INNOVATIVE DATA COLLECTION AND MODELING METHODS FOR LONG-DISTANCE PASSENGER TRAVEL DEMAND ANALYSIS

The Pennsylvania State University ❖ University of Maryland
University of Virginia ❖ Virginia Polytechnic Institute and State University
West Virginia University

University of Maryland ❖ Department of Civil and Environmental Engineering ❖ College Park, Maryland 20742
Phone: 301-405-2881 ❖ Fax: 301-405-2585
www.pti.psu.edu/mautc

INNOVATIVE DATA COLLECTION AND MODELING METHODS FOR LONG-DISTANCE PASSENGER TRAVEL DEMAND ANALYSIS

FINAL REPORT

Contract No. DTRT07-G-0003

Prepared for

Mid-Atlantic Universities Transportation Center

By

Lei Zhang, Assistant Professor
Yijing Lu, Ph.D, Graduate Research Assistant

Department of Civil and Environmental Engineering,
University of Maryland
College Park, MD 20742

October 2012

This work was sponsored by the Mid-Atlantic Universities Transportation Center. The contents of this report reflect the views of the authors, who are responsible for the facts and the accuracy of the data presented herein. This report does not constitute a standard, specification, or regulation.

1. Report No.

UMD-2010-01

2. Government Accession No.

3. Recipient's Catalog No.

4. Title and Subtitle

Innovative Data Collection and Modeling Methods for Long-Distance Passenger Travel Demand Analysis

5. Report Date

October 2012

6. Performing Organization Code

7. Author(s) Lei Zhang and Yijing Lu

8. Performing Organization Report No.

9. Performing Organization Name and Address

University of Maryland
College Park, MD 20742

10. Work Unit No. (TRAIS)

11. Contract or Grant No.

DTRT07-G-0003

12. Sponsoring Agency Name and Address
US Department of Transportation
Research & Innovative Technology Admin
UTC Program, RDT-30
1200 New Jersey Ave., SE
Washington, DC 20590

13. Type of Report and Period Covered

Final

14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

After the Intermodal Surface Transportation Efficiency Act was established in 1991, an increasing number of state highway agencies and federal agencies have started to develop and implement statewide or national travel demand models to meet policy and legislative development needs, and to predict the future travel demand. Up to date, more than 35 states have conducted modeling developments at statewide level (Cohen, Horowitz, & Pendyala, 2008; Giaimo & Schiffer, 2005; Horowitz, 2006, 2008; Souleyrette, Hans, & Pathak, 1996). However, a lack of up-to-date multimodal and inter-regional travel survey data hinders researchers' or analysts' ability to quantitatively conduct reliable and effective evaluation of long-distance travel infrastructure investment and management at statewide level. Meanwhile, in Europe travel demand modeling at national level has received more attention in the last two decades. From the perspective of geography and population size, the European national travel demand model, to an extent, can be taken to be a statewide model in the U.S. Among the efforts making on long-distance passenger travel modeling, the travel data collection is found to play a critical role in the success of the travel demand modeling at both statewide and national levels.

17. Key Words

Travel demand analysis, modeling methods

18. Distribution Statement

No restrictions. This document is available from the National Technical Information Service, Springfield, VA 22161

19. Security Classif. (of this report)

Unclassified

20. Security Classif. (of this page)

Unclassified

21. No. of Pages

22. Price

1. INTRODUCTION

After the Intermodal Surface Transportation Efficiency Act was established in 1991, an increasing number of state highway agencies and federal agencies have started to develop and implement statewide or national travel demand models to meet policy and legislative development needs, and to predict the future travel demand. Up to date, more than 35 states have conducted modeling developments at statewide level (Cohen, Horowitz, & Pendyala, 2008; Giaimo & Schiffer, 2005; Horowitz, 2006, 2008; Souleyrette, Hans, & Pathak, 1996). However, a lack of up-to-date multimodal and inter-regional travel survey data hinders researchers' or analysts' ability to quantitatively conduct reliable and effective evaluation of long-distance travel infrastructure investment and management at statewide level. Meanwhile, in Europe travel demand modeling at national level has received more attention in the last two decades. From the perspective of geography and population size, the European national travel demand model, to an extent, can be taken to be a statewide model in the U.S. Among the efforts making on long-distance passenger travel modeling, the travel data collection is found to play a critical role in the success of the travel demand modeling at both statewide and national levels.

The most recent sources of long-distance passenger flow data in the U.S are the 1995 American Travel Survey (ATS) conducted by the Bureau of Transportation Statistics (BTS) and 2001/2009 National Household Travel Survey.

The 1995 American Travel Survey obtained detailed long-distance travel (>100 miles) information from more than 80,000 households. The long-distance travel information was needed "to identify characteristics of current use of the nation's transportation system, forecast future demand, analyze alternatives for investment in and development of the system, and assess the effects of Federal legislation and Federal and state regulations on the transportation system and its use" (BTS 1995). The 1995 ATS long-distance travel data has successfully supported a few demand modeling studies, including the development of long-distance passenger travel modules in many statewide travel models and the development of a four-step national travel demand model (conducted by Virginia Tech and funded by NASA). However, the ATS data are 15 years old, and have limited sample size, which is inadequate and needs to be updated for long-distance passenger travel analysis in the U.S.

The NPTS/NHTS collected data on all trips taken during a designated travel day, regardless of the trip length. However, recognizing the rarity of long distance trips and the difficulty in capturing long distance travel during a single day data collection window, the NPTS survey was modified in 1990 to expand the data collection window from a single day to a 2-week period so that trips longer than 75 miles or more one way can be captured. In 2001, the NPTS was combined with the ATS, which focused specifically on long distance travel behavior. The combined survey created the nation's inventory of daily and long-distance travel, and was re-named to the National Household Travel Survey (NHTS). The 2001 NHTS redefined the collection of long distance trips to round-trips taken during a four-week period where the farthest point of the trip was at least 50 miles from home. The 2001 survey, which sampled approximately 66,000 households, produced only 45,165 trips longer than 50 miles. By 2008, the NHTS returned to the single day data collection scheme where data on trips taken by all members of a household during a designated travel day were collected. The 2008 survey sampled 155,000 households, consisting of a nationwide sample of 30,000 households supplemented by regional add-ons totaling 125,000 additional households, with interviews conducted from April 2008 through May 2009. However, the survey's limited sample size means that the use of data expansion factors to disaggregate data below large geographic regions of the country (New England, Mid-Atlantic, etc.) soon runs into statistical problems, preventing the direct creation of detailed long-distance passenger OD flow tables.

The traditional long distance travel survey at household level can collect most of the information required for travel analysis and modeling. However, it places large burden on the respondents with relatively high cost, and the travel data reporting and measurement errors would decrease the data reliability. In addition, the low frequency of the long-distance travel for most of the households makes it difficult and costly to acquire a sufficiently large sample of long-distance travel. Consequently, advanced travel survey methods using GPS technology, smartphone, social media and etc., which can overcome the weakness of the traditional survey, can provide the temporal-spatial information of travel more accurately. It attracts travel researchers and analysts to explore and test the feasibility of long-distance travel survey based on the advanced technologies. However, the travel survey methods based on GPS, smartphone, and etc. cannot provide all the long-distance trip information such as travel mode, trip purpose, travel time, and etc. Therefore, while such travel survey methods are explored and tested, the practical post-processing methods which can generate the missing travel characteristics (e.g. trip purpose, travel mode, and etc.) are needed to supplement the data directly from the GPS/smartphone/social media-based survey.

