Investigating the Transferability of Individual Trip Rates: A Decision Tree Approach

Mehran Fasihozaman Langerudi
Ph.D. Student
Department of Civil and Materials Engineering
University of Illinois at Chicago
842 W. Taylor St.
Chicago, IL 60607
Phone: 312-996-0962
Fax: 312-996-2426
Email: mfasih2@uic.edu

Taha Hossein Rashidi, Ph.D.
Postdoctoral Fellow
Department of Civil Engineering
University of Toronto
Galbraith Building, Room 319C
Phone: 647-638-5450
Email: taha.hosseinrashidi@utoronto.ca

Abolfazl (Kouros) Mohammadian, Ph.D.
Associate Professor
Department of Civil and Materials Engineering
University of Illinois at Chicago
842 W. Taylor St.
Chicago, IL 60607
Phone: 312-996-9840
Fax: 312-996-2426
Email: kouros@uic.edu

Submitted to the 92th Annual Transportation Research Board Meeting, Jan 2013-
Washington D.C.

Submission date: 8/1/2012.

Word Count: 5643+ (11 x 250) = 8393

Abstract

Transferring trip rates to areas without local survey data is a common practice which is typically performed in an ad-hoc fashion using household-based cross-classification tables. This paper applies a rule-based method called decision tree to develop individual-level trip generation models for eight different trip purposes as defined in the National Household Travel Survey data (NHTS 2009) in addition to their daily vehicle miles traveled (VMT). For each trip purpose, the models are then obtained by finding the best-fitted statistical distribution to each one of the final decision tree clusters while considering the correlation between different trip purposes. The rule-based models utilize several socio-demographic and land-use explanatory variables and are sensitive to changes in demographics. The performance of the models are then tested and validated in a transferability application to Phoenix Metropolitan Region. These models can be employed in a disaggregate microsimulation framework to generate trips with different purposes at individual or household level. They can also be used as an alternative solution for trip generation step of a conventional four step travel demand model.

Keywords:
Data Transferability, Trip Rate, Decision Tree, Travel Demand, Vehicle Miles Traveled (VMT)

1. Introduction

Application of disaggregate data transferability methods for modeling household and individual travel attributes is increasing for data simulation purposes especially in small and medium sized cities. Detailed travel attributes like number of trips, distance traveled, and modes used for each individual are critical requirements of any disaggregate travel demand analysis (1, 2). Such disaggregate travel attributes are usually obtained from costly travel surveys; however, data transferability approaches are seen as reliable alternative solutions especially for smaller communities where data collection is more costly and challenging. Data Transferability is broadly referred to any approach which utilizes data or models from one context to generate data or models to be use in another context (3). This can be used either in spatial context like generating a model or data for a region on the basis of data that is obtained from another region or in a temporal context like forecasting data for a region based on an existing data from the same region. However, since transferability applications over the years have had mixed success results, transportation community has been skeptical of transferring data. Therefore, any alternative approach should be properly validated to ensure its appropriateness for the specific application.

This paper attempts to present a robust transferability method which is based on authors' previous successful experiences with similar methods. The method fundamentally applies a decision tree clustering analysis by using *exhaustive CHAID algorithm* (18) for the variables of interest. The variables are number of individual daily trips for different trip purposes and individual daily vehicle miles traveled. The models for each trip purpose are devised in a way to account for the correlation with the other trip purposes. The clusters are then fitted by the best probability distributions that can be later used for simulation. The current data transferability method is validated using the Phoenix-Mesa-Scottsdale region data where adequate number of observations were available as part of the add-on dataset in NHTS 2009.

The rest of the paper is structured as follows. First an overview of the previous studies on the topic of data transferability is presented. Then, various data sources that were available in this study are discussed. Following that, supporting methodology for model development is described and followed by testing results and a brief discussion about validation of models and their goodness-of-fit. Finally, conclusion and future research directions are discussed.

2- Background

Data transferability models are basically built upon data mining methods which can explore the data and detect the interdependencies and correlations among variables (4, 5).

In the literature, various models have been proposed to transfer disaggregate travel attributes using statistical methods. Mahmassani et al. (6) studied spatial transferability of trip frequency at three levels: area wide, zonal and household levels, for small urban areas in the State of Indiana. They compared cross-classified tables of trip frequencies among urban areas and their distributions for different trip purposes across different socio-economic groups. Walker and Olanipekun (7) used regression based cross-classification to model trip generation in a spatial transferability application. For the same application, Wilmot (8) used multiple linear regression models. Unlike the regression models that generate continuous results for discrete variables, Zhao (9) applied discrete choice models to come up with discrete results which could also account for more behavioral process of trip generation. Mahmassani and Sinha (10), Ben-Akiva and Bolduc (11), Mohammadian and Zhang (2) used Bayesian Updating approach to transfer spatial travel attributes. Mohammadian and Zhang (2) modeled trip rate and trip distance per person with gamma distribution in a spatial transfer application. However, assumptions for parameters of prior distribution in the Bayesian Updating approach add to the uncertainty of validity of the updated results. Long et al. (12) applied Small Area Estimation models to estimate household and census tract level travel characteristics like number of work trips for small and midsize metropolitan areas where few travel samples are available from various data sources. Small Area Estimation is referred to any statistical approach that uses a larger data to estimate model parameters for a small area where the sample size is too small. One of the limitations in their models is the inappropriate distribution of the household level travel attributes that does not match with the census tract level control total estimates.

The dependency between the travel attributes is another challenging issue which has been typically ignored in the transferability models (13, 19). For example, the number of recreational trips for an individual in a day might be dependent on the work trips for the individual in that day. This means that modeling the number of daily recreational trips and work trips independently could infer estimation bias to the results. There are exceptional studies that have attempted to study disaggregate trip rates for different trip purposes in the transferability context (14). Complexity of models, limited explanatory variables, and lack of accurate disaggregate models are among other limitations of transferability studies in the literature (2, 5). In an effort to improve the inefficiencies of previous travel attribute data transferability models, Rashidi and Mohammadian (14) presented household travel attribute models using an *Exhaustive CHAID* data mining algorithm to address several of these limitations.

CHAID is a type of decision tree technique based upon adjusted significance testing which can be used for prediction, classification and for detection of interactions between variables. It is a data mining algorithm that classifies the data into homogeneous clusters based on the dependent attribute variable. It can be used as an alternative to multiple linear regression or logistic regression analysis especially when the data is not appropriate for regression analysis. The *Exhaustive CHAID* has the same Splitting and Stopping steps as the *CHAID* but the Merging step is more exhaustive than *CHAID* algorithm. The splitting relies on the Chi-square test or F-test for either nominal dependent variable or continuous dependent variable respectively. It is noteworthy to mention that the modeling approach of this study is based upon assumption of continuous dependent variables which is practical and common in modeling discrete variables.

Following the previous research by Rashidi and Mohammadian (14), this paper applies *Exhaustive CHAID* algorithm with significant modifications compared to the previous approach. The current study attempts to explore and discuss a more disaggregate and policy sensitive individual-based data transferability approach by using a broad set of socio-demographic and land use variables. Utilizing the most recent National Household Travel Survey (NHTS 2009), the modeling approach of the previous data transferability framework is further enhanced in the current work by using a wide range of probability density functions (pdf) instead of only using normal/gamma distributions. Unlike the previous study which was developed at household level, the models of this paper are developed at individual level for more detailed trip purposes to better capture the behavior of household members. Furthermore the models take in to account the correlation between trip purposes in a directly sequential way versus the previous research which assumed an order among trip purposes with independent generation of trip rates for different trip purposes and the sequential subtraction of the rates from total number of trips.

These modifications revealed salient improvements in the simulated results compared to the results of the former data transferability approach, showing the effectiveness of the refined modeling framework. Travel attribute variables, modeled in this paper, include the total number of daily trips for eight different trip purposes as defined in the NHTS 2009 dataset, and the daily vehicle miles traveled. A summary of NHTS 2009 trip purposes are defined as work, school/religious/daycare, medical/dental services, shopping/errands, meals, family personal business/obligations, recreation/social and transporting someone. The return- home trips are not modeled in this study.

