

ABSTRACT

Title of Document: SHORT TERM TRAVEL BEHAVIOR PREDICTION
THROUGH GPS AND LAND USE DATA

Cory Krause

Directed by: Professor Lei Zhang
Department of Civil and Environmental Engineering

The short-term destination prediction problem consists of capturing vehicle Global Positioning System (GPS) traces and learning from historic locations and trajectories to predict a vehicle's destination. Drivers have predictable trip destinations that can be estimated through probabilistic modeling of past trips.

This dissertation has three main hypotheses; 1) Employing a tiered Markov model structure will permit a shorter learning period while achieving similar accuracy results, 2) The addition of derived trip purpose information will increase accuracy of the start of trip and in-route models as a whole, and 3) Similar methodologies of travel pattern inference can be used to accurately predict trip purpose and socio-economic factors.

To study these concepts, a database of GPS driving traces (120 participants for 70 days) is collected. To model the user's trip purpose, a new data source was explored: Point of Interest (POI)/land use data. An open source land use/POI dataset is merged with the GPS dataset. The resulting database includes over 20,000 trips with travel characteristics and land use/POI data. From land use/POI data, and travel patterns, trip purpose is calculated with machine learning

methods. A new model structure is developed that uses trip purpose when it is available, yet falls back on traditional spatial temporal Markov models when it is not.

The start of trip model has an overall increase of accuracy over other start of trip models of 2%. This comes quickly, needing only 30 days to reach this level of accuracy compared to nearly a year in many other models. When adding trip purpose and the start of trip model to in-route prediction methods, the accuracy of the destination prediction increases significantly: 15-30% improvement of accuracy over similar models between 0-50% of trip progression. Certain trips are predicted more accurately than others: work and home based trips average of 90% correct prediction, whereas shopping and social based trips hover around the 50% mark. In all, the greatest contribution of this dissertation is the trip purpose methodology addition and the tiered Markov model structure in gaining fast results in both the start of trip and in-route models.

SHORT TERM TRAVEL BEHAVIOR PREDICTION THROUGH GPS AND LAND USE
DATA

By

Cory Krause

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
2015

Advisory Committee:
Professor Lei Zhang, Chair/Advisor
Professor Daniel J. Dailey
Professor Paul Schonfeld
Professor Chris Davis
Professor Ali Haghani

© Copyright by
Cory Krause
2015

Acknowledgements

I would have been unable to complete my Dissertation without the support and guidance of colleagues and friends from the University of Maryland and Federal Highway Administration.

My deepest gratitude extends to my advisor, Dr. Lei Zhang, for his great effort in providing the best possible atmosphere for advanced research. Dr. Paul Schonfeld has been a strong guiding presence throughout my academic career, earning the deep admiration that students from University of Maryland have for him. Dr. Ali Haghani and Dr. Chris Davis, I thank you for your dedication to my committee and continued encouragement throughout the program. Many thanks are extended to Dr. Dan Dailey, not only for his help with Chapter 8, but for many Funny Times over the years.

Additionally, I have been lucky enough to benefit from professional relationships in the engineering community outside of University of Maryland. The opportunity to join FHWA and gain colleague's support has been invaluable. My thanks go out to Dr. Gene McHale, Dr. Govind Vadakpat, Dr. Joe Bared, Dr. Taylor Lochrane, and Dr. Ximiao Jiang. Whether they were providing technical advice, or giving me the opportunity to lead exciting research efforts, they have all opened doors I never thought possible.

I have also benefitted from wonderful friends and fellow students in the PhD program at University of Maryland. Thanks to Nayel, Chenfeng, Yijing, Mercedeh, Sahar, Arefeh, Dilya, and others who I have studied and collaborated with. It is an honor to be in the same class as you all.

Finally, none of this would be possible without the love and support of my family. Many thanks to my mom and dad for making higher education available to me. Of course, I would like to thank my smarter, harder working better half Allison who stood by me through all the night classes, exams, and stressful presentations. If I can do this, I can only imagine what she will accomplish.

Table of Contents

Table of Figures:	viii
Table of Tables:	x
List of Definitions	xii
List of Abbreviations	xiii
List of Variables.....	x
CHAPTER 1: Introduction	1
1.1 Motivation.....	1
1.2 Objective	2
1.3 Contribution of Work.....	3
1.4 Outline:	4
CHAPTER 2: Literature Review	7
2.1 Travel Behavior	7
2.1.1 Four step model.....	7
2.1.2 Representing Travel Behavior	10
2.2 GPS Travel Survey	11
2.3 Trip Purpose Designation	11
2.4 Destination Prediction Lit Review	13
2.5 Markov Modelling	17
2.5.1 Markov Model Explanation	18
2.5.2 Markov Order.....	19
2.5.3 State-Transition Diagram	19
2.5.4 Graphical Representation	21
2.6 Literature Review Conclusions	23
CHAPTER 3: Data.....	25
3.1 GPS Data.....	25
3.1.1 Pilot Survey.....	26

3.1.2 GPS Device.....	26
3.1.3 Postcard and E-mail Campaign.....	27
3.1.4 Website Design	30
3.1.5 Survey Design.....	32
3.1.6 Respondent Sampling	35
3.1.7 GPS Mailing.....	36
3.1.8 Travel Survey.....	37
3.1.9 GPS Data Checking	39
3.2 Open Street Map	42
3.3 Data Processing.....	44
3.3.1. Data Excluded from the Analysis	44
3.3.2 Scripting for GPS.....	44
3.3.3 Calculating Trip Characteristics	45
3.4 Data Conclusions	46
CHAPTER 4: Tiered Time Origin Markov Model.....	47
4.1 Introduction:.....	47
4.2 Model:	47
4.2.1 Data Elements	47
4.3 Formulation.....	51
4.3.1 Variables	51
4.3.2 Probability of location destination by tier.....	52
4.3.3 Mathematical Formulas	52
Time & Origin:	52
Time:	52
Origin:	53
Most Frequently Visited:	53
Most Frequent Day:	54
4.4 Matrix size	54
4.5 Array filling and algorithm example.....	56
4.6 Results:.....	61
Tier 2: Time of Day and Origin	61
Tier 3: Time of Day	61
Tier 4: Origin	61
Tier 5: Most Common Location	61