In this report, the post-processing methods (machine learning methods) to automate the trip purpose estimation is developed for long-distance travel, and available datasets including travel survey data and other supplementary data are employed to test and validate the method. This research aims to provide the support tool for long-distance travel data collection and sound methodology for post-processing the GPS-, smartphone-, and social media-based travel survey data in future. Alternative trip purpose categorization schemes for long-distance travel have been developed. Furthermore, the model performance under different purpose categorization is tested in order to provide comprehensive information to assist the design of future long-distance travel surveys.

2. LITERATURE REVIEW

More and More travel researchers are exploring GPS-based travel survey methods, and different methods have been developed to derive the trip purpose as accurately and effectively as possible mainly in the area of regular intra-regional travel. Wolf et al. (2001) pioneered the procedures of trip purpose detection based on a set of deterministic rules and a sample of 19 respondents who both successfully collected travel data with the GPS data logger and returned a completed a paper trip diary and demonstrated the possibility of detecting trip purposes in Atlanta, Georgia, given a detailed GIS database of land use. They found that mixed-use land use parcels such as shopping center, office building and strip mall posed a major challenge on accurate trip purposes detection. In addition to the GIS land use data, respondent's socio-economic characteristics such as household composition, possession of travel modes, and home and work addresses can be helpful to derive trip purpose as well. Subsequent work by Schönfelder et al. (2003) in Europe further developed the procedures. They used multi-stage hierarchical matching procedure, calculating a cluster center of stop ends by combining trip ends, identifying trips with obvious purposes, and establishing relationships between trip purposes and activity temporal information as well as the socio-demographics of the respondents. Stopher et al. (2008) presented a set of heuristic rules to derive trip purpose of 43 trips collected in Sydney with the help of not only the parcel-level land use data but also the geo-coded addresses of the respondent's workplace or school, and the two most frequently used grocery stores. Bohte et al (2008) developed a GPS-based travel data collection method combining GPS devices, GIS technology and a web-based validation procedure, and derived the trip purposes based on the heuristic rules. Chen (2010) followed Schönfelder's approach to cluster trip ends into activity locations. Supplemented by the GIS data and respondent's socio-demographic characteristic, deterministic rules were used to classify trip purposes for trips occurred in low-density area. For those trips in high-density area trip purposes cannot be deterministically decided, and the Multinomial Logit model is employed to calculate the probability that a trip served a particular purpose, with only four trip purposes considered.

The method of deriving trip purpose based on GPS/GIS-based data was further explored with artificial intelligence or machine learning. Griffin et al. (2008) constructed a decision tree to derive trip purposes, and the procedure was implemented in the C4.5 environment with 50 randomly generated trips which are simulated following a series of assumptions. Different from Griffin's method, Deng et al. (2010) employed a number of attributes (not only the attributes from the GPS data, such as time stamp, spatial-temporal indices of trips and attributes from GIS data, but also the social-demographic and socio-economic characteristics of respondents) to construct a decision tree to derive the travel models and trip purposes. The decision tree is implemented in the C5.0 machine learning environment with a homogenous set of 226 GPS trip records collected from 36 respondents in Shanghai. A detailed description of previous researches in trip purpose detection for regular intra-regional travel based on GPS-based travel survey data is presented in Table 1.

In 2001 NHTS trips of 50 miles or more from home to the farthest destination traveled are defined as long distance travel. The definition has changed from the one used in 1995 American Travel Survey (ATS) which defined long distance travel as trips of 100 miles or more and commuting trips were excluded. A long-distance trip includes the part of the trip to the final destination, the return trip home and any overnight stops made along the way or stops to change the travel modes. Similar to regular daily trips, long-distance travel includes trips by all modes such as private vehicle, airplane, bus, train, and ship, and long distance travel includes all purposes, such as commuting, business, pleasure, and personal or family business. However, compared to daily intra-regional trips, the trip purposes of long distance are more focused on business, pleasure, and visiting. Table 2 indicates the different trip-purpose categories of long distance travel from various sources or studies. Until now, little research has been done to identify trip purposes for long distance travel, while it's feasible to estimate the trip purposes provided the trip start/end destination from long distance travel survey, land use information and other sources.

3. METHODOLOGY

The trip purpose detection system is illustrated in Figure 1 in the Appendix. It consists of four parts including input, learning process, output, and validation. Model inputs include travelers' geospatial location data which are reconstructed based on GPS inputs for the derivation of trip characteristic information, travel recall surveys that provide the individuals' social-demographic and economic attributes, and GIS land use data. The learning process module employs machine learning methods and implements trip purpose detection algorithms. After trip purposes are derived based on the machine learning methods, the validation module will evaluate the classifier performance and the reliability of the results.

In the learning process, multiple machine learning methods (e.g. decision tree learning, Meta-learning, Support Vector Machine) have been employed and tested for trip purpose imputation. The purpose is to find the classifier with the best performance. Furthermore, alternative trip purpose categorization schemes for long-distance travel have been developed and tested step by step from binary-class to multi-class (Table 3).

3.1 Decision Tree Learning

Machine learning approach is employed to automate the trip purpose detection procedure. It takes a series of inputs to construct a decision tree classifying trip purposes. The input attributes include individual's trip characteristics derived from the GPS data such as trip start/end time, trip destination location, and activity duration, GIS-based land use type, as well as individual's social-demographic attributes.

The widely used decision tree algorithm in practice is C4.5 introduced by J. Ross Quinlan in 1993, an extension of his earlier ID3 algorithm. C4.5 algorithm employs the information gain to split each node, choosing the attribute at each node that produces the purest daughter node to split on. The information is a measurement of purity. The daughter nodes in the sub-tree will be split based on the same procedure, until all the instances at a node reach the same classification. Given a training data set S and attribute set A ($a_1,a_2,\dots a_n$), different attributes could create different branches and partition the data set S into different subdivisions(V_1, $V_2,\dots V_n$). The number of leaf nodes (L) in subdivision V_i varies by the split attribute. The information gain of each attribute in the attribute set A will be calculated and the attribute with the largest information gain will be chosen to split on. The information gain is represented in Formula 1.

$$\text{Gain}(S, a_i)=\text{Info}(S) -\text{Average } [\text{Info}(L_1,V_i), \text{Info}(L_2,V_i), \dots, \text{Info}(L_n,Vi)] \qquad (1)$$

Gain(S, a_i) represents the information gain of the attribute a_i in the data set S. Info(S) refers to the information value of the data set S. (L_i, V_i) represents the leaf node L_i in subdivision V_i, Info(L_i, V_i) is the information value of leaf node L_i in subdivision V_i resulting from the data split on attribute a_i. The term Average [Info(L_1,V_i), Info(L_2,V_i), …, Info(L_n,Vi)] on the right hand side in the formula is a weighted average linked to the number of instances at each leaf node. It represents the amount of information expected to be necessary to determine the class of a new instance, given the tree structure. The information gain of each attribute in attribute set A based on data set S can be generated, and the attribute with the largest information gain will be selected to be split on.