3. Data

National Household Travel Survey 2009 (NHTS) was the major data source for modeling exercise of this study. The NHTS is the main authoritative data source for American travel behavior which allows analysis of daily travel by all modes, characteristics of people traveling, their households and their vehicles. Since our previous study (14) was conducted on NHTS 2001; not only the new methodology but also the new NHTS dataset (2009) was a real motive for this research. The data includes the national sample along with the add-on data; however, in order to account for oversampling in the add-on areas, NHTS 2009 has applied weighting factors to adjust the data. Table 1 shows the explanatory variables used in this research. The variables are selected based on their common association with trip generation. From trip purpose summary variable in NHTS 2009 dataset, it is interesting to note that shopping trips have the highest frequency by including almost twenty percent of the records (after return-home trips which are not the focus in this study). Ranked second in frequency, are recreation and social trips with including 12 percent of the records. The next frequent trip purposes are work, meal, transport, school/religious, family and medical trips respectively. Although total daily work trips are less than daily shopping and recreational trips, for workers who are approximately consisting half of the records, work trips are the most frequent daily trips. For building the decision trees, final trip weight is applied to each trip record to adjust the data.

Table 1. Explanatory variables used in this study

Variable Definition	Categories/Frequency Percentage							
Race of HH respondent	White	African American	White and African American	Asian Only	Hispanic/Mexican	Other		
	85%	**6%**	**3%**	**2%**	**1%**	**3%**		
Driver status of the subject	Driver	Not Driver						
	91%	**9%**						
Respondent age	<18	18-30	30-40	40-55	55-70	>70		
	13%	**6%**	**10%**	**27%**	**28%**	**15%**		
Respondent gender	Male	Female						
	46%	**54%**						
Subject worker status	Worker	Not Worker						
	57%	**43%**						
Number of drivers in HH	0	1	2	3	4	>4		
	3%	**22%**	**43%**	**7%**	**22%**	**4%**		
Derived total HH income	<10k $	10k< <20k	20k< <30k	30k< <40k	40k< <50k	50k< <60k	60k< <70k	>70k
	6%	**11%**	**12%**	**11%**	**10%**	**9%**	**7%**	**36%**
Count of HH members	1	2	3	4	5	6	>6	
	11%	**41%**	**17%**	**18%**	**8%**	**3%**	**2%**	
Life Cycle classification for the HH	1 adult, no children	2+ adults, no children	1 adult, youngest child 0-5	1 adult, youngest child 0-5	1 adult, youngest child 6-15	2+ adults, youngest child 6-15	Others	
	5%	**20.5%**	**0.5%**	**12%**	**2%**	**21%**	**40%**	
Count of HH vehicles	0	1	2	3	4	5	6	>6
	5%	**27%**	**41%**	**18%**	**6%**	**2%**	**0.5%**	**0.5%**
Count of adults at least 18 years old	1	2	3	4	>4			
	25.6%	**62.8%**	**8.9%**	**2.3%**	**0.4%**			
Number of workers in HH	0	1	2	>2				
	29%	**34%**	**30%**	**7%**				
Highest grade completed	Less then high school graduate	High school graduate	Some college or Associate's degree	Bachelor's degree	Graduate Degree			
	8%	**28%**	**28%**	**21%**	**15%**			
Size of urban area (Population)	50,000 - 199,999	200,000 - 499,999	500,000 - 999,999	1 million or more without subway or rail	1 million or more with subway or rail	Not in an urbanized area		
	14%	**9%**	**8%**	**19%**	**11%**	**39%**		
Household in urban/rural area	Urban	Rural						
	70%	**30%**						

In order to validate the transferability capability of the models, one of the NHTS 2009 add-on regions, the Phoenix Metropolitan Area was chosen with a sample size of more than 4200 households. Table 2 shows the mean and standard deviation of the individual travel attributes modeled in this study, both for the national and the Phoenix Metropolitan Area. One of the advantages of the modeling methodology of this study is that trips that are not frequently occurring during a day can also be modeled with an acceptable prediction potential. Traveling to medical centers and attending family related meetings are not frequent activities and modeling the occurrence of these types of activities requires careful attention. It will be shown in the results section that the modeling approach can acceptably estimate even these less frequent activities. Daily trips for different purposes as well as vehicle miles traveled for each trip are extracted from the travel day trip file which included 1,167,321 trip records. For analyzing each trip purpose, the national observations are split into two divisions. Seventy percent of the observations are randomly selected for the calibration and training purpose and the remaining records are used to validate the model.

Table 2. Individual Travel attributes

Variable	Mean		S.D.	
	National	Phoenix	National	Phoenix
Number of Daily Family Trips	**0.16**	**0.09**	**0.49**	**0.34**
Number of Daily Meal Trips	**0.33**	**0.36**	**0.58**	**0.6**
Number of Daily Medical Trips	**0.09**	**0.04**	**0.32**	**0.22**
Number of Daily Recreational Trips	**0.53**	**0.46**	**0.85**	**0.75**
Number of Daily School/Religious Trips	**0.21**	**0.24**	**0.48**	**0.5**
Number of Daily Shopping Trips	**0.86**	**0.44**	**1.2**	**0.83**
Number of Daily Transporting-someone Trips	**0.25**	**0.29**	**0.71**	**0.78**
Number of Daily Work Trips (for workers)	**0.9**	**0.85**	**1.06**	**0.96**
Daily VMT	**37.2**	**31.8**	**56.16**	**43.5**

4. Methodology

The primary focus of this paper is on trip generation. Vehicle Miles Traveled (VMT) is also modeled as a complementary part to the trip generation models. Eight different trip purposes including work, school, medical, shopping, meal, family, recreation, and transporting-someone trips plus VMT are modeled using *Exhaustive CHAID* rule based data mining approach.

Data mining approaches have now been increasingly applied in diverse academic fields for predicting purposes. Data mining is defined as an analytical approach used to explore large datasets to achieve consistent and systematic interdependencies among the variables. With the availability of storage of large datasets, data mining approaches are gaining more popularity among researchers. Among data mining approaches are rule based methods which construct the relationship between the dependent variable and the independent ones over certain rules. Such rules cause the data to be categorized in a number of clusters in a way that data in each cluster share homogenous attributes. Some of the data mining methods applied for obtaining optimal decision tree are Chi-squared Automatic Interaction Detector (CHAID) (20), Classification and Regression Tree

(C&RT) (21), and Quick, Unbiased and Efficient Statistical Tree (QUEST) (22). In this study exhaustive CHAID is utilized to obtain optimal decision trees. *Exhaustive CHAID*, as a comprehensive data mining decision tree algorithm, recognizes the relationship between the dependent variable and most significant independent variables efficiently and helps the tree to grow branches based on distinguishing homogenous leaves. As an example, if the number of daily work trips made by an individual as the dependent variable is mainly dependent on the household size, the *Exhaustive CHAID* tree grows with several branches of household size categories. In this study, number of daily trips is modeled as a continuous variable; therefore, ANOVA F-test is used to calculate F-statistic to test if means of branches are statistically different. It could be argued that Decision Tree is a good alternative approach to conventional regression models since it has the capability of modeling various categories of data.

As mentioned earlier, this paper attempts to model an individual's trip rates per day for eight different trip purposes. In order to do that, a number of socio-demographic and land use explanatory variables are selected from NHTS 2009 dataset as the explanatory variables. Counting daily trips has been considered as a common transferable travel variable in some of the previous studies (2, 10). In order to consider potential correlation among trip purposes, each trip purpose tree can first be clustered by the previous modeled trip purposes meaning that to generate trips, a sequence among trip purposes is assumed (14). This follows the traditional assumption of classifying trips into sequence of mandatory, maintenance and discretionary trips. It is assumed that mandatory trips including work and school are modeled first, then the maintenance trips including family, shopping and medical are modeled by clustering on the basis of the number of mandatory trips, and finally discretionary trips, meal, transporting someone, and recreational trips are modeled conditional to the number of maintenance trips. The proposed methodology maintains the correlation among various trip purposes. It is noteworthy to remind that recent research suggests a dynamic activity planning process and criticizes the assumption of fixed order of activity planning (15). Incorporating dynamic order of activities in a trip rate transferability model would be a challenging research and remains as a future task.

