>
<td></td>
<td>1,621</td>
<td>126</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>SoV (1,894)</td>
<td>Carpool (294)</td>
<td>Transit (23)</td>
</tr>
<tr>
<td></td>
<td>92</td>
<td>151</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>Test Dataset</td>
<td>Predicted Mode Choice</td>
<td>Individual Match Rate (%)</td>
</tr>
<tr>
<td></td>
<td>SOV (1,820)</td>
<td>Carpool (210)</td>
<td>Transit (43)</td>
</tr>
<tr>
<td></td>
<td>1,629</td>
<td>69</td>
<td>26</td>
</tr>
<tr>
<td></td>
<td>SoV (1,895)</td>
<td>Carpool (328)</td>
<td>Transit (13)</td>
</tr>
<tr>
<td></td>
<td>138</td>
<td>128</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>(a) Confusion matrix classified by the DT model</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Test Dataset</td>
<td>Predicted Mode Choice</td>
<td>Individual Match Rate (%)</td>
</tr>
<tr>
<td></td>
<td>SOV (1,831)</td>
<td>Carpool (262)</td>
<td>Transit (49)</td>
</tr>
<tr>
<td></td>
<td>1,667</td>
<td>107</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>SoV (1,895)</td>
<td>Carpool (328)</td>
<td>Transit (13)</td>
</tr>
<tr>
<td></td>
<td>118</td>
<td>127</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(b) Confusion matrix classified by the NN model</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Actual Mode Choice</td>
<td>Predicted Mode Choice</td>
<td>Individual Match Rate (%)</td>
</tr>
<tr>
<td></td>
<td>Training Dataset</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>SOV (1,721)</td>
<td>Carpool (302)</td>
<td>Transit (54)</td>
</tr>
<tr>
<td></td>
<td>1,633</td>
<td>124</td>
<td>29</td>
</tr>
<tr>
<td></td>
<td>SoV (1,894)</td>
<td>Carpool (294)</td>
<td>Transit (23)</td>
</tr>
<tr>
<td></td>
<td>81</td>
<td>176</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Test Dataset</td>
<td>Predicted Mode Choice</td>
<td>Individual Match Rate (%)</td>
</tr>
<tr>
<td></td>
<td>SOV (1,831)</td>
<td>Carpool (262)</td>
<td>Transit (49)</td>
</tr>
<tr>
<td></td>
<td>1,667</td>
<td>107</td>
<td>30</td>
</tr>
<tr>
<td></td>
<td>SoV (1,895)</td>
<td>Carpool (328)</td>
<td>Transit (13)</td>
</tr>
<tr>
<td></td>
<td>118</td>
<td>127</td>
<td>10</td>
</tr>
<tr>
<td></td>
<td>(b) Confusion matrix classified by the NN model</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TRB 2003 Annual Meeting CD-ROM   Paper revised from original submittal.
<table>
<thead>
<tr>
<th>Actual Mode Choice</th>
<th>Training Dataset</th>
<th>Predicted Mode Choice</th>
<th>Individual Match Rate (%)</th>
<th>Aggregate Match Rate (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>SOV (1,901)</td>
<td>Carpool (99)</td>
<td>Transit (50)</td>
<td>Bicycle (104) Walk (219)</td>
</tr>
<tr>
<td>SOV (1,894)</td>
<td>1,632</td>
<td>38</td>
<td>30</td>
<td>63</td>
</tr>
<tr>
<td>Carpool (294)</td>
<td>186</td>
<td>50</td>
<td>9</td>
<td>15</td>
</tr>
<tr>
<td>Transit (23)</td>
<td>15</td>
<td>1</td>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>Bicycle (40)</td>
<td>16</td>
<td>3</td>
<td>2</td>
<td>15</td>
</tr>
<tr>
<td>Walk (122)</td>
<td>52</td>
<td>7</td>
<td>3</td>
<td>11</td>
</tr>
<tr>
<td></td>
<td>Test Dataset</td>
<td>Predicted Mode Choice</td>
<td>Individual Match Rate (%)</td>
<td>Aggregate Match Rate (%)</td>
</tr>
<tr>
<td></td>
<td>SOV (1,922)</td>
<td>Carpool (87)</td>
<td>Transit (45)</td>
<td>Bicycle (116) Walk (203)</td>
</tr>
<tr>
<td>SOV (1,895)</td>
<td>1,643</td>
<td>37</td>
<td>26</td>
<td>80</td>
</tr>
<tr>
<td>Carpool (328)</td>
<td>213</td>
<td>35</td>
<td>14</td>
<td>21</td>
</tr>
<tr>
<td>Transit (13)</td>
<td>5</td>
<td>3</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Bicycle (28)</td>
<td>12</td>
<td>3</td>
<td>1</td>
<td>6</td>
</tr>
<tr>
<td>Walk (109)</td>
<td>49</td>
<td>9</td>
<td>9</td>
<td>42</td>
</tr>
</tbody>
</table>

(c) Confusion matrix classified by the MNL model
### TABLE 5  Summary of the Features and Capabilities of the DT, NN and MNL Models

<table>
<thead>
<tr>
<th>Model</th>
<th>DT Model</th>
<th>NN Model</th>
<th>MNL Model</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Application</strong></td>
<td>Classification and Prediction</td>
<td>General purposes such as classification, estimation, prediction, segmentation, association and others</td>
<td>Discrete choice prediction</td>
</tr>
<tr>
<td><strong>Scope</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Model</strong></td>
<td>Tree structure</td>
<td>MLF three-layer structure with variable number(^1) of nodes on the middle layer</td>
<td>Three-Layer structure with fixed number(^2) of nodes on the middle layer</td>
</tr>
<tr>
<td><strong>Topology</strong></td>
<td>Recursive partitioning</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Estimation</strong></td>
<td>Fast (Approximately 1-2 sec)</td>
<td>Extremely slow (approximately 1-1.5 hr)</td>
<td>Moderate (Approximately 8-10 sec)</td>
</tr>
<tr>
<td><strong>Efficiency(^3)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Prediction</strong></td>
<td>76.8%</td>
<td>78.2%</td>
<td>72.9%</td>
</tr>
<tr>
<td><strong>Performance(^4)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Interpretability</strong></td>
<td>Explicit decision trees/\texttt{if-then rules}</td>
<td>Implicit “black box”</td>
<td>Explicit utility functions</td>
</tr>
<tr>
<td><strong>Software Tool</strong></td>
<td>See5/C5.0</td>
<td>MATLAB R11</td>
<td>STATA 7</td>
</tr>
</tbody>
</table>

---

1. The number of nodes on the middle layer in the NN model needs to be specified prior to the model estimation.
2. The number of nodes on the middle layer in the MNL model equals to the number of the modes.
3. The estimation time is based on the performance of a typical Pentium-III PC with 256 MB memory.
4. The index of prediction performance is the individual match rate for the travel mode choice.
LIST OF FIGURES

FIGURE 1 Tree Structure of the Induced DT Model
FIGURE 2 Topology of the NN Model
FIGURE 3 Topology of the MNL Model
FIGURE 4 Prediction Performance Comparison among the DT, NN and MNL Models
FIGURE 1 Tree Structure of the Induced DT Model

1 The primary leaves with relatively high individual match rate for each mode alternative are underlined.

2 One of the primary branches for the carpooling mode is highlighted in bold-italic type.
FIGURE 2 Topology of the NN Model
FIGURE 3 Topology of the MNL Model
FIGURE 4 Prediction Performance Comparison among the DT, NN and MNL Models

(a) Individual prediction performance

(b) Aggregate prediction performance