Under this basic framework, each attribute in set A would be split recursively so that the information gain can reach the maximum value at each node of the tree, until all the instances at each leaf node will have only one classification.

Decision Tree Pruning

Pruning a decision tree is a technique that reduces the size of the tree by cutting off some nodes from the tree which have litter power in instances classification. Employing pruning in decision tree model could improve the computational efficiency and accuracy, reduce the complexity of the tree and avoid the problem of the data set over-fitting. The pruning methods applied to the trip purpose decision tree in the research are post-pruning and on-line pruning. Post-pruning, a bottom up pruning strategy, is executed based on a built decision tree. The relative frequencies of leaf nodes are calculated and compared, and any leaf node with dominant classification will result in a pruning of the parent node. Afterwards, error estimates of the replacement node and the old parent node would be compared to evaluate whether the pruning is advantageous. On-line pruning is different from the post-pruning in the time of pruning, and the former one implements pruning while the decision tree is being built. When a split is made on a certain node which we discussed in the Trip Purpose Estimation part, several children leaf nodes will be generated. Once a child leaf node owns less than a minimum number of instances, the parent node and its children leaf nodes will be compressed into a single node. The pruning process continues until the completion of the entire tree.

Validation

The method to estimate the error rate of machine learning technique is the 10-fold cross-validation. The full sample size is randomly divided into 10 parts each one of which has the same proportion of classes as that in the full data set. Each part is held alternately and the remaining nine parts are trained by the learning algorithm, then the error rate of the held one part can be calculated. The learning procedure is repeated 10 times with different training sets. At last, an overall error rate can be acquired by averaging the 10 error rates.

3.2 Meta-Learning

Meta-learning is a learning process itself, and it’s learning from the learned knowledge. It means learning from the classifiers produced by the inducers and from the classifications of these classifiers on training data. The main idea of meta-learning is to execute a number of base learning processes on a number of data subsets, and to integrate the knowledge of the separately learned classifiers through an extra level of learning to boost the overall predictive accuracy. Ensemble methods, one type of meta-learning algorithms, are typically employed for classification. They combine the results of multiple base classifiers. Bagging or Bootstrap Aggregating, one of ensemble methods, is also emphasized and used in this research. It builds data subsets by bootstrap sampling, trains the multiple classifiers based on these data subsets, and predicts (tests) by majority vote for classification and by averaging for regression. Bagging method works when larger variance exists in the training data set and the base classifier is overfitted. Bagging can decrease the variance without changing the bias. However, if the base classifier is under-fitting, bagging will not help much.

4. DATA

Travel Survey Data

The travel survey data employed to help derive the trip purpose imputation is the 1995 American Travel Survey. The emphasis of the more recent 1995 American Travel Survey (1995) is to gather both cross-section and longitudinal information. Key cross sectional estimates include the origins and destinations of trips, the proportions of people traveling on various transportation modes, intermodal connections, reasons for trips, trip duration, trip distance, and person and household characteristics that may influence aggregate travel demand during a particular time period. Longitudinal estimates require the collection of information about travel behavior or the members of households and persons over the entire survey year. A probability-based sample of households from each of the 50 states and the District of Columbia with more than 80,000 total households were contacted between April 1995 and March 1996. The sampled households were interviewed four times during this period, at approximately three month intervals. In addition to data on a household's members and their individual characteristics, detailed information about each trip taken by each member of the household was collected quarterly. The main trip characteristics collected included the purpose of trip, means of transportation, origin, destination, intermediate stops, travel dates, trip duration, number of nights away, side trips originating at the final destination, and types of lodging used at intermediate stops, final destination, and side stops. Travel distances for each trip were assigned based on transportation network routing algorithms. Most interviews were conducted by telephone, with respondents mailed a travel dairy, a map, and instructions on why and how the survey was being performed. Using census developed household expansion factors origin-destination (OD) trip matrices were developed at the State-to-State and Metropolitan Area-to-Area and Area-to-State levels, for the nation's 55 largest metropolitan areas. The sample size limited further spatial breakouts, and also limited most of these O-D flows to major OD travel pairs.

The 1995 ATS survey collected the long-distance travel information of the household members, and it is composed of four data subsets including household trip, household characteristics, personal trip and personal characteristic. Due to the specific objective of the research, the personal trip, personal characteristic and household characteristic data are adopted. The personal trip data in the 1995 ATS includes 556026 trip records, which include both domestic and abroad long-distance trips. All of the trips are employed to help derive the trip purpose. In addition to the primary long-distance trip characteristics, additional information including stops to the destination, stops from the destination, and the side trips at the destination are provided. These include the stop location at metropolitan area level and state level, travel mode used to the stop, reason for the stop, number of nights at the stop, and the lodging type at the stop.

Supplementary Data

Sources of land use data include land use type and intensity at state, zone, parcel, block and even building levels from local, metropolitan and state planning agencies, and graphic/digital land use information and other geospatial information. Owning to the long-distance travel's specific feature of wide coverage, national coverage of land use data is required. Since the 1995 ATS data doesn't contain any geo-coded address information of the trip, land use data at more aggregate level are adequate and suitable under the premise of providing the destination state or metropolitan area. The report currently employs the NOAA Coastal Assessment and Data Synthesis System as the land use data source at the national level. It provides the area and the corresponding percentage of different land use types by state. Total 39 land use types which can be combined into 10 land use classes are provided. According to the particular objective of the research, the 10 land use classes would be over utilized and are further aggregated into 3 land use covers including urban, agriculture, and nature. In order to better derive the long-distance trip purposes, supplementary data such as travel and tourism statistics data as well as Gross State Product (GSP) data are collected and employed. It's hypothesized that people who go to the state with higher travel and tourism population are more likely to travel for pleasure and visiting. Similarly, states with higher GSP tend to have more enterprises and easier accessibility which results in the higher possibility of attracting business trips. The travel and tourism statistic data are collected from U.S Census Bureau. It provides the yearly recreation visits in national parks and state parks by state. Meanwhile, the Gross State Product data in 1995 are obtained for each state from Bureau of Economic Analysis.

To derive trip purposes for long distance travel, various model input variables are proposed in four categories: trip-related variables, respondent characteristics, land use attributes, and other supplementary data. Detailed information about the model variables can be seen in Table 4.

5. RESULTS ANALYSIS

Based on the 1995 ATS data, models from binary class to multi-class are developed to estimate the trip purpose and provide methodological sound support to assist the design of GPS-, social media-, and smartphone-based long-distance travel survey. Multiple classifiers are tested in each step, and results from the best one with the highest classification accuracy will be presented. Firstly, a binary classification model is developed with two long-distance trip purposes: business and non-business. The results of the classifier with the highest classification accuracy are shown in Table 5. It indicates that the encouraging classification results can be obtained for non-business trips at 96.1% accuracy level, and for business trips with 70.1% accuracy. Meanwhile, the non-business trips are over-predicted (Figure 2) with more business trips wrongly classified into non-business trips. Overall, Model 1 successfully estimated trip purposes for 90.31% for all long-distance trips in the 1995 ATS.