As noted earlier, the sequential decision tree models are developed on seventy percent of the national trip file. Once the decision trees are developed, the distinguished homogeneous clusters are fitted by the best distribution. The distribution fitting process is performed by employing a statistical and mathematical package called EasyFit (23) to test several probability density functions and selects the best fitted distributional forms. In order to evaluate the quality of the fitting exercise, the Kolmogorov-Smirnov (K-S) test (16, 17) is used. In this case, the K-S test is applied to examine whether the two probability distributions (transferred and observed) are equal. A simulation code is then implemented to generate number of trips for each one of the eight trip purpose categories based on random number generation from the corresponding inverse distributions. These simulated values are then summed up for each individual to be evaluated against the observed total daily trips for the individual. For simulating the trip rates for the mandatory trip purposes, the simulation code is executed directly while the trip rate simulation for maintenance trip purposes requires the result of the mandatory trip rate as an input. Subsequently, simulation of the trip rates for the discretionary trip purposes is executed after the maintenance trip rate is obtained from the previous simulation step.

The rest of the paper presents the daily trip rates modeling results for each of the trip purposes, total daily trip rates and eventually daily VMT model which are followed by validation and goodness of fit section.

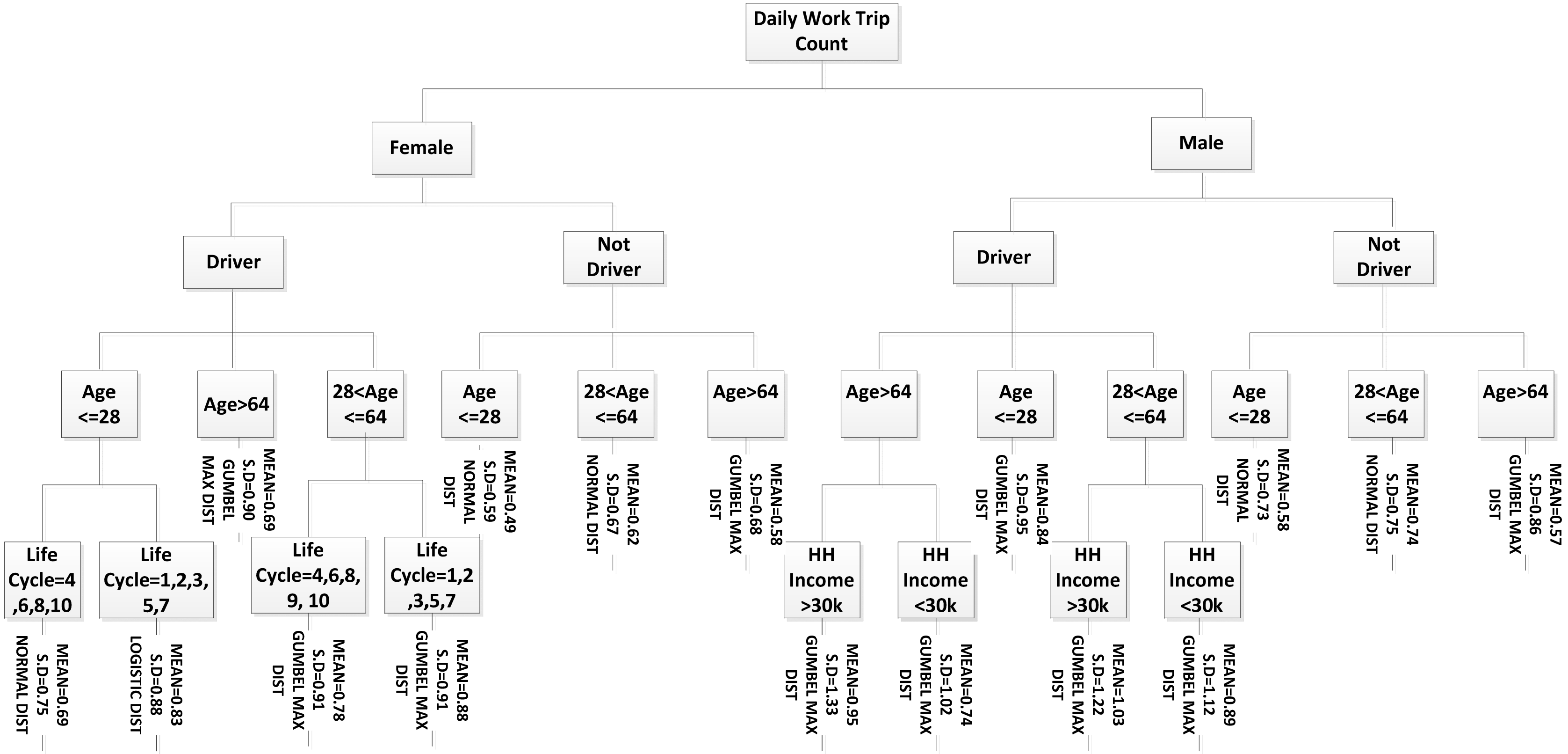
5. Model Results (Train and Test)

In this section, the daily trip rate decision trees for trip purposes are presented. Moreover, an example of improving the trip rate models by using the best fitted distribution against the simplified approach of using normal/gamma distribution is shown to emphasize the importance of the presented methodology. Also, several decision trees for various trip purposes are displayed and discussed. The decision trees are developed based on the following three rules:

1) There must be a minimum number of 500 records in a child node and a minimum number of 1000 records for the parent node.
2) Each tree is composed of maximum 4 depth levels.
3) P-value of 0.05 is considered as a test on the difference between the node mean values.

Figure 1 presents the best fitted decision tree which is developed for the number of daily work trips for employed individuals. For estimating the number of daily work trips, variables of gender, driver status, age, household income and household life cycle classification for an individual were found to be significant. The most significant variable in forming a work trip is gender. On average, male individuals have higher daily work trips than female individuals. Following that, being a driver is the next significant variable in number of daily work trips. Driver individuals have significantly higher number of daily work trips on average throughout the nation. An individual’s age is the other significant variable. The national data shows that the middle group of age in the range of 28 and 64 has the highest number of daily work trips on average for both genders. It is interesting that female individuals under the age of 28 have higher average daily work trips than female higher-aged group while male individuals under the age of 28 have almost the same average daily work trips than male higher-aged group. The tree split stops at age variable for the non-driver individuals because the sample size for the non drivers would be small; therefore, no further significant variable is detected. On the other side of the tree which includes female driver individuals, Life Cycle is an important variable while on the male driver individual side, household income plays an important role. These clusters make sense because the female individuals who either have children or live with their partners tend to have less work trips. On the male side, higher income individuals make more daily work trips. It should be noted that after fitting distributions to end nodes of the tree, it was found that majority of the best fitted distributions are Gumbel Max while there are a few nodes with Normal and Logistic distributions.