Location by differing accuracy definitions:	61
True accuracy:.....	62
4.7 Tiered Time-Origin Model Conclusions.....	63
CHAPTER 5: Markov Model with Future Trip Purpose Information.....	64
5.1 Introduction.....	64
5.2 Methodology	64
5.3 Trip Purpose Determination.....	65
5.3.1 Trip Purpose Model Formulation.....	67
5.4 Results:.....	68
Tier 1: Trip Purpose	68
5.4.1 Accuracy as algorithm learns trip behavior	70
5.4.2 Accuracy by day of week.....	72
5.5 Applications	73
5.6 Conclusions:.....	74
CHAPTER 6: Markov Model with Derived Trip Purpose and Training Sets	75
6.1 Introduction.....	75
6.2 Markov Model (Baseline)	75
6.3 Trip Purpose.....	77
6.4 Results.....	83
6.4.1 Prediction Accuracy by Trip Purpose	84
6.4.2 Training sets of differing length	85
6.4.3 Prediction Variation	87
6.4.4 Confusion Matrix	88
6.4.5 Combination 5-15-30 trip learning models.....	92
6.4.6 Comparison with state-of-the-art models.....	93
6.5 Regression Analysis.....	94
6.6 Conclusion	99
CHAPTER 7: In-Route Destination Prediction	101
7.1 Introduction.....	101
7.2 Lit review – In-route prediction.....	104
7.3 Methodology	107
7.3.1 Roadway Link Allocation	107
7.3.2 Destination Prediction.....	109

7.4 Model Formulation	110
7.4.1 Trip Purpose:.....	112
7.4.2 Origin:	115
7.4.3 None:.....	115
7.5 Results.....	117
7.5.1 Raw Results	117
7.5.2 Best Model Creation	120
7.5.3 Decreased Accuracy over time	122
7.5.4 Accuracy by Trip Purpose	123
7.6 Prediction Validation	125
7.7 Conclusions.....	125
CHAPTER 8: Socio-Economic Prediction through GPS Travel Data	127
8.1 Introduction.....	127
8.2 Data.....	127
8.3 Methodology	128
8.3.1 Variables	128
8.4 Linear Regression	132
8.4.1 Survey Data.....	133
8.4.2 Trip Data	134
8.5 Machine Learning	136
8.5.1 Trip data.....	136
8.5.2 Survey Data.....	137
8.6 Conclusions.....	139
CHAPTER 9: Conclusions	141
9.1 Realistic Implementation	141
9.2 Further Applications	142
9.3 Closing	145
Bibliography	195
APPENDIX.....	147
Appendix A: Data processing code (VBA)	147
Raw data to trip format:	147
Create unique OD identifiers	149

Converting Lat/Long to distance:	151
Appendix B, GPS Survey demographic information:.....	153
Surrounding Area.....	153
Pilot Study Participants	155
Appendix C: GPS Forms and information mailed to survey participants:	159
Travel Diary	159
GPS Installation	164
Return Shipping	167
Appendix D: 30 Day trip purpose training model	168
Appendix E: Purpose Graphs.....	185
Appendix F: In-route prediction accuracy by day	189
Appendix G: Raw In-route Model accuracies	191
Appendix H: Income Prediction Model (J48 in WEKA).....	192
Appendix I: Census Metadata File.....	194

Table of Figures:

Figure 1- Travel Behavior Representation Ben-Akiva (2008)	10
Figure 2- Classification of Markov Processes (Ibe[2013]).....	18
Figure 3- State Transition Diagram Example	20
Figure 4- Bayesian Graph of example problem.....	22
Figure 5: Front side of postcard	27
Figure 6: Back side of postcard	28
Figure 7: Email sent to individuals in the project area	30
Figure 8: Front page of the survey website.....	31
Figure 9: Participation page to start the survey	33
Figure 10: Participation form first page.....	34
Figure 11: Participation form second page	35
Figure 12: Example of participation form results	36
Figure 13: Travel Survey First Page	38
Figure 14: Travel Survey Second Page.....	39
Figure 15: An example of data error.....	40
Figure 16: Correct data as recorded by the GPS device.	41
Figure 17- Zoom in of Long Distance trip.....	42
Figure 18- GPS point buffers and the spatial join to important POI data around the University of Maryland.....	43
Figure 19- Graphical Representation of Tiered Time Origin Model.....	49
Figure 20- Model Accuracy by Distance threshold	63
Figure 21- Trip Purpose Estimation based on GPS Data, GIS Land Use and Participant Demographics	65
Figure 22-Model Accuracy by distance threshold & location availability	69
Figure 23- Percent of accurately identified trip destinations by GPS participant.....	70
Figure 24- Percent of correctly estimated trip locations as Survey continues	71
Figure 25- Accuracy by day of the week	72
Figure 26- Graphical Representation of Tiered Trip Purpose Model	76
Figure 27- Trip Purpose Estimation Based on GPS, GIS, and Machine Learning Methods	79
Figure 28- Setting the variable list and predicting variable (purpose) in WEKA for the J48 Algorithm.....	81
Figure 29- Algorithm Accuracy: With and Without Trip Purpose Module.....	83
Figure 30- Prediction Accuracy by Trip Purpose	84
Figure 31-Accuracy by Trip Purpose Learning Period starting at trip 3 increasing to trip 130. ..	86
Figure 32- Model Accuracy with combined learning method	93
Figure 33- Comparison of accuracy by trip purpose for Tiered Time-Origin versus Trip Purpose Model	102
Figure 34- Alvarez-Garcia model participant averaged.....	106
Figure 35: GPS points in roadway layer for link allocation	108
Figure 36: Links matched to GPS points based on shortest path between points.....	109
Figure 37- Estimation Accuracy 25, and 50% trip progression.....	118
Figure 38- Estimation Accuracy, 75, and 90% trip progression.....	118
Figure 39- Trips estimated; 25 and 50%	119
Figure 40- Trips estimated; 75 and 90%	119

Figure 41- Model accuracy comparison with Alvarez-Garcia model.....	121
Figure 42- Accuracy of in-route model 90% trip completion, .85 route match.....	122
Figure 43- Accuracy of in-route model 90% trip completion, .85 route match (trips 1-30)	123
Figure 44- Accuracy by Trip Purpose (in-route model)	124
Figure 45; Scaled ZeroLag Covariances over census tract locations. 100 income variables	129
Figure 46- Scaled-ZeroLag Covariances over Census tract locations for 100 employment variables.	130