Trip purpose imputation models with more than two purposes have also been tested. It should be noted that in the majority of long-distance travel models, only three trip purposes are defined usually along the lines of business, pleasure (leisure/vacation), and other personal purposes. While the combined business/pleasure trip maintains to be treated as business trips in model 2, the non-business trip is split into pleasure and personal business trips. The best classifier results for these 3 trip purposes are presented in Table 6. Results present an overall predictive accuracy of 81.87%, with pleasure trips acquiring the highest performance of 91.5% and personal business trips obtaining the lowest accuracy of 51.7%. Almost half of the personal business trips are wrongly classified as pleasure trips, leading to the under-prediction of personal business trips

(Figure 3). Moreover, as the number of trip purpose categorization increases from binary to three-class, the decision tree grows larger. Another 3-trip-purpose scheme we tested includes business, non-business and combined business and pleasure (B/P) trips (Model 4 in Table 3). The classification results are shown in Table 8. The overall performance can reach up to 90.22%. Among all the trips, the combined B/P trips have the weakest predictive power with only 30.50% of accuracy and almost 60% of the trips are classified as pleasure trips resulting in the under-predication of combined B/P trips (Figure 4). As the estimation procedure goes forward (Table 3), four trip purposes are tested to examine the impact of more than three trip purposes designing in the future advanced long-distance travel survey on trip purpose classification. The decoding structure for business and personal business trips remains the same as that in model 2, while the pleasure trips are further split into leisure and social visiting trips (Model 3 in Table 3). Table 7 shows the results of the four-trip-purpose imputation. The overall accuracy decreased to 76.98% from 81.87% for model 2. Compared to only one pleasure trip category, the separate categorization of pleasure trips in terms of leisure and social visiting trips would deteriorate the predicative accuracy of pleasure trips. Meanwhile, the personal business trips still remain the lowest classification performance.

More models (Model 5 and 6) are developed and evaluated to see whether the trip purpose scheme with combined B/P trips treated as non-business or pleasure trips can improve the classification performance. The binary classification for recoded non-business and business trips is developed and estimated (Model 5). The results of the model are represented in Table 9. An overall accuracy of 91.86% is achieved, which is 1.5% higher than the classification accuracy of model 1. Furthermore, the reconstructed business, pleasure, and personal business trips are utilized to learn the three trip purpose classification (Model 6). The results (Table 10) indicate that the predicative accuracy (82.82%) is increased to a small extent, when the combined B/P trips are coded as pleasure trips. Due to the uncertainty of the pleasure part in the combined B/P trip, it’s risky to define the combined B/P trip as either leisure trip or social visiting trip for four-trip-purpose. Therefore, we stopped at three-trip-purpose classification.

Travel survey using advanced technologies such as GPS and smartphone cannot record the travel party information, unless the survey is designed to be an interactive GPS or smartphone survey. In order to provide comprehensive information for future travel survey design and evaluate the effect of trip party information on trip purpose derivation, another binary classification for non-business and business trips are re-estimated without any trip party attributes. The binary classifier is learned based on model 5 which decoded the combined business and pleasure trips as non-business trips. The results are represented in Table 11. The overall classification performance (88.98%) is reduced by almost 3% and the decision tree grows larger, compared to the binary classifier with trip party information (model 5).

6. CONCLUSION

This research demonstrates and evaluates the feasibility of automating the trip purposes estimation for long-distance travel. Machine learning methods and algorithms are employed and tested to find out the best one for the trip purposes classification. In addition, alternative trip purpose categorization scheme is generated in order to provide comprehensive and reliable assistance for future advanced long-distance travel survey design which will utilize the emerging technologies such as GPS, smartphone, social-media, and Bluetooth. Multiple classifiers are learned for each trip purpose categorization using all the trip records in the 1995 ATS dataset, and the best one with the strongest predicative power is analyzed.

The estimation results show that in general, as the number of categories increases, the performance of trip purpose imputation tends to deteriorate, and the decision tree is inclined to be more complex. According to the classification results, it's found out that non-business trips or pleasure trips can achieve satisfactory results, with higher classification accuracy than business trips. Moreover, based on the results comparison of different trip purpose categorizations, it's more appropriate to decode the reported combined business and pleasure trips to non-business trips for binary classification and to pleasure trips for three-class classification. Unsatisfactory results can be seen for business trips and personal business trips, which could be explained by the reported errors which are inevitable in the traditional travel survey and some similar characteristics possessed by personal business trip and pleasure trip such as travel party, travel mode, lodge type of destination, duration and etc. More information about respondents' travel at the destination at urban level and detailed land use data would be helpful to distinguish business trip and personal business trip from other trips based on high-quality travel survey data. One more model without the travel party information is developed and tested to examine the role of such information in long-distance trip purpose imputation and to assist the long-distance travel survey design in future. As expected, the predictive accuracy of the model without the trip party attributes will decrease, however, to a small degree, by almost 3%. Consequently, it can be concluded that the trip party information could affect the long-distance trip purpose estimation, but not significantly.

REFERENCE

Axhausen, KW., Schonfelder, S., Wolf, J., Oliveira, M., Samaga, U., 2003. 80 weeks of GPS-traces: approaches to enriching the trip information. Transportation Research Record, 1870, 46 - 54

Bohte, W., Maat, K., 2009. Deriving and validating trip destinations and modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transportation Research Part C 17, 285–297

Chen, C., Gong, H., b, Lawson, C., Bialostozky, E., 2010. Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study. Transportation Research Part A 44, 830–840

Cohen, H., Horowitz, A., & Pendyala, R. 2008. Forecasting statewide freight toolkit. Washington, DC: Transportation Research Board.

Deng, Z., Ji, M., Deriving Rules for Trip Purpose Identification from GPS Travel Survey Data and Land Use Data: A Machine Learning Approach. 2010. Traffic and Transportation Studies. p.768-777

Doherty, S., N. Noël, M. Lee-Gosselin, C. Sirois, M. Ueno, and F. Theberge. 2001 . Moving beyond observed outcomes: Integrating Global Positioning Systems and interactive computer-based travel behavior surveys. Transportation Research E-Circular, C 26, 449-466.

Du, J. and L. Aultman-Hall. 2007 . Increasing the accuracy of trip rate information from passive multi-day GPS travel datasets: Automatic trip end identification issues. Transportation Research Part A 41 (3), 220-232.

Giaimo, G.T.,& Schiffer, R. (Eds.). 2005, August. Statewide travel demand modeling: A peer exchange. Transportation Research Circular, #E-C075.

Gonzalez, A.P., Weinstein, S.J., Barbeau, J.S, Labrador, A.M., Winters, L.P., Georggi, L.N., Perez, R., Automating mode detection using neural networks and assisted GPS data collected using GPS-enabled mobile phones. 2008. 15thWorld Congress on Intelligent Transportation Systems, New York, NY. Paper # 30267

Griffin, T., Huang, Y., 2005. A Decision Tree Based Classification Model to Automate Trip Purpose Derivation. In the Proceedings of the 18th International Conference on Computer Applications in Industry and Engineering, Honolulu, Hawaii

Horowitz, A.J. 2006. Statewide travel forecasting models (NCHRP Synthesis #358). Transportation Research Board.