Figure 1. Decision tree clusters for work trip per day



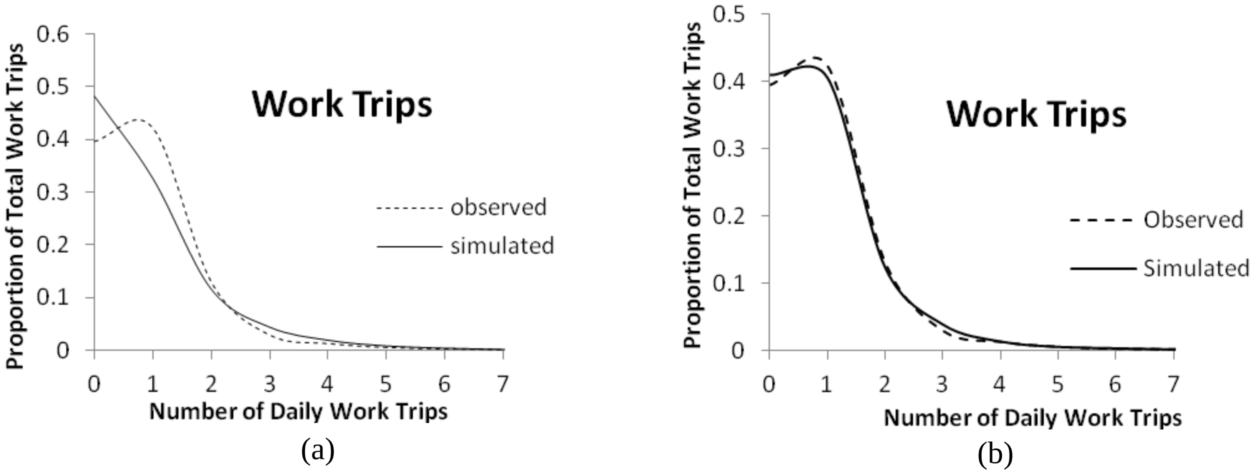
Daily Work Trip Count
Female
Male
Driver
Not Driver
Driver
Not Driver
Age <=28
Age>64
28<Age <=64
Age <=28
28<Age <=64
Age>64
Age>64
Age <=28
28<Age <=64
Age <=28
28<Age <=64
Age>64
Life Cycle=4 ,6,8,10
Life Cycle=1,2,3, 5,7
MEAN=0.69 S.D=0.90 GUMBEL MAX DIST
Life Cycle=4,6,8, 9, 10
Life Cycle=1,2 ,3,5,7
MEAN=0.49 S.D=0.59 NORMAL DIST
MEAN=0.62 S.D=0.67 NORMAL DIST
MEAN=0.58 S.D=0.68 GUMBEL MAX DIST
HH Income >30k
HH Income <30k
MEAN=0.84 S.D=0.95 GUMBEL MAX DIST
HH Income >30k
HH Income <30k
MEAN=0.58 S.D=0.73 NORMAL DIST
MEAN=0.74 S.D=0.75 NORMAL DIST
MEAN=0.57 S.D=0.86 GUMBEL MAX DIST
MEAN=0.69 S.D=0.75 NORMAL DIST
MEAN=0.83 S.D=0.88 LOGISTIC DIST
MEAN=0.78 S.D=0.91 GUMBEL MAX DIST
MEAN=0.88 S.D=0.91 GUMBEL MAX DIST
MEAN=0.95 S.D=1.33 GUMBEL MAX DIST
MEAN=0.74 S.D=1.02 GUMBEL MAX DIST
MEAN=1.03 S.D=1.22 GUMBEL MAX DIST
MEAN=0.89 S.D=1.12 GUMBEL MAX DIST

Note: Life cycle classifications are defined as the following: (It seems that NHTS have included the retired workers who are still working.)

01 = one adult, no children
02 = 2+ adults, no children
03 = one adult, youngest child
04 = 2+ adults, youngest child
05 = one adult, youngest child
06 = 2+ adults, youngest child
07 = one adult, youngest child
08 = 2+ adults, youngest child
09 = one adult, retired, no
10 = 2+ adults, retired, no

As can be seen in Figure 1, normal/gamma distributions are present as the best fitted distributions in few terminal nodes. To further elucidate the significance of searching for the best distribution at the terminal nodes, a simple analysis is presented. Figure 2 displays the daily work trip model results based on the best fitted distributions of Figure 1 and the pre-assumed gamma distribution. Gamma distribution is typically selected in the literature to avoid simulating negative values (2). The validation results presented in Figure 2 are prepared by using the 30% test data which was put aside for the validation purpose. The figures schematically show the superiority of the new simulation approach and the best fitted distributions (as shown in Figure 1) over the simplified application of a pre-assumed gamma distribution.

Figure 2. Daily work trip models comparison between general gamma distribution (a), and best fit distribution (b)

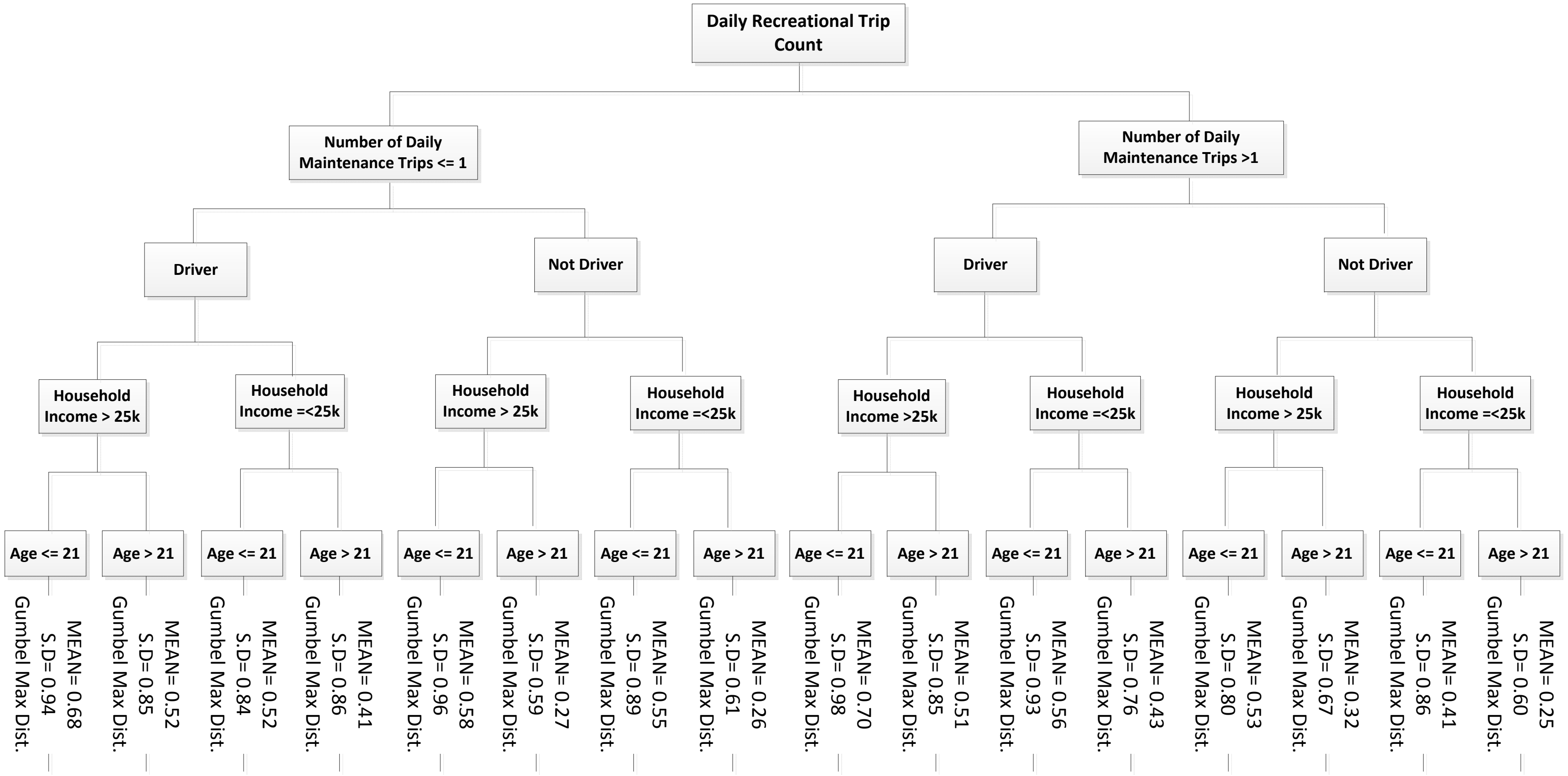


(a)

(b)

Similar modeling approach was undertaken to model the other seven trip purposes that are reported in the NHTS 2009 dataset. Due to word count limitations only a selection of the developed models are presented in complete detail. Figure 3 shows the details of the decision tree model for the daily recreational trips for individuals. Since recreational trips are categorized as discretionary trips, the tree is initially clustered by the total number of maintenance trips to account for the possible correlation between recreational and maintenance trips. In a similar way, the impact of mandatory trips is considered when maintenance trip trees are being generated and the correlation between recreational and mandatory trips is considered in an indirect modeling approach. Therefore, the interdependencies among trip counts for different trip purposes are maintained through a sequence. The *Exhaustive CHAID* algorithm branches the tree into two leaves by the number of maintenance trips either equal/less than one or more than one recreational trips. In the next level, the tree is branched by driver status variable. Being a driver comes out a significant variable in the number of daily trips for recreational trips as well as most of the other trip purpose trees. If an individual drives, the higher number of daily trips by the individual is expected which is consistent with common sense.