Table of Tables:

Table 1: Email campaign tracking report numbers.....	28
Table 2- Input Variables for Trip Purpose Estimation (phase 1).....	66
Table 3- Participant Income Range.....	66
Table 4- Input Variables for Trip Purpose Estimation (Phase 2).....	79
Table 5- Estimated Trip Origin Land Use Distribution	91
Table 6- Regression Statistics- TTOM	94
Table 7- Regressions Statistics- Trip Purpose Model.....	94
Table 8- Regression Model- without Trip Purpose.....	95
Table 9- Regression Model- with Trip Purpose Tier	98
Table 10- Literature Review model data comparison.....	105
Table 11- Accuracy of Trip Purpose (exact values)	124

List of Variables

U = set of all users in the survey

i = Trip number (i signifies the current trip)

P = trip purpose category (Home, Work, Other, Driving, Social/Recreation, Shopping, School/Daycare)

T = set of all time steps (24 total time steps)

V = set of all visited locations

l = location to be identified as visited location

List of Definitions

Location- A geographical position which has 3 consecutive 1 minute interval GPS points whose velocity is equal to zero

Trip- A series of consecutive GPS points which includes no more than 2 consecutive points whose velocity is 0 miles per hour.

Origin- A location which begins a trip string

Destination- A location which ends a trip string

Distance- The Haversine equation calculation of distance from GPS point to GPS point using the radius of the earth as the mean radius (6371km)

Time interval- A one hour period of the day beginning at the 0th minute of each hour, and ending in the 59th minute. There are 24 time intervals per day

Tier- A level of formulation within the hierarchy of destination prediction methodology. Each higher tier has a quantified greater accuracy than those below.

Accuracy- The total number of correct estimations divided by the total number of predictions made

Correct estimation- If the predicted destination of the end of a trip is within the designated distance threshold of the actual end location of the trip, it is designated as a correct estimation. Unless otherwise stated, the default distance threshold is 300 meters.

Array- an ordered arrangement of a travel attribute. Each entry is accompanied with the destination location it is associated with in that instance. Multiple arrays in the research exist: Trip Purpose, User, Time, Origin, Day of Week, Time & Day.

List of Abbreviations

HPY- Hierarchical Pitman-Yor

HPH- Hierarchical Putman-Yor Prior Hourly Model

HPD- Hierarchical Pitman-Yor Prior Daily Model

HPHD- Hierarchical Putman-Yor Hourly Daily

MFV- Most Frequent Visit Model

OMM- Order 1 Markov Model

ITS – Intelligent Transportation System

MM - Markov Model

HMM- Hidden Markov Model

HHMM– Hierarchical Hidden Markov Model

POI- Point of Interest

OSM- Open Street Map

GPS- Global Positioning System

HHMM- Hierarchical Hidden Markov Model

MMC- Mobility Markov Chain

n-MMC- n^{th} order Mobility Markov Chain

MMM- Mixed Markov Model

MFHD- Most Frequent Hour Day Model

V2V- Vehicle to Vehicle

V2I- Vehicle to Infrastructure

FHWA- Federal Highway Administration

TTOM- Tiered Time Origin Model

CHAPTER 1: Introduction

1.1 Motivation

Moving vehicle data is becoming more prevalent and accessible as traveling with GPS enabled smart phones and in vehicle systems has become the norm. With this increase in vehicle tracking, many new applications are becoming available. Vehicle to infrastructure (V2I) and vehicle to vehicle (V2V) communication are now primary focuses of research at institutions such as the Federal Highway Administration and National Highway Traffic Safety Administration. These research focuses are dependent on up-to-date and accurate vehicle tracking information so that processes can be implemented to react to vehicles on the network. FHWA has begun to explore the area of destination prediction through its Enabling Advanced Traveler Information Services (EnableATIS) project. The advantages to this work are clear: if we are able to accurately predict the future location of a vehicle, smarter suggestions of travel routing can be given. Also, as computer systems in-vehicle become more advanced, travel information and history can be stored in vehicle and relayed to the traffic management center (TMC) when the need arises. This may occur any time from engine startup to when the system sees congestion along your determined route. The TMC would get not just up-to-date information, but future vehicle location; the ability to see network conditions before they occur. This research has applications for the individual user, the network as a whole, and transportation management centers.

Along with the sophistication of ITS infrastructure, this data need is also increasing. One branch of transportation research that is still in its infancy is the topic of vehicle location prediction, or the accurate estimation of where a vehicle is going to be before it gets to its location. Clearly the

applications for this research are multifaceted. In the future, intersections will be adapting to traffic flows to more efficiently phase signals to allow for highest inputs, ramp meters will adapt to changes in demand to allow the correct amount of vehicles onto the highway network, toll roads will shift prices to optimize travel times and increase revenue. What is needed for all of these systems is a way to know where vehicles are, where they are coming from, and where they are going before the vehicle arrives at the location. With this kind of advanced knowledge, systems can react more intelligently by knowing about changing flow rates in advance and reacting to these changes before they occur. The applications also greatly extend to behavior modeling. If the trip destination is able to be predicted accurately, this may lead to more advanced prediction of mode choice, travel route, trip chaining, etc. The knowledge of where a vehicle is going before arrival, or even before any trip information is given, would be a great advancement to transportation systems.

1.2 Objective

The objective of this research is to develop the first trip purpose based short-term destination prediction model using collected GPS, point of interest, and demographic data. There is much to be gained through the accurate prediction of the future location of a vehicle or individual. There is a clear operational advantage of knowing traffic patterns before they occur. If the market penetration is large enough, and vehicle predictions are shared en-masse with Intelligent Transportation Systems, traffic patterns can be known before they occur.

The results portions will put the developed algorithm in comparison to those that have been developed in the past. The final aim is to increase the accuracy (correct prediction of the destination

of the vehicle) when the trip commences. The process for delivering a high accuracy algorithm will be shown from data conception through multiple iterations of Markov modelling.

Each chapter will cover how the approach works under that stage of the research, but on the whole, the technique of machine learning and artificial intelligence algorithms through the WEKA programming suite is used to define trip purpose based on user surroundings and behavior, and a series of Markov models are used to learn about trip behavior and predict future destination locations. By altering the way predictions are made, along with what data is provided to the learning processes, an accurate destination prediction system is created. Finally, how the variables interact with the model is explored via a simple linear regression, showing that the trip purpose model benefits from advanced information in a way that current advanced Markov models cannot.