Horowitz, A.J. 2008. White paper: Statewide travel demand forecasting. Requested by AASHTO and presented at the conference on meeting federal surface transportation requirements in statewide and metropolitan transportation planning.

Rates, E. 2007 . Atlanta Commute Vehicle Soak and Start Distributions and Engine Startsper Day Impact on Mobile Source. Atlanta.

Schönfelder, S., K. Axhausen, N. Antille, M. Bierlaire, and E. Lausanne. 2002 . Exploring the potentials of automatically collected GPS data for travel behavior analysis - a Swedish data source. GI-Technologien für Verkehr und Logistik 13, 155-179.

Schuessler, N., Axhausen, K.W., 2009. Processing raw data from Global Positioning Systems without additional information. Transportation Research Record 2105,28–36.

Souleyrette, R.R., Hans, Z.N., & Pathak, S. 1996. Statewide transportation planning model and methodology development program. Ames: Iowa State University.

Stopher, P.R., Greaves, S., FitzGerald,C.,2005. Developing and deploying a new wearable GPS device for transport applications. Paper presented to the 28th Australasian Transport Research Forum, Sydney, 28–30 September.

Stopher, P., Clifford, E., Zhang, J., FitzGerald, C., 2008a. Deducing Mode and Purpose from GPS data. Working Paper of the Austrian Key Centre in Transport and Logistics. University of Sydney, Sydney, Australia.

Stopher, P., FitzGerald, C., Zhang, J., 2008b. Search for a Global Positioning System device to measure personal travel. Transportation Research Part C 16(3), 350–369.

Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary: an experiment to derive trip purpose from GPS travel data. In 80th Annual Meeting of the Transportation Research Board, Washington DC., p.24.

Wolf, J., R. Guensler, and W. Bachman (2001). Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. Transportation Research Record: Journal of the Transportation Research Board 1768 (1), 125-134.

Witten, I. H., Frank, E., 2005, Data Mining: Practical Machine Learning Tools and Techniques, Second Edition

Zhu, S., D. Levinson, and H. Liu. 2010. Measuring Winners and Losers from the new I-35W Mississippi River Bridge. In The 89th Annual Conference of Transportation Research Board, Washington D.C.

LIST OF TABLES AND FIGURES

Table 1 Previous Study on Trip Purpose Imputation

Author	Trip Purpose	Variables	Methodology	Data Source	Land Use Type	Validation	Model Performance
Jean Wolf et al. (2001)	1. go to work 2. go to school 3. personal business 4. return home 5. shop 6. social / recreation 7. eat 8. drop off / pick up	1.Trip destination coordination 2.Arrival Time 3.Activity Duration 4.Land Use Type Code	A set of deterministic rules	19 participants both collected in-vehicle GPS data and completed paper travel diary, 151 trips were detected trip purpose	A derived Land Use database by property polygon, center point of polygon and street Address;(25 categories of land use type)	Reported trip purposes from the participants' travel diary were compared with derived trip purposes	10 trips(7%) purposes were incorrectly derived due to inaccurate land use assignment
Schonfelder et al. (2003)	1.Pick up / Drop off 2. Private business 3. Work related business 4. School 5. Work 6. Daily shopping 7. Long-term shopping 8. Leisure 9.Home 10. Other	1.Location of parked vehicle; 2.Activity Duration; 3. Time of day; 4.Day of week; 5.Frequency of visits 6.Socio-economic variables	Multi-Stage Heuristic rules (Each POI and Land Use are given a certain probability of trip purpose)	186 private vehicles with minimum socio-economic data for at least 30 days in Swedish, while only 39 vehicles were selected to impute the trip purpose;	1.Home location; 2.POI, cluster center's buffer of 300m 3.Polygon Land Use type, cluster center's buffer of 200m	Trip purpose shares derived from GPS data are compared with those of the 2000/2001 Swedish national travel survey data	The shares of the trip purposes principally show the same pattern as the (weighted) Swedish reference data, except the Private Business, Work Related Business, and Daily Shopping.

Stopher et al. (2008)	1.Home-Based Work 2.Home-Based education 3.Home-Based Shopping 4.Home-based eat meal 5.Home-Based personal/medical 6.Home-based social/recreational 7.Home-based pick-up/drop-off 8.Home-based other 9.Non-home-based work-other 10.Non-home-based other-other	1.Activity Duration 2.Geocode addresses listed in Land Use Type 3.Frequency of visits over each week 4.Trip destination coordination	A set of Heuristic rules (Trip ends within buffer of 200m are considered having the same location)	Two projects: 1).56 days of wearable GPS data from 21 respondents to Sydney Household Travel Survey 2). 245 days of wearable GPS data for an evaluation of a pilot Travel Behavior Change Program	1.Home Address; 2.Address of each workplace for each working household member; Address of each 3. Educational establishment for household members engaged in education; 4.Address of the two most frequently used grocery stores; 5.Parcel-level GIS	Checking with supplementary data of people’s trip purpose	

Bohte, et al. (2008)	1. Work
	2. Study
	3.Shop
	4.Social Visit
	5.Recreation
	6.Home
	7.Other
		1.Activity Duration;
		2.Distance between the end points and home/work geo-coded address;
		3.Distance between the end point and POI;
		4.Whether the end point is within Shop Center polygon
		5. Individual Characteristic variable
			Heuristic Rule (learning process through the feedback of the respondents)	1104 respondents’ completed the entire project with handheld GPS data logger for one-week in the Netherlands	1. Home and Work Address, with a buffer of 100m and 50m separately;
					2.POI data, trip ends’ buffer of 50m;
					3.Polygon Land Use data;
						Respondents are asked to correct/add the trips derived from the GPS data through web-based interface;
						Trip purposes derived from GPS data are compared with the Dutch Travel Survey (one-day paper recall survey)
							Trip purposes share from both data set are similar, and the number of tours per day is almost equal in both dataset. The main difference lies in the number of trips per tour.

Chen, et al (2009)	1. Home-Based Work/school 2.Home-Based Personal Business 3.Home-Based Social Recreation 4.Home-Based Shopping 5.Non-Home-Based Work/school 6.Non-Home-Based Personal Business 7.Non-Home-Based Social Recreation 8.Non-Home-Based Shopping	1.Time of day, 2.History Dependence, 3.Land use characteristics	1.Deterministic matching between O/D and corresponding land use type in low-density area, 2. Multinomial logit model	25 participants carrying personal GPS for one weekday; The other 24 participants carrying personal GPS for five weekdays in New York City	1. Business listings 2.Frequently Visited Locations 3.Polygon land use Land use buffer of 50m, 150 and 250m are used to estimate the MNL, while 250m is most significant	Trip purposes from participant’s everyday travel diary are used to validate the derived trip purpose	67% and 78% prediction rates for home-based trips and non-home-based trips according to MNL
Griffin et al. (2005)		1. Time of Day (Aggregate Values of point ends in cluster) 2.Activity Duration (Aggregate Values) 3. Earliest Arrival Time	1.Dbscan cluster algorithm 2.Decision Tree (C4.5 algorithm)	50 randomly generated trips based on users’ information from questionnaire.		Percentage of correctly classified points within the cluster	An accuracy of well over 90% for specific cluster sizes is very significant
Deng, et al. (2010)	1.go to work 2.go to school 3.go home 4.pick-up/drop off 5.shopping/recreation 6.business visit 7.others	1. Weekdays, weekend days, time of a day 2. Socioeconomic Variables, 3.Trip Distance 4.Activity Duration 5.Speed	Decision Tree learning (C5.0 algorithm)	226 trips from 36 respondents carrying personal GPS in a three-day period in Shanghai	10 Land Use types according to China’s Urban Land-use Classification Scheme and Standards	Derived trip purposes are compared with the reported trip purposes from web-based recall survey	Classification accuracy of 87.6% was achieved

* Empty cell indicates that the exact information cannot be found in their research.