Figure 3. Decision tree clusters for recreational trip rate per day.


Daily Recreational Trip Count
Number of Daily Maintenance Trips <= 1
Number of Daily Maintenance Trips >1
Driver
Not Driver
Driver
Not Driver
Household Income > 25k
Household Income =<25k
Household Income > 25k
Household Income =<25k
Household Income >25k
Household Income =<25k
Household Income > 25k
Household Income =<25k
Age <= 21
Age > 21
Age <= 21
Age > 21
Age <= 21
Age > 21
Age <= 21
Age > 21
Age <= 21
Age > 21
Age <= 21
Age > 21
Age <= 21
Age > 21
Age <= 21
Age > 21
MEAN= 0.68 S.D= 0.94 Gumbel Max Dist.
MEAN= 0.52 S.D= 0.85 Gumbel Max Dist.
MEAN= 0.52 S.D= 0.84 Gumbel Max Dist.
MEAN= 0.41 S.D= 0.86 Gumbel Max Dist.
MEAN= 0.58 S.D= 0.96 Gumbel Max Dist.
MEAN= 0.27 S.D= 0.59 Gumbel Max Dist.
MEAN= 0.55 S.D= 0.89 Gumbel Max Dist.
MEAN= 0.26 S.D= 0.61 Gumbel Max Dist.
MEAN= 0.70 S.D= 0.98 Gumbel Max Dist.
MEAN= 0.51 S.D= 0.85 Gumbel Max Dist.
MEAN= 0.56 S.D= 0.93 Gumbel Max Dist.
MEAN= 0.43 S.D= 0.76 Gumbel Max Dist.
MEAN= 0.53 S.D= 0.80 Gumbel Max Dist.
MEAN= 0.32 S.D= 0.67 Gumbel Max Dist.
MEAN= 0.41 S.D= 0.86 Gumbel Max Dist.
MEAN= 0.25 S.D= 0.60 Gumbel Max Dist.

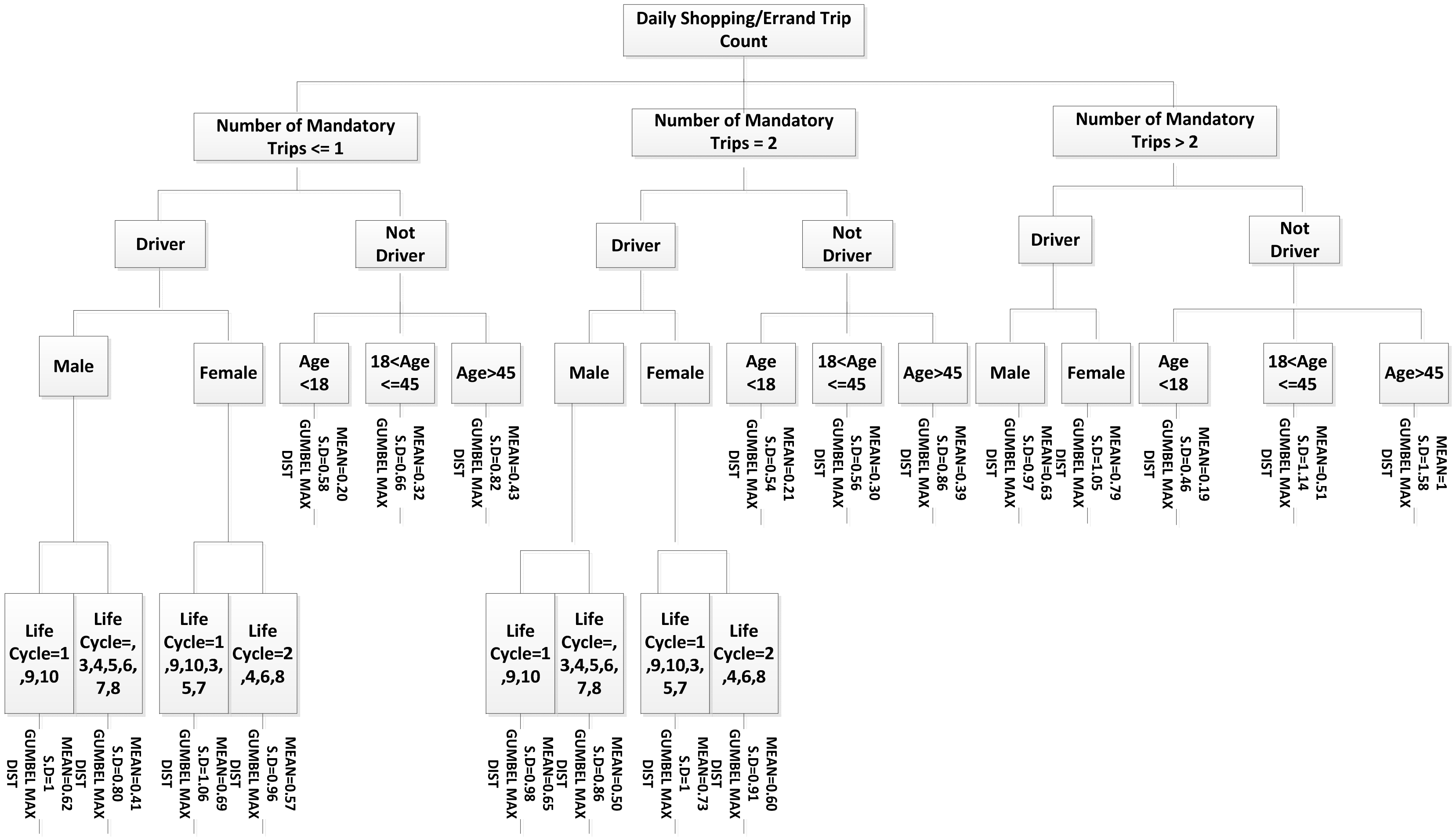
Household income clearly has an impact on the number of recreational trips. Individuals with higher household income have more resources to attend recreational and social events. This fact can be observed in the terminal tree nodes. Finally, age is the final variable which is significant in the number of daily recreational trips. Individuals under the age of 21 make more recreational trips than the older group. All of the best fitted distributions for the final nodes are Gumbel Max.

As mentioned earlier in the data section, shopping/errand trip is the most frequent trip purpose among all people. Figure 4 displays the decision tree for the shopping trips. The tree associated with shopping trip is initially clustered by the number of mandatory trips to consider the correlation between the number of mandatory and shopping trips. It seems that the number of shopping trips generally increases as individuals make more mandatory trips. This might be caused by the higher opportunity for accomplishing maintenance activities when an individual can chain trips. Like the previous models, being a driver directly affects the number of shopping trips. For individuals who do not drive, age is the next significant factor. According to the trained data, the individuals in the age group over 45 make the most daily shopping trips. The middle age group and the teenager group are the next. It is intutive that an individual under 18 makes the least daily shopping trips on average because of less responsibility and wealth. For the driver categories, it is important if the individual is a male or a female. Female individuals generate significantly higher daily shopping trips. Life cycle classification is the next level that shows association with number of shopping trips. Single female individuals usually generate more shopping trips than the other female individuals probably because shopping trips are shared with the partners for the latter group. On the other hand, single male individuals with no children and retired male individuals generate more shopping trips than other males. All of the best distributions are Gumbel Max for the terminal nodes.

Figure 5 shows the validation result for recreational and shopping trips as well as the validation results of the daily trip rate for few other trip purposes that are modeled in the same fashion as discussed earlier. To be concise, the decision trees for the other trip purposes are not included in this paper; but the testing results of the trip purposes of medical, meal, family and transport models are shown.