1.3 Contribution of Work

This work will show that significant contribution to the research has already been made:

- 1) Empirical research on vehicle destination prediction has generally been constrained to the computing and sensor fields. While advanced probabilistic and trajectory modelling of GPS traces has begun the research branch of short term vehicle prediction, it has not moved away from these rudimentary beginnings. New concepts that have great impact on driving destination need to be explored in the realm of transportation engineering. Many of these concepts are explored in this dissertation; driver demographics, trip purpose, land use types, soak time. For real world implementation of short-term destination prediction, these concepts are brought into the body of work. This is the first time trip purpose is used in short-term destination prediction.
- 2) Currently, it is standard practice to implement long training periods to maximize model accuracy. Thus, the models are structured to use long training periods to fill large Markov

Model training sets. The current state-of-the-art start of trip prediction algorithm has a 315 day training period. This is not implementable in the real world, where personal users of traveler information systems require same day information, and traffic management centers cannot wait a year for accurate roadway volumes. The proposed modelling framework is structured to make destination predictions immediately. By selecting the module of the overall model to use based on the data that is available, it is able to maximize accuracy on limited data. After the trip purpose module is turned on at the 5, 15, and 30 trip marks, significant accuracy improvements are seen. In less than a tenth of the learning time used for the current state of the art model, the trip purpose model surpasses it in terms of prediction accuracy.

- 3) There are currently no in-route models that employ the use of point of interest/land use data, the derivation of trip purpose, and subsequent prediction based on said purpose. This research uses a smaller subset of possible destinations by starting the in-route model with destinations only fulfilling the selected purpose. By restricting the purpose in-route, results are both faster and more accurate.
- 4) A new area is studied in attempting to determine socio-economic variables in real time using only the historic GPS location of the participant. By using Census tract information of the locations visited a better understanding of the types of location an individual travels can be determined and hence the likely income level of the participants.

1.4 Outline:

The general outline of the dissertation is as follows. Chapter 2 will first complete an in depth literature review of five areas that greatly impact this research: Travel behavior modelling will start at the basic four step model and explore up to agent based modeling. Next, GPS devices and

their use as a travel survey will be discussed. This includes previous large GPS surveys, and how they were used in travel behavior modelling. Then, trip purpose designation is reviewed, particularly focusing on the new branch of research in using GPS survey data to identify trip purpose. Finally, from trajectory based, grid systems, probabilistic and Markov models, to data linkage and state-of-the-art accuracy levels, a literature review of destination prediction algorithms will be completed.

Data collection is covered in Chapter 3. A full scale GPS survey was conducted with 263 participants including postcard mailer, pilot run (20 GPS devices), online surveys (including website and survey design), representative sampling, GPS device calibration, travel diary forms, instructions for GPS use, data processes, and cleaning. All point of interest data and its' linkage to the GPS travel data is also fully explained. The final dataset is shown with a data dictionary for each variable.

Chapter 4 will explain how the baseline Tiered Time Origin Model is developed. By setting up multiple models to work in conjunction, and only using the highest rated one when available, results show that reasonable results can be found early on in the learning process. The Tiered Time Origin model is the first of its' kind to implement multiple Markov models for prediction with limited trip learning.

Chapter 5 regards the Tiered Time Origin model with advanced trip purpose information prediction algorithm. This paper aims to be the first destination prediction algorithm to bring into account trip purpose for the prediction of the next stop of the user. How trip purpose is derived is also explained with machine learning rule derivation in the Appendix. The results section shows model accuracy in terms of distance and destination uniqueness.

Chapter 6 works upon the previous step by advancing the Markov Model to derive trip purpose with only start of trip information and differing training sets for the trip purpose derivation. Results are shown by training set length, and trip purpose type. A major improvement to previous models is proven. Comparisons are made between the Tiered Time Origin Model and the Trip Purpose Model. Regression analysis shows the increase in accuracy for the trip purpose model is largely due to land use variation and certain types allocating to easier to predict trip purposes.

Chapter 7 consists of taking into account in-route information to constantly update the prediction model as the trip is being taken. In-route GPS points will be added to the existing framework. Once they are linked to the road network, more advanced predictions can be made based on historic route information much the same way that historic origin and time patterns are used. By using destination locations that satisfy only the given trip purpose derived at the start of the trip, more accurate estimations are made.

Chapter 8 is an exploratory research effort in predicting socioeconomics via only GPS travel data. Through taking the average conditions of the travel locations that an individual visits throughout the GPS survey, a more accurate representation of the individual's socio-economics can be established than a basic questionnaire.

Chapter 9 concludes the research effort by summarizing results and discussing overall contributions to the field.

CHAPTER 2: Literature Review

This literature review will focus on a variety of fields that have impacted this research. First, a basic overview of travel behavior theory is done and how these concepts are explored in the dissertation. GPS devices and their recent use as a replacement for travel surveys are overviewed. How trip purpose has been derived from GPS data is the next step of the literature review. The most impactful section is that of destination prediction algorithms. They will be explained ranging from basic trajectory models to in-route spatial-temporal models, and the data intensive nature of their calculation. Finally the theory of Markov model completes the literature review.

2.1 Travel Behavior

Travel Behavior modelling has been a staple in the transportation planning field for decades. It is important to understand how travel behavior has been forecasted in the past to move forward into new research areas such as short-term travel behavior prediction. This section will review the basic concepts that are used in travel behavior modelling that are also used in short term behavior prediction applications.

2.1.1 Four step model

There are 4 basic steps:

1) Trip Generation (Frequency)

By using factors such as household size, income car ownership, residential density, and accessibility, trip production values are generated. Attraction values are the number of trips expected to terminate at a location and are developed based on land-

use and employment categories (Industrial, commercial, services, etc.). Through regression models, the historic trip rates are compared to production and attraction values to determine future expectations. These future expectations are also heavily dependent upon growth factors in the area such as new businesses or homes and increases in population. The end result of trip generation step is the number of trips originating from a zone and the number of trips destined to a zone.