Table 2 Long-Distance Travel Trip Purpose Categorization

Study	**Trip Purposes**
TSAM (Ashiabor, Baik *et al*. (2007-2008)), U.S	Business/Non-Business
Koppelman (1990), U.S	Business/Non-Business
1995 American Travel Survey	Business, pleasure, leisure, personal business
Jin and Horowitz (2008), U.S	Work, return home, personal business, recreation.
Oregon, U.S	Home-based, work-based
Michigan, U.S	HB work/biz, HB soc/rec/vac, HBO, NHB work/biz, NHB
Maryland, U.S	HB work, journey to work, journey at work, school, HB shop, HB other.
Cambridge Systematics (2006), U.S	Business, commute, recreation, other
Volpe Center (2008), U.S	Business/non-business
Bhat (1995), U.S	Paid business
Dutch National Model System (LMS)	Work, HB business, NHB business, shopping, education, other
Great Britain National Transport Model (NTM)	HB work, business, HB education,
Italian Decision Support System (SISD)	Commute, business, education, leisure and tourism, and other.
Norway National Transport Model 4 (NTM 4)	Work, business, social, recreation, services and other.
Swedish National Model System (SAMPERS)	Private, business.
Danish National Transport Model (PETRA)	Home, work, errand, and leisure.
BVWP (Austria)	Work, business, school, shopping, leisure, other.
VALIDATE (Germany), 2005	Home, work, business, shopping, and other.
Switzerland National Travel Demand Model	Home, work, education, business, shopping, and leisure.
MATISSE (France)	Business, private.
STREAMS (EU)	Commuting-business, personal business-education, visiting, domestic holiday, and international holiday.
STEMM (EU)	Business, private, and vacation.
TRANS-TOOLS (EU)	Business/home-work, holiday, and other
Yao and Morikawa (2005), Japan	Business and non-business.

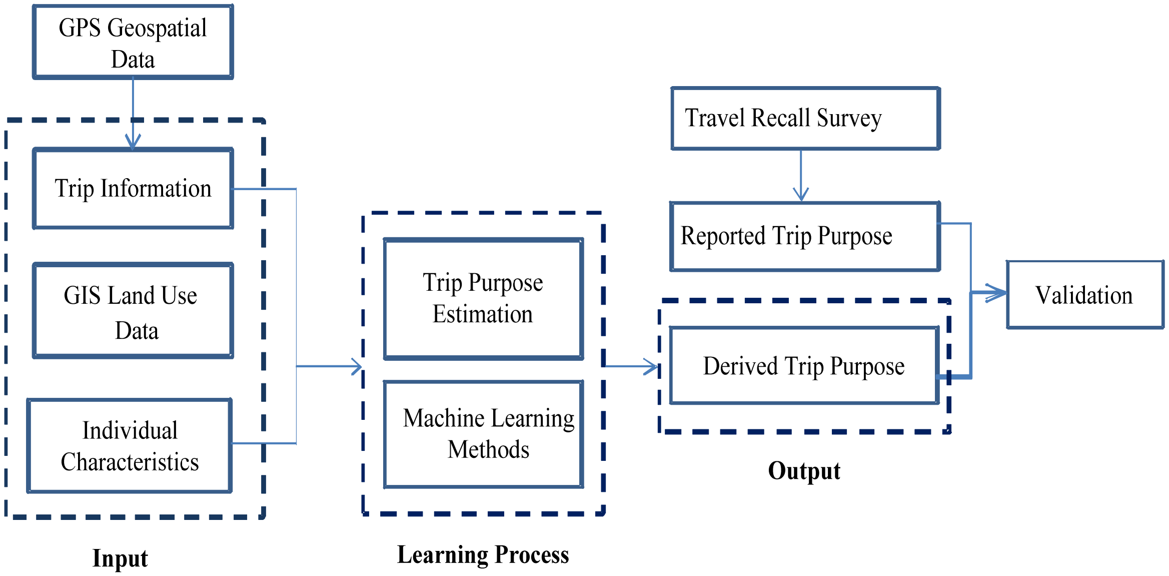

Figure 1 Trip Purpose Learning System for Long-distance Passenger Travel Survey

Table 3 Long-Distance Trip Purposes Categorization in this Project

Reported Trip Purpose	**Decoded Trip Purpose (Model 1)**	**Decoded Trip Purpose (Model 2)**	**Decoded Trip Purpose (Model 3)**	**Decoded Trip Purpose (Model 4)**	**Decoded Trip Purpose (Model 5)**	**Decoded Trip Purpose (Model 6)**
Business	Business	Business	Business	Business	Business	Business
Combined Business/Pleasure	Business	Business	Business	Combined B/P	Non-Business	Pleasure
Convention,Conference, or Seminar	Business	Business	Business	Business	Business	Business
School-related activity	Non-Business	Personal Business	Personal Business	Non-Business	Non-Business	Personal Business
Visit relatives or friends	Non-Business	Pleasure	Social Visit	Non-Business	Non-Business	Pleasure
Rest or relaxation	Non-Business	Pleasure	Leisure	Non-Business	Non-Business	Pleasure
Sightseeing, or to visit a historic or scenic attraction	Non-Business	Pleasure	Leisure	Non-Business	Non-Business	Pleasure
Outdoor recreation	Non-Business	Pleasure	Leisure	Non-Business	Non-Business	Pleasure
Entertainment	Non-Business	Pleasure	Leisure	Non-Business	Non-Business	Pleasure
Shopping	Non-Business	Pleasure	Leisure	Non-Business	Non-Business	Pleasure
Personal, family or medical	Non-Business	Personal Business	Personal Business	Non-Business	Non-Business	Personal Business
Other	Non-Business	Deleted	Deleted	Non-Business	Non-Business	Deleted

Table 4 Proposed Model Variables for Long-Distance Trip Purpose Estimation

Variable Name	**Description**
HHIncome	Household Income
Age	Respondent's Age
Race	Respondent's Race
EducAttainment	Respondent's education level
Activity	Activity of Respondent
TrParty	Travel party size
TrPrHousePercent	Percentage of Adult Household Members in Travel Party
TrPrTyCh	Children Under 18 Years in the Travel Party
Weekend	Whether it's a weekend trip
NiteDest	Number of nights at destination
LodgDest	Lodge type at destination
TransportOriginDest	Principal Transportation from Origin to Destination
InternationalDestFlag	U.S. or International Destination Flag
StopsTo	Number of Stops to Destination
SideTrps	Number of Side trips
Sex	Respondent's gender
Side1state	Side trip 1 destination locates in the same state as the main trip or not
SidetripDest1Lodgn	Lodge type at side trip 1 destination
SidetripDest1Reasn	Trip purpose of side trip 1
SidetripDest1Transportation	Transportation mode to side trip 1 destination
DestRegion	The region where the destination state falls in
Tourism	National Park recreation visits by state
GSP	Gross State Product
Urban	Percentage of urban land use cover by state
Agriculture	Percentage of agriculture land use cover by state
Nature	Percentage of natural land use cover by state