The result for the simulated recreational trips per day is consistent with the observed test data. The testing result shows less than three percent over generating the number of 1 daily recreational trip while underestimating the number of 2-3 daily recreational trips by less than three percent. Although three percent of the total number of recreational trips could be a significant error, still, the result shows significant improvements compared to the previously proposed transferability models. The simulated and observed number of shopping trips are also presented in figure 5. It seems that the simulated model slightly overestimates the number of zero and one daily shopping trips and underestimates the number of two and three daily shopping trips. However, such a small difference is limited to less than three percent of the total shopping trips.

Figure 4 Decision tree clusters for shopping trip rate per day.


Daily Shopping/Errand Trip Count
Number of Mandatory Trips <= 1
Number of Mandatory Trips = 2
Number of Mandatory Trips > 2
Driver
Not Driver
Driver
Not Driver
Driver
Not Driver
Male
Female
Age <18
18<Age <=45
Age>45
Male
Female
Age <18
18<Age <=45
Age>45
Male
Female
Age <18
18<Age <=45
Age>45
MEAN=0.20 S.D=0.58 GUMBEL MAX DIST
MEAN=0.32 S.D=0.66 GUMBEL MAX
MEAN=0.43 S.D=0.82 GUMBEL MAX DIST
MEAN=0.21 S.D=0.54 GUMBEL MAX DIST
MEAN=0.30 S.D=0.56 GUMBEL MAX DIST
MEAN=0.39 S.D=0.86 GUMBEL MAX DIST
MEAN=0.63 S.D=0.97 GUMBEL MAX DIST
MEAN=0.79 S.D=1.05 GUMBEL MAX DIST
MEAN=0.19 S.D=0.46 GUMBEL MAX DIST
MEAN=0.51 S.D=1.14 GUMBEL MAX DIST
MEAN=1 S.D=1.58 GUMBEL MAX DIST
Life Cycle=1 ,9,10
Life Cycle=, 3,4,5,6, 7,8
Life Cycle=1 ,9,10,3, 5,7
Life Cycle=2 ,4,6,8
Life Cycle=1 ,9,10
Life Cycle=, 3,4,5,6, 7,8
Life Cycle=1 ,9,10,3, 5,7
Life Cycle=2 ,4,6,8
MEAN=0.62 S.D=1 GUMBEL MAX DIST
MEAN=0.41 S.D=0.80 GUMBEL MAX DIST
MEAN=0.69 S.D=1.06 GUMBEL MAX DIST
MEAN=0.57 S.D=0.96 GUMBEL MAX DIST
MEAN=0.65 S.D=0.98 GUMBEL MAX DIST
MEAN=0.50 S.D=0.86 GUMBEL MAX DIST
MEAN=0.73 S.D=1 GUMBEL MAX DIST
MEAN=0.60 S.D=0.91 GUMBEL MAX DIST

Note: For Life Cycle classifications refer to the work decision tree figure.

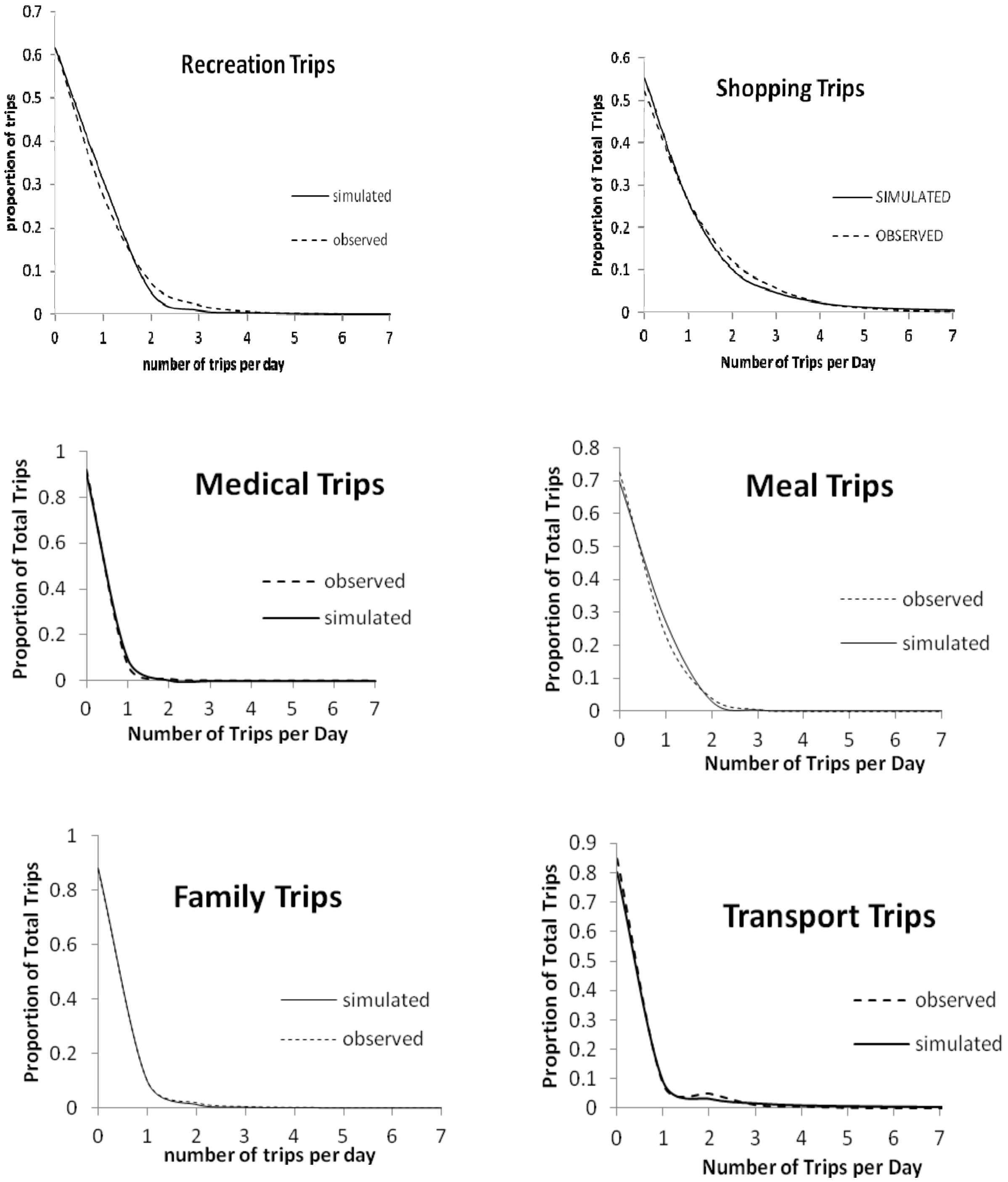
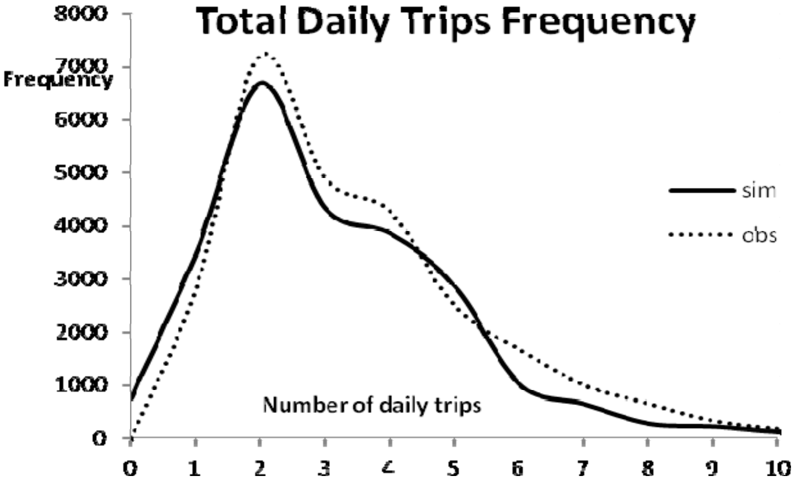


Figure 5. The results of simulated daily trip rates for various trip purposes versus the observed trip rates over the test data

Eventually, the total daily trip rates are tested over 5% of the total population (NHTS data) and the result is presented in Figure 6. Trip rates for different trip purposes are calculated for each individual from the developed models and then summed up to come up with the total number of trips per day for an individual. Figure 6 shows an acceptable model simulation graph over the observed data. The simulation slightly overestimates the number of zero and one total trips for an individual and slightly underestimates the number of 2+ daily trip rates. The difference between the simulated and the observed graphs is more apparent because of the aggregation error. Although this is expected, the result shows an acceptable pattern. All the efforts so far have been to graphically accept the models until later section of the paper which helps to statistically validate the models.