2) Trip Distribution (Destination)

The second step of the four step model is trip distribution. It uses the previously developed trip productions and attractions to create a link between the two. A Trip Matrix is derived to account for every trip generated for each origin destination pair. A Gravity model aims to formulate relative accessibility by comparing distance and travel cost between locations. Mathematically, the gravity model often takes the form:

$$T_{ij} = K_i K_j T_i T_j f(C_{ij}) \quad (1)$$

$$\sum_j T_{ij} = T_i \quad \sum_i T_{ij} = T_j \quad T_i \quad (2)$$

$$K_i = 1 / \sum_j K_j T_j f(C_{ij}) \quad , \quad K_j = 1 / \sum_i K_i T_i f(C_{ij}) \quad (3)$$

where

- T_{ij} = Trips between origin i and destination j
- T_i = Trips originating at i
- T_j = Trips destined for j
- C_{ij} = travel cost between i and j
- K_i, K_j = balancing factors solved iteratively.
- f = distance decay factor, as in the accessibility model

The Gravity model is implemented by city planners with factors regarding costs and attractiveness of locations to match up trip demand with supply. The final trip matrix will be solved iteratively until convergence is met. Using Dynamic systems, these values may change as the trip is taking place, or as information about locations changes.

3) Modal Split (Mode)

The third step in the four step model is modal split. The mode of transportation is now identified for each individual trip that is to be taken. A probabilistic Logit model is built that looks at the relative attractiveness of all mode options. The probability of each approach is derived and the most likely alternative is attributed to each trip. The Modal Split can be a simple auto-transit Logit model, or be explained via a Nested Logit structure that explores more nuanced relationships between modes.

4) Assignment (Route)

The trip assignment by route is the final step in four step traffic assignment. There are a variety of options for assigning route. Assignment can be completed by various methods. Deterministic (shortest Path, minimum cost), Stochastic (Discrete Choice) and also by Equilibrium (User Equilibrium, or System Optimal).

2.1.2 Representing Travel Behavior

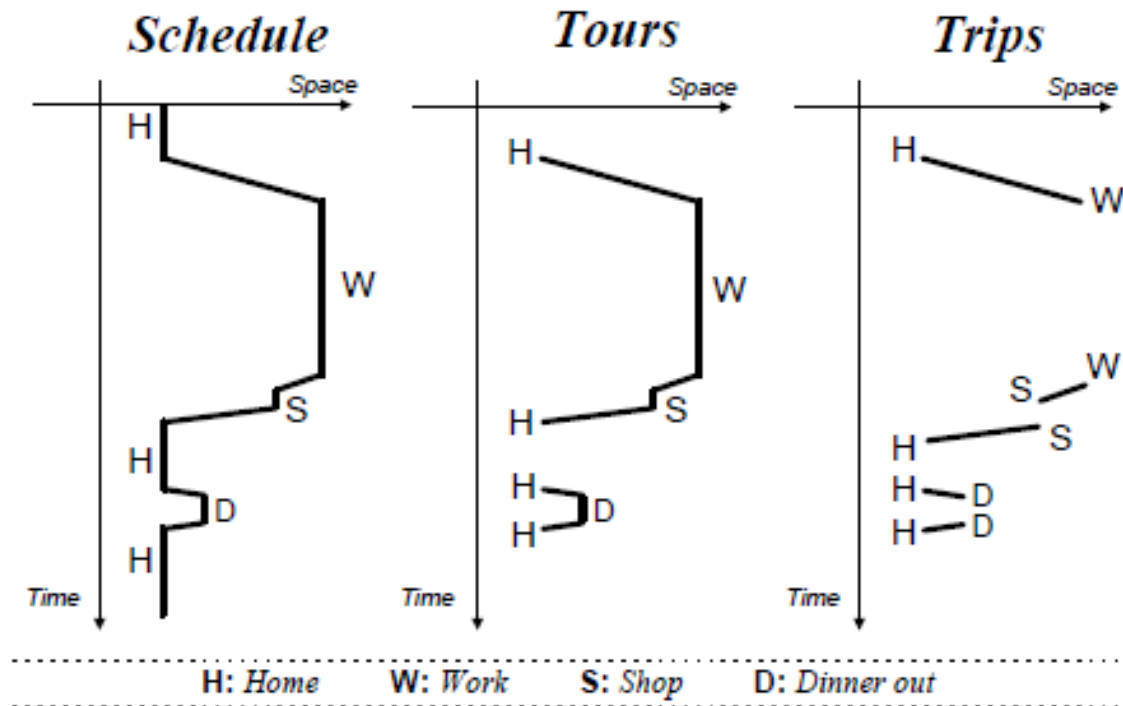


Figure 1- Travel Behavior Representation Ben-Akiva (2008)

The manner of travel can also be represented in a number of ways. The four step model approaches the problem in terms of trips, but is a simplistic way of viewing travel behavior that has some limitations. First, the model identifies demand in terms of trips, rather than activities. An activity may require multiple trips to accomplish, yet the four step model does not take this into account. Next, the behavior that is modeled in previous steps are unaffected by the choices that come after. Based on these limitations, travel behavior research moved on to tour based modelling. With tour based modeling, trip chains are possible, making it easier to understand what the next location of a trip chain will be when taken into account past locations. Instead of seeing home to work then work back home as two separate trips, it can be seen as a trip chain whose previous locations have an impact of future ones. Many of these ideas are carried over to the short term destination prediction algorithm. Classifying trips by the time of day, day of week, trip purpose, etc. all have basis in traditional travel behavior modelling.

2.2 GPS Travel Survey

There is a wealth of research conducted on GPS travel surveys. This work began with Wolf (2001) when data loggers were used to replace or supplement electronic travel diaries. The Commute Atlanta project is the largest GPS based travel survey in the country (Schonfelder et al. 2005).

2.3 Trip Purpose Designation

The derivation of trip purpose is an extremely important aspect of this research, as the trip purpose derived is used as the basis for the destination selected by the short term destination prediction algorithm. This section covers the history of trip purpose prediction and the expected accuracy of the machine learning model used in the final algorithm.

Since the initial use of GPS devices as a form of travel diary replacement, several research papers have aimed to derive trip purpose using only GPS data and the area's point of interest/land use data: Axhausen (2003 & 2008), Griffin (2005), McGowen (2006), Bohte (2009). Beyond the GPS and land use data, it has been found that the user's socio-economic characteristics such as household composition, demographics, and location of the user's home and work is very helpful for the categorization of GPS strings to trip purpose. Stopher et al. (2008) derived heuristic rules using 43 trips with land use data and geographical coordinates of user's most frequently travel home, work, and shopping locations.