Table 5 Prototype Model (Model 1) Results

Non-business	Business	Actual Purpose	TP Rate
415473	16950	non-business	96.1%
36932	86671	business	70.1%
Overall Accuracy:90.31%			
Number of Leaves: 27473; Size of the tree: 35643			

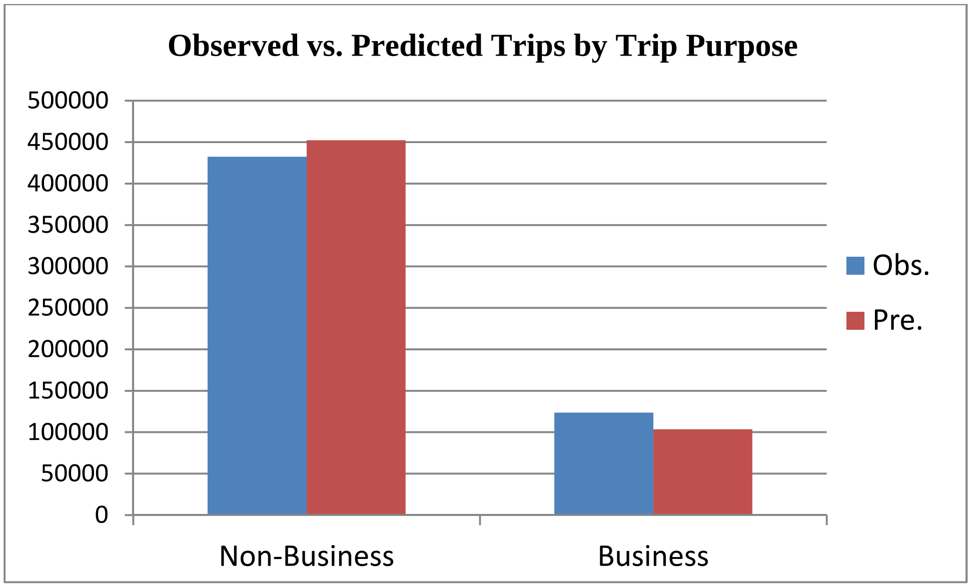

Figure 2 Observed trips vs. Predicted trips by trip purpose for Model 1

Table 6 Model 2 Results

Pleasure	Business	Personal Business	Actual Purpose	TP Rate
315520	17032	12328	Pleasure	91.5%
25656	94419	3528	Business	76.4%
35955	6286	45276	Personal Business	51.7%
Overall Accuracy: 81.87 %				
Number of Leaves: 91241; Size of the tree: 108145				

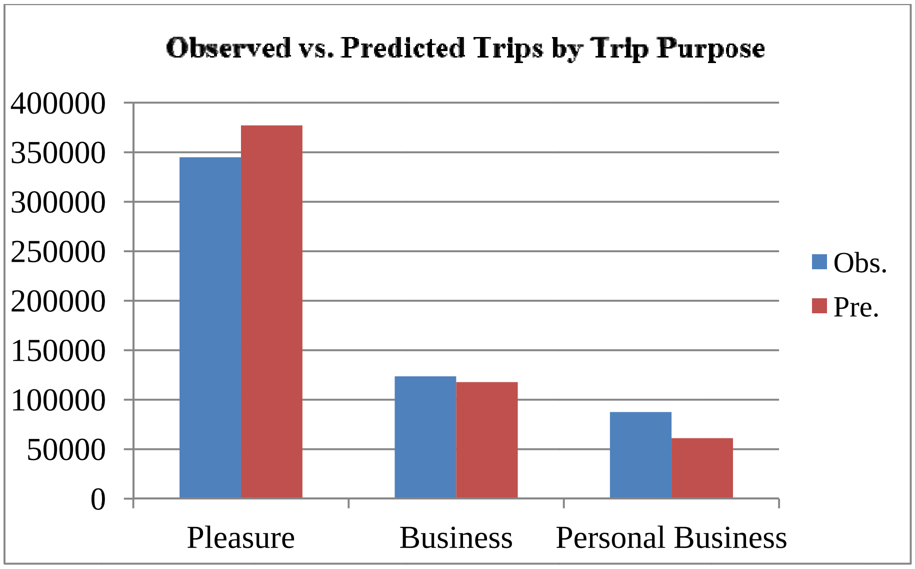

Figure 3 Observed trips vs. Predicted trips by trip purpose for Model 2

Table 7 Model 3 Results

Leisure	Social Visit	Business	Personal Business	Actual Purpose	TP Rate
136656	14359	13395	8665	Leisure	79%
13871	144403	6519	7012	Social Visit	84.1%
14544	7430	97597	4032	Business	79%
14894	16231	7051	49341	Personal Business	56.4%
Overall Accuracy: 76.98 %					
Number of Leaves: 137660; Size of the tree: 163883					

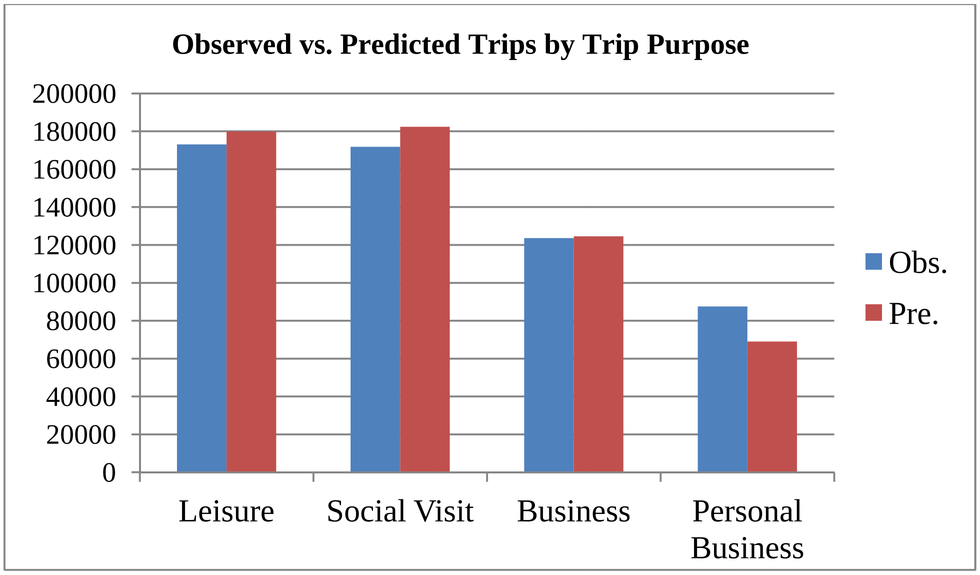

Figure 4 Observed trips vs. Predicted trips by trip purpose for Model 3

Table 8 Model 4 Results

Non-business	Business	Combined B/P	Actual Purpose	TP Rate
417072	14288	1063	Non-business	96.50%
28740	80249	397	Business	73.40%
8557	1330	4330	Combined B/P	30.50%
Overall Accuracy:90.22%				
Number of Leaves: 30342; Size of the tree: 38672				