Figure 6. Total daily trip frequency test result



Calculated trip distance for vehicle trips (VMT) per day is another variable which is modeled for an individual. This variable is extracted from the travel day trip file; but, each record represents the VMT associated with a special trip purpose for an individual; therefore, the travel day trip file is aggregated to calculate the total VMT for all trip purposes. Figure 7 shows the decision tree for individual-based daily VMT. After the individual's gender which is the most significant factor in determining VMT, urban or rural classification of the individual's residential location affects VMT. According to the data clusters, the average daily VMT for individuals living in rural areas is higher than those living in urban areas probably because people in urban areas are closer to locations of interest while rural residents need to drive further distances. It is noteworthy that being a worker or not is also an important factor on the recorded daily VMT. Being a worker is generally associated with higher VMT which does clearly make sense.

Figure 7. Decision tree clusters for Individual daily VMT

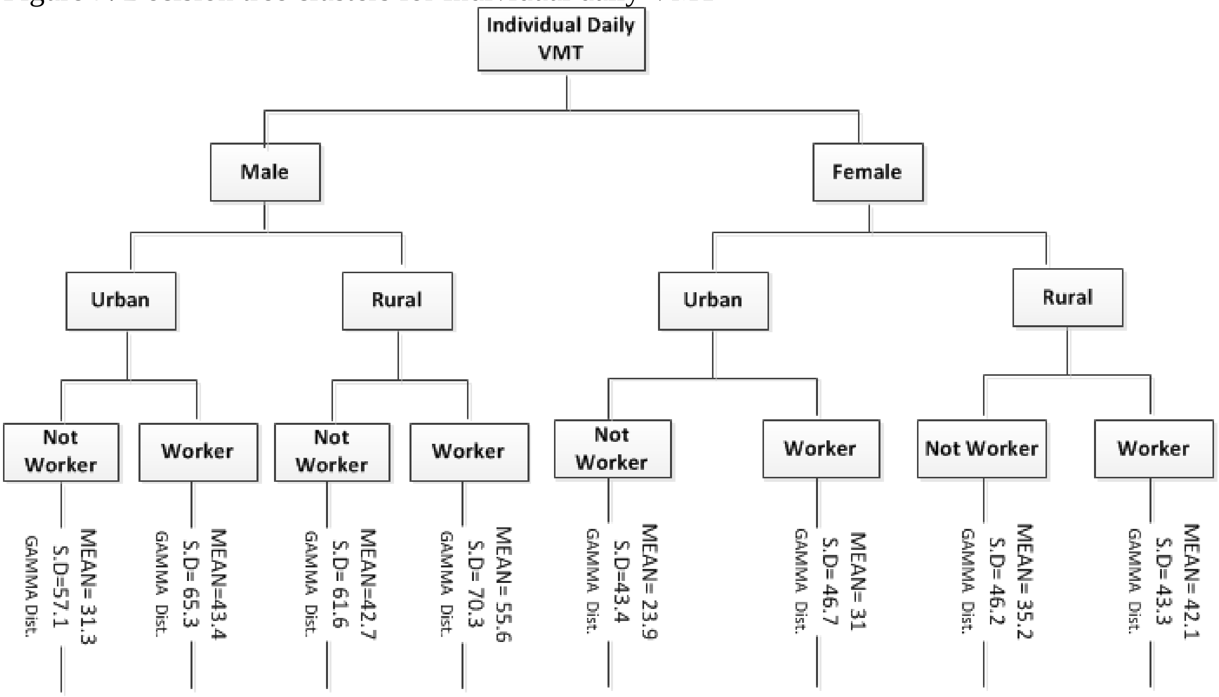
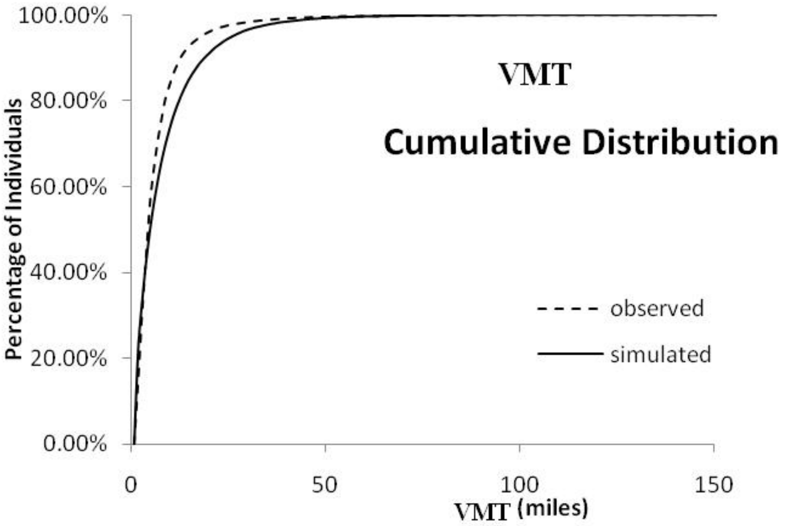

Individual Daily VMT
Male
Female
Urban
Rural
Urban
Rural
Not Worker
Worker
Not Worker
Worker
Not Worker
Worker
Not Worker
Worker
MEAN= 31.3 S.D=57.1 GAMMA Dist.
MEAN=43.4 S.D= 65.3 GAMMA Dist.
MEAN=42.7 S.D= 61.6 GAMMA Dist.
MEAN= 55.6 S.D= 70.3 GAMMA Dist.
MEAN= 23.9 S.D=43.4 GAMMA Dist.
MEAN= 31 S.D= 46.7 GAMMA Dist.
MEAN= 35.2 S.D= 46.2 GAMMA Dist.
MEAN= 42.1 S.D= 43.3 GAMMA Dist.

Figure 8 illustrates the VMT cumulative distribution result for both of the observed and simulated data. The maximum difference between the two graphs is within the first 50 miles which is the range where most of the daily trips take place. Although a cumulative error difference of less than 10 percent between the two graphs in some points could not be ignored, the result is comparable with other transferability models for VMT. The VMT decision tree model seems to slightly overestimate the individual daily VMT which needs to be addressed in future work.

Figure 8. Individual daily VMT test result



6. Transferability Application

One of the common applications of spatial transferability is to transfer data from a larger geography to smaller regions, and to validate the performance of the transferred data in the new context. In this study, the national decision tree models are applied to one of the NHTS 2009 add-on regions, Phoenix Metropolitan Area. Phoenix was selected because a reasonable sample with more than 4200 household records was available for validation purpose. The mandatory trip models, including work and school trip, are first executed and the numbers are summed up as the total number of mandatory trips for the individuals to be used by the maintenance trip models which are medical, shopping and family trips. Then, the total number of maintenance trips is summed up to be used by the discretionary trip models of recreational, transporting others, and meal trips. Finally, the total number of daily trips for each individual is summed up and it was compared with the observed data. Two-sample Kolmogorov–Smirnov test was utilized to test whether the two underlying probability distributions of the transferred and the observed data for all trip purposes are equal. In this case the K-S statistic is $D_{n,n'}$so that:

$$D_{n,n'} = \sup_x |F_{1,n}(x) - F_{2,n'}(x)|,$$

$n = observed\ sample\ size \quad n' = Transferred\ sample\ size$;

$F_{1,n}$ is the cumulative distribution function for the observed data;

$F_{2,n'}$ is the cumulative distribution function for the transferred data; and

sup x is the supremum of the set of distances.