Chen (2010) was able to cluster the ends of trip locations into similar activity types. These deterministic rules could more accurately classify trip purposes in low density areas. However, this approach has difficulty in determining trip purpose for high-density locations due to the clustering algorithm. A Multinomial Logit Model was then used to give the probability that a location and trip serves one of the 4 trip purposes employed.

Deng et al. (2010) began to derive rules using a machine learning decision tree approach: the overall accuracy was 87.6% for 226 trips. They were the first to use GPS data, social-demographic and socio-economic characteristics to construct a decision tree of trip purpose. Most recently reports show that accuracy between 80 and 85% can be found by employing a random-forest machine learning approach (Montini et al., 2014) on larger datasets consisting of over 150 participants for a week's time.

The methodology from Lu (2013) was employed in this research due to the similarity in datasets and access to the model. Using closest Point Of Interest information, the model has an accuracy of 80.58%. The land use categorization was employed in the same manner to retain the accuracy found in this research. Lu (2014) further explored trip purpose estimation for urban travel by using NHTS Add-on Data. The model shows accuracy above 80% for trip types Home, Work, School, and Shopping, but unsatisfactorily for types Social, other, and driving. This approach was used in Chapter 5 of this research to get baseline trip purpose estimation, yet had to be changed in chapter 6 to implement pre-destination estimation. The methodologies will be shown in-depth in those chapters.

The fields of trip purpose derivation and vehicle destination prediction have been gaining much interest with advances being made. Despite the gains that could be made in vehicle destination prediction algorithms with the use of trip purpose estimation, no research has been done. This paper presents a new branch in vehicle destination prediction with the addition of trip purpose as a classifier to base predictions. Results presented here show that this additional information is invaluable, with large gains in prediction accuracy.

2.4 Destination Prediction Lit Review

There are many ways in which vehicle prediction algorithms currently operate. Trajectory data and historic GPS points are used to create a basis for future trips. Probabilistic models that attempt to predict future vehicle destinations with GPS data by Ashbrook and Starner (2003). A Markov Model (MM) is used to find the probabilistically most likely next location based on a subset of previous locations.

Liao, Fox and Kautz (2004) developed multimodal destination prediction. The hierarchical model is able to increase accuracy with Bayesian inference while including different modes other than personal vehicle. The research does not require a map database for linking locations to roadways and inferring previous trips' route to determine location.

Some models have moved away from learning from travel patterns and historic trip traces and instead use only the trajectory data being presented from the current trip (Xue et al, 2013). This approach has utility because less data is used and post processing of the data is decreased. The drawback is lower accuracy of the model and the inability to increase accuracy through travel data. Karbassi and Barth (2003) began the use of vehicle position tracking to more accurately predict the route and arrival time for public transportation vehicles. In their application, the modelers have advanced knowledge of the intended vehicle destination. The algorithms derived promising results in route and time estimation, with improvements being possible with increased point frequency and accuracy.

Destination prediction studies used historical vehicle trajectories in two general ways: firstly using external information to improve the accuracy of predicted destinations, and secondly by customizing each prediction to the individual user.

The set of information that can be added to the vehicle trajectory is ever increasing. This includes travel time, trip length, road conditions, driving habits, time-of-day, day-of-week, and velocity. (Horvitz and Krumm, 2006, 2007, 2012)

(Terada et al, 2006) were the first to bring trip purpose into post processing. This is done by estimating the destination location for the ongoing trip, then assigning a trip purpose to the given trip depending on the land use of the estimated location. The trip purpose portion of the modeling is brought in only after the estimation is made, and is used to give suggestions for other locations in the area of similar type. They acknowledge however that “from the start to the middle stage, the probability of a correct destination is too low to provide services according to predicted destinations.” Only the ends of trips are being supported by their methodology, when a user is close to their destination. This dissertation increases the early stages of prediction to provide services on predicted destinations.

(Terada et al, 2008) built upon previous work by adding map matching to the algorithm. The map matching method cuts the path into intervals and calculates the shortest path to help determine the destination of the trip. The trip purpose application in their 2006 research was not further explored in this work.

(Alvarez-Garcia et al, 2010) proposed a destination prediction model that uses current location at the beginning of the trip. The modeling structure is most similar to the one used in this research’s approach in that the support map of the algorithm is generated purely by the GPS points and is thus independent from a street map database.. The Alvarez-Garcia paper does not give a prediction at the start of trip, only for 25% of the way through the trip (48.54% accuracy). As a comparison, the results in this work will show an accuracy of 52.34% at the very start of the trip, not 25%; a 4% increase, earlier in the trip.

Lei et al. (2011) use a sub trajectory synthesis which employs an offline Markov model to predict the probability of any given trajectory that is made online.

Recent papers have advanced the field through more advanced algorithms via map-matching, sub trajectory synthesis, and other data mining approaches. These methods greatly increase predictive accuracy but require in-route information and do not report higher predictive accuracy at the beginning of the trip. To increase the baseline accuracy, new methods, and data elements are needed. This paper explores these needs through the use of trip purpose characteristics.

Currently, multiple papers have looked into destination prediction from the computer science fields of ubiquitous computing, artificial intelligence, data mining, and other similar specialties. Early papers have focused more on the trajectory of the routes and predict end location from trajectory and a classification of common end nodes, while more recent research has explored temporal and spatial information for trip end prediction. The user's travel patterns were learned and routines from the data as well as the transportation network were found for a single individual (Liao et al. 2003). This type of work has the ability to estimate when a user has missed a mode switch due to the implementation of roadway networks and the user's normal mode switch (at parking lots, bus stops, etc.) Using GPS data, the mode of travel has been explored, and under what circumstances switching between modes occurs (Zhang et al. 2008). The most current papers in the field of Spatial-Temporal Trajectory Models specify a future time period, then attempt to estimate the location based on current trajectories and common nodes near future locations. The work creates important nodes based on their frequency, and then uses Markov chains to estimate how they will move between them (Lei et al. 2011). Similarity between movements and potential to move between these specified frequent nodes are used as variables for deciding the future locations. Research has been done that deals with the accuracy issues in GPS data, and builds an algorithm

designed for imprecise trajectories (Yang 2006). Tables in (Gao 2012) show that the current most advanced Spatial-Temporal location prediction models yield an accuracy ranging from 46-50%.