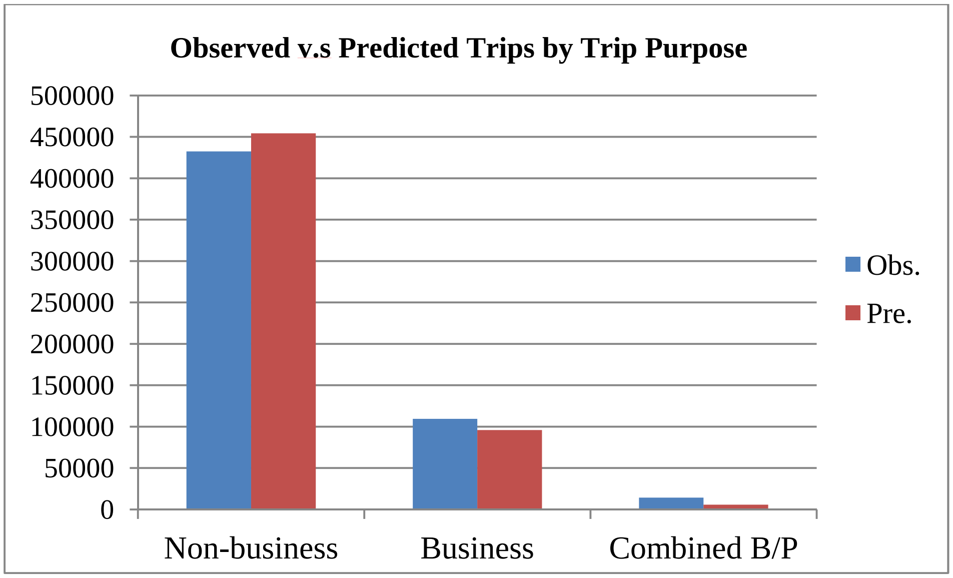

Figure 5 Observed trips vs. Predicted trips by trip purpose for Model 4

Table 9 Model 5 Results

Non-business	Business	Actual Purpose	TP Rate
432079	14561	non-business	96.7%
30693	78693	business	71.9%
Overall Accuracy:91.86%			
Number of Leaves: 24109; Size of the tree: 30120			

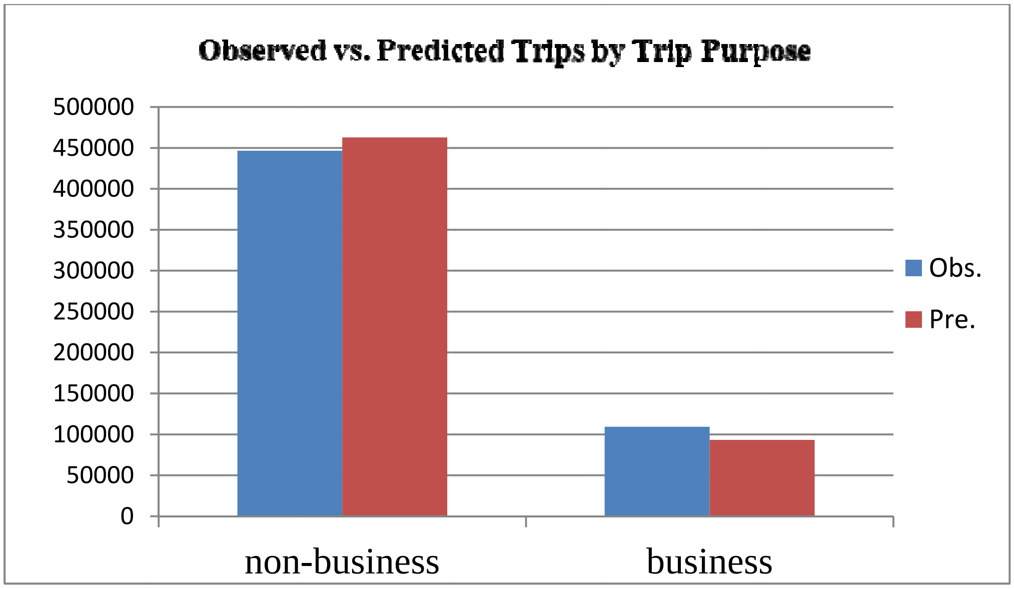

Figure 6 Observed trips vs. Predicted trips by trip purpose for Model 5

Table 10 Model 6 Results

Pleasure	Business	Personal Business	Actual Purpose	TP Rate
332435	15029	11633	Pleasure	92.6%
22158	84496	2732	Business	77.2%
38696	5249	43572	Personal Business	49.8%
Overall Accuracy: 82.82 %				
Number of Leaves: 86176; Size of the tree: 100786				

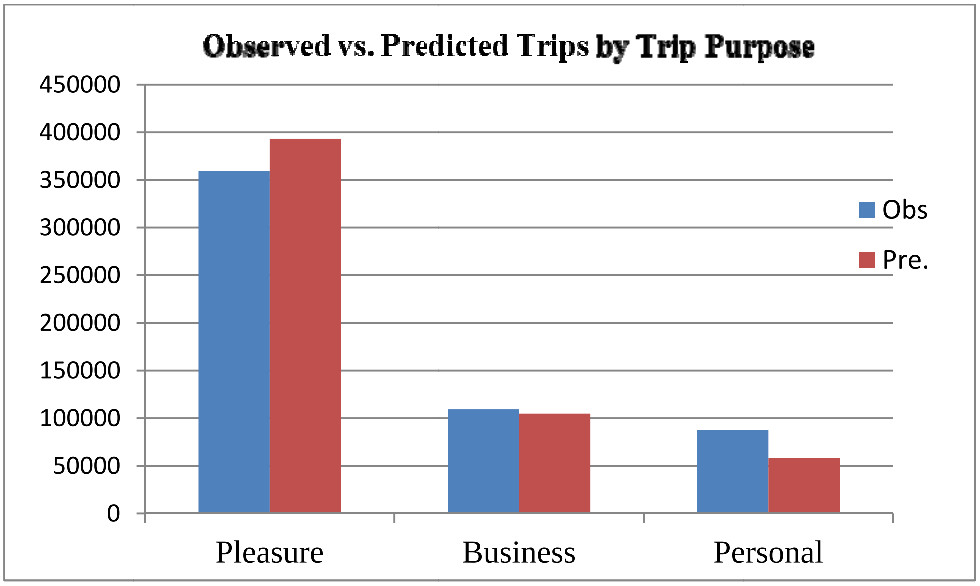

Figure 7 Observed trips vs. Predicted trips by trip purpose for Model 6

Table 11 Compared Model Results

Non-business	Business	Actual Purpose	TP Rate
427853	18787	non-business	95.80%
42498	66888	business	61.10%
Overall Accuracy:88.98%			
Number of Leaves: 24493; Size of the tree: 30792			