It is clear that n and n' are equal in this case because the number of transferred sample and the observed one are equal. For conducting this statistical test, transferred and observed trip data for various trip purposes were fitted by the best distribution with the help of EasyFit software and the corresponding cumulative distribution functions were extracted to maximize the absolute value of the subtraction function ($D_{n,n'}$). Table 3 shows the K-S statistic results given the null hypothesis that the observed and transferred distributions are similar. The hypothesis concerning the distributional type is rejected at the selected significance level (0.05) if the test statistic,$D_{n,n'}$ is bigger than the critical value taken from the K-S table. The values show that all of the test statistics are smaller than the critical statistic meaning the null hypothesis regarding the equality of the two distributions could not be rejected at the 0.05 significance level. This example could show the capability of the tree models in transferring trip counts from a national level dataset to smaller geographies like a metropolitan area.

Table 3 Comparison of the observed and transferred distributions and their K-S statistics

Decision Tree	Observed Distribution	Transferred Distribution	Test Statistic	n (sample size)	Critical Statistic =1.36×((n+n')/nn')^0.5	Reject?
Work	GUMBEL MAX DIST. σ=0.84 μ=0.48	GUMBEL MAX DIST. σ=0.76 μ=0.41	0.022	3868	0.031	NO
School	GUMBEL MAX DIST. σ=0.38 μ=0.01	GUMBEL MAX DIST. σ=0.38 μ=0.02	0.014	8139	0.021	NO
Medical	GUMBEL MAX DIST. σ=0.16 μ= -0.05	GUMBEL MAX DIST. σ=0.17 μ= - 0.05	0.01	3946	0.031	NO
Shopping	GUMBEL MAX DIST. σ=0.65 μ= 0.08	GUMBEL MAX DIST. σ=0.64 μ= 0.19	0.021	3945	0.031	NO
Recreational	GUMBEL MAX DIST. σ=0.58 μ= 0.15	GUMBEL MAX DIST. σ=0.61 μ= 0.22	0.016	4521	0.029	NO
Family	GUMBEL MAX DIST. σ=0.27 μ= - 0.03	GUMBEL MAX DIST. σ=0.27 μ= - 0.06	0.018	3945	0.031	NO
Transport	GUMBEL MAX DIST. σ=0.60 μ= 0.06	GUMBEL MAX DIST. σ=0.53 μ= 0.07	0.022	4521	0.029	NO
Meals	GUMBEL MAX DIST. σ=0.46 μ= 0.09	GUMBEL MAX DIST. σ=0.47 μ= 0.12	0.017	4521	0.029	NO
Total Daily Trip	GAMMA DIST. α = 2.77 β= 1.28	GAMMA DIST. α = 2.52 β= 1.11	0.008	37406	0.010	NO
Daily VMT	GAMMA DIST. α = 0.52 β= 61	GAMMA DIST. α = 0.47 β= 58	0.019	5477	0.026	NO

7. Conclusion

This study introduced an individual-level travel attribute transferability model that applies the exhaustive CHAID data mining algorithm, a Decision Tree approach is implemented to cluster the 2009 NHTS dataset into homogenous groups based on the number of daily trips for different purposes. Then each group is modeled by fitting the best statistical distribution to the cluster. Consequently, inverse CDF functions are used to simulate interested variables for each individual. The models are designed in a way that the correlations between trip rates for different trip purposes are taken in to account in a sequential manner.

Models are calibrated using 70 percent of the data while validation is performed on the remaining 30 percent. To validate the approach, models are utilized to transfer trip counts from national NHTS to Phoenix Metropolitan Area. The results show that the simulated dataset is consistent with the observed data in the Phoenix add-on sample, suggesting that the models are capable of estimating travel attributes with a high level of confidence. There are numerous immediate applications for these models. They can replace trip generation step in traditional 4-step models. In addition, they can be integrated with activity generation step in the activity-based models. Finally, daily VMT for an individual was modeled as a function of socio-demographic, land-use, and built environment variables. The approach allows researchers to accurately estimate

VMT that can be used to examine various issues like environmental effects and energy consumption, as well as person-level carbon footprints.

8. Reference

1. Axhausen, K., Garling, T.(1992), Activity-based approaches to travel analysis: conceptual frameworks, models, and research problems. Transp. Rev. 12(4), 323–341
2. Zhang, Y., Mohammadian, A. (2008), Bayesian updating of transferred household travel data.Transportation Research Record. 2049, 111–118 (2008)
3. Volpe National Transportation Systems Center (2004), U.S. Department of Transportation. *Summary Report for the Peer Exchange on Data Transferability*. http://media.tmiponline.org/clearinghouse/tmip/peer_exchange/2004-12-16/
4. Stopher, P.R., Greaves, S., Bullock, P. (2003), Simulating household travel survey data:application to two urban areas. Proceedings of the 82th Annual Meeting of the Transportation Research Board, Washington DC (2003)
5. Reuscher, T.R., Schmoyer, R.L., Hu, P.S. (2002), Transferability of nationwide personal transportation survey data to regional and local scales, Transportation Research Record, 1817, 25–35
6. Mahmassani, H.S., Bevilacqua, O. and Sinha, K. (1979). Framework for Transferring Travel Characteristics of Small Urban Areas. Transportation Research Record 730, 1979.
7. Walker W.T. and O. A. Olanipekun (1989). Interregional Stability of Household Trip Generation Rates from the 1986 New Jersey Home Interview Survey, Transportation Research Record 1220 (TRR) , 1989, pp.47-57.
8. Wilmot, C.G. (1995). Evidence of Transferability of Trip Generation Models. *Journal of Transportation Engineering* 9:405–410.
9. Zhao, H. (2000). Comparison of Two Alternatives for Trip Generation, paper presented at the 79th Annual Meetings of the Transportation Research Board, Jan. 9–13, 2000, Washington, DC.
10. Mahmassani, H.S., Sinha, K.C.(1981). Bayesian updating of trip generation parameters. Transp. Eng. J. 107(TE5), 581–589
11. Ben-Akiva, M., and Bolduc, D. (1987). Approaches to Model Transferability and Updating: The Combined Transfer Estimator, Transportation Research Record No. 1139, 1-7.
12. Long L., Lin J., and Pu W. (2009). Model-Based Synthesis of Household Travel Survey Data in Small and Midsize Metropolitan Areas, Transportation Research Record: Journal of the Transportation Research Board, No. 2105, Transportation Research Board of the National Academies, Washington, D.C., 2009, pp. 64–70.
13. Wilmot, C.G., Stopher, P.R. (2001), Transferability of transportation planning data. Transportation Research Record 1768, 36–43
14. Rashidi, T.H. , Mohammadian, A.(2011), Household travel attributes transferability analysis: application of a hierarchical rule based approach, Published online: 17 April 2011, Springer Science+Business Media, LLC. 2011, 38:697–714
15. Auld, J. A. and Mohammadian A. (2009). Framework for the development of the Agen-based Dynamic Activity Planning and Travel Scheduling (ADAPTS) model. Transportation Letters: The International Journal of Transportation Research, 1 (3), 243-253.

16. Chakravarti I.M., R.G. Laha, J. Roy, 1967, Handbook of Methods of Applied Statistics, vol. I, John Wiley and Sons.

17. Eadie, W.T., Drijard D., James F.E., Roos M. and Sadoulet B., (1971), Statistical Methods in Experimental Physics. Amsterdam: North-Holland, 269-271.

18. Biggs, D.B., de Ville, B., Suen, E. (1991), A method of choosing multi-way partitions for classification and decision trees. J. Appl. Stat. 8, 49–62

19. Golob, T.F. (2000), A simultaneous model of household activity participation and trip chain generation. Transportation Research B 34, 355–376

20. Kass, G.V.(1980) An exploratory technique for investigating large quantities of categorical data. Appl. Stat. 29, 119–127

21. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.(1984), Classification and Regression Trees. Wadsworth, Belmont, CA

22. Loh, W.Y., Shih, Y.S. (1997), Split selection methods for classification trees. Stat Sin 7, 815– 840 (1997)

23. Mathwave (2004). *Mathwave Data Simulation and Analysis*. (n.d.). Retrieved 2004, from EasyFit – Distribution Fitting Made Easy: http://www.mathwave.com/ last accessed July 2011