Much of the work comes in the form of database management as well. An entire subset of research in locations prediction deals with the arrangement of data for the use of data mining. Pfoser (2006) deals with the modeling of databases and its systems to make data more viable for traffic congestion alerts. Boulmakoul (2012) arranges data into a Meta-model that allows for efficient data sharing and easier trajectory querying.

Karbassi and Barth (2003) use historic car sharing data to estimate the route that a vehicle will take between known start and end locations. Torkkola et al. (2007) look at historic GPS traces and predict standard routes that have previously been taken.

Krumm and Horvitz (2006, 2007) also attempt to predict the location of the driver's destination in-route. They employ road network and historic travel pattern integration into their trajectory models. Patterson (2003) states that future work of trip purpose linking with personal calendars could lead to increased accuracy in prediction models, but no further development in this area can be found.

Krumm (2009) has led the way in in-route turn prediction by employing an nth-order Markov Model to probabilistically predict the future road segment of a trip based on the previous road segments that the user has taken on that trip. Dissimilar from previous research, predictions are not based on historic travel patterns, or with the beginning or end location of the current trip. The model is able to predict the next step from the previous step's observation.

The literature shows the best start of trip model in Gao et al.'s (2012) HPHD with an accuracy of 50.05%. The dataset includes 3,373 locations and a 315 day training set. The Markov Model (Tiered Time Origin Model: Chapter 4) in this paper uses both most frequent hour and day

categorization and has an accuracy of 45.7% with the GPS data in this paper. The purpose model, with a 30 trip training period and 41,304 locations has an after training set accuracy of 52.74%, improving the best accuracy model in the literature by 2.69%. Taking into account the network size (including long distance trips in 22 states) and over ten times the number of locations, the improvements may be larger than the accuracy improvements suggest.

2.5 Markov Modelling

A stochastic process is Markov if the conditional probability distribution of future states of the process (conditional on both past and present values) depends only upon the present state, not on the sequence of events that preceded it.

The nth order Markov model aims to predict the next location (either final destination, or next stop, next link, etc...) by creating a probability function on the previous locations. The basic 1st order Markov model depends only on the current location, creating a histogram of previous locations that have derived from the current location, and then the most likely scenario is chosen from there. If there exists a 2nd order Markov model, the probability function is derived from the historic travel patterns of the two locations available in order.

$X(i)$ denotes the location, with i as the time variable. The next step to be predicted would be $X(1)$, while the current step is indicated by $X(0)$. Looking back at historic travel data, information on previous steps ($X(-1)$, $X(-2)$, etc...) can be helpful in the current step prediction. The nth order markov model then has the form of

$$P_n[X(m)] = P[X(m)|X(-n+1), X(-n+2), \dots, X(0)] \quad (4)$$

Where m is the numbered next predicted location.

2.5.1 Markov Model Explanation

The Markov Process that best explains the problem explored in this thesis as a Discrete-time Markov chain.

		State Space	
		Discrete	Continuous
Time	Discrete	Discrete-time Markov Chain	Discrete-time Markov Process
	Continuous	Continuous-time Markov Chain	Continuous-time Markov Process

Figure 2- Classification of Markov Processes (Ibe[2013])

The time aspect of the problem is defined as a jump procedure. Since the time between the time steps (1 minute intervals) is discrete, the process is called a jump chain. Each jump is the time in between the start of the trip at the origin location, and the destination GPS point. At the start of the Markov chain, the time period is unknown. If the time periods are known, the chains are not considered Markov chain jumps. The time between jumps to another state is known as a holding time. In this research, the holding time is the soak time of the vehicle (the time that the engine stops to the next time it is turned on). The Markov chain starts back up from the new location (the previous jump's destination) and jumps to the next state (both time and space).

The Markov Chain of the system as a whole is actually rather simple. The agent (the driver of the automobile) is in a current state. This state includes the time and location of the vehicle. There are both knowns; land use types, day of the week, demographic information, and unknowns; trip purpose type, number of people in the vehicle, etc. Based on the parameters of this state, a

probability function exists that explains the probability of driving (jumping) to the next state (destination). After this jump, the model moves to the next step of the chain where another jump will be made. The overarching application of the thesis is to take much of the known variables that can be found and linked to the GPS data, and form some knowledge base of the unknown (or hidden) aspects of the agent's state. Based on an improved knowledge base, the probability of future states can be more accurately determined leading to improved models for those state locations. This is the first time that land use data will be used as a known variable in the destination prediction Markov process to make some assumptions about hidden variables (trip purpose).

2.5.2 Markov Order

The Markov chain that is being modeled with the approach in the paper is a 1st order Markov chain. The probability of the next step occurring is simply a function of the current step. This is true because only the known location information of the current location is used for the estimation of the future step. If for instance, the previously known location was brought into the estimation, the process would be a 2nd order Markov chain. The classification of the model changes regarding which phase of the research is being explored. 2nd order processes are brought in later in the paper, but the initial destination procedure is made with only current location state information known.

2.5.3 State-Transition Diagram

A state-transition diagram can be formed to imitate the problem explored in this thesis. In the example case imagine that each state (1, 2, 3, and 4) is equal to a location for a participant (Home, Work, School, and Shopping). The below example will show how some biases exist in deriving the next state using the current state. The models proposed in this research aims to tease out these biases from historic travel patterns to predict the next state.

Take for instance a theoretical problem of rolling a 4 sided die. The user does not know if the die is balanced or biased, so the experimenter rolls the die 100 times to find out what numbers are rolled. Using these values of numbers rolled a state transition diagram can be developed. Using basic math, the participant knows that there should be a 25% chance of rolling each number. The user rolls 40 4's, 20 1's, 20 2's, and 20 3's. The known variables are the sides of the die, whereas the unknown Markov properties are whether or not the die is biased, and how so. Below is a basic state-transition diagram for the die:

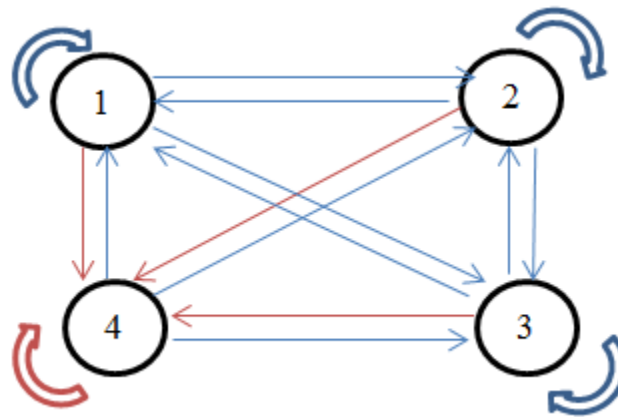


Figure 3- State Transition Diagram Example

Red arrows signify a 0.4 chance of occurring, whereas blue lines indicate a 0.2 chance of occurring. The state-transition diagram shows that the 4 sided die is biased towards rolling a 4. The destination prediction problem can be drawn similarly, although in a much more complex network. Depending on the current state that the vehicle is in (location), there is some unknown chance of the vehicle traveling to another location during the next step. After normal travel is performed, a reasonable state transition diagram can be drawn and future predictions can benefit from advanced information about the nature of the Markov chain. By adding information such as land use types

and trip purpose, the probabilities become clearer (like the biases in a die), and deviations are less hidden.

Finally, it should be noted that the destination point of a trip is not necessarily the origin of the following trip. This is due to common problems with GPS signal and the time that is required to obtain an accurate GPS fix. For this reason, the Origin Destination identification code for the origin of a trip will not necessarily match the destination from the trip before. This can lead to problems down the line for the 2nd order Markov Model when a check is done for the previous location identifier. For this reason, if an identifier code is only a single integer away from the next trip, it is considered the same value in running the Markov Model.

2.5.4 Graphical Representation

The problem proposed can best be explained through a Bayesian Network. An undirected or cyclical Bayesian Network is also considered a Markov Network.

Take a simplified example of a desire to determine where an individual's next trip will be. We want to find the joint probability function that a participant's next location will be "School". We will use two variables to build our Bayesian Network: Time and Origin (T & O respectively). The third variable, which we are attempted to estimate, is Destination (D). The joint probability function is thus:

$$P(D,T,O) = P(D|T,O)P(T|O)P(O) \quad (5)$$

Now variables D, O, & T may have several states. In the real application of the model, there are over 5000 unique origins and destinations, but for this example we will use 3: Home, Work, and School. Time has 24 states (one for each hour of the day) in the actual application, but will be limited to am and pm for this example.

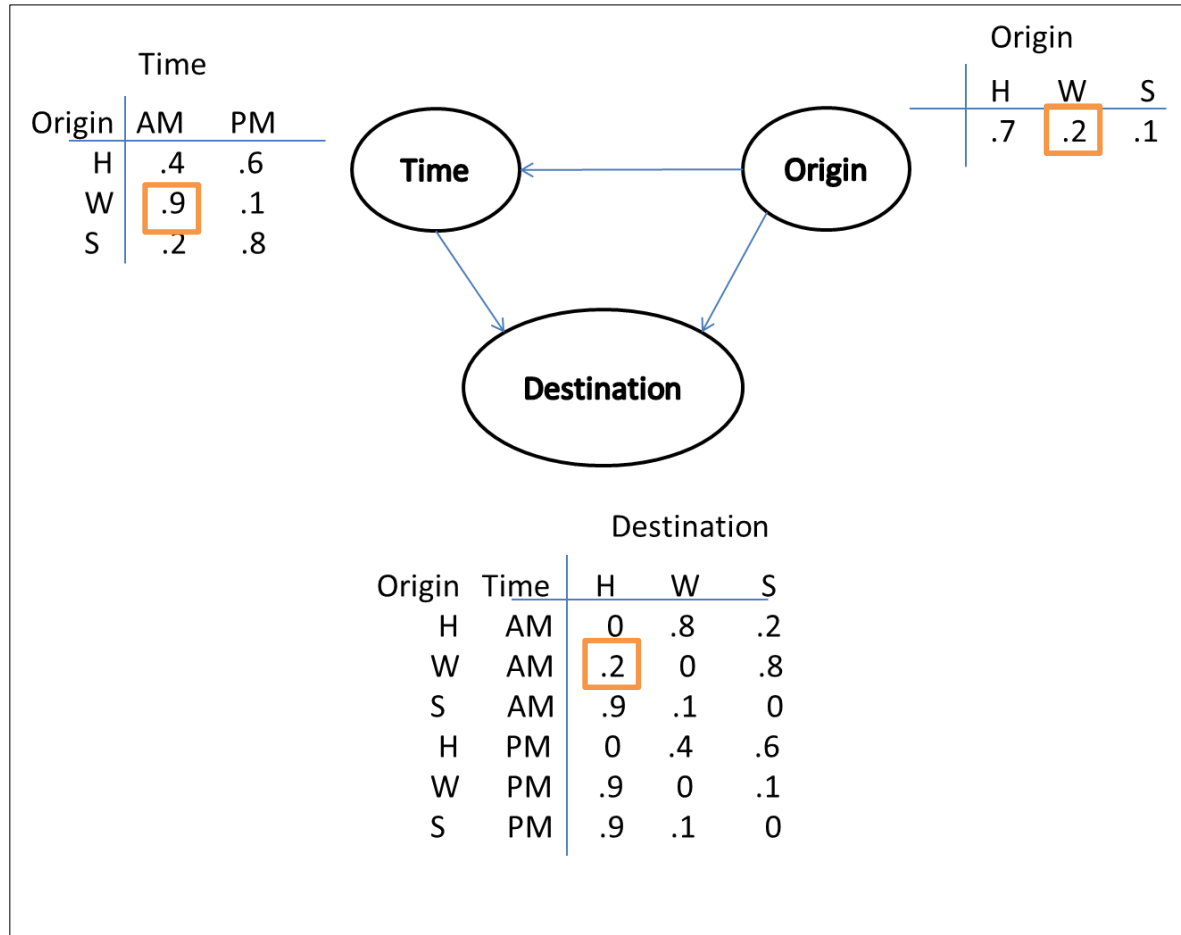


Figure 4- Bayesian Graph of example problem

Suppose we want to know the probability that the user's destination is Home, given that the user is at Work. The following conditional probability function can be calculated and summed for all variables.

$$P(O=W|D=H) = \frac{\sum_{T \in \{am, pm\}} P(D=H, T, O=W)}{\sum_{T, O \in \{H, W, S\}} P(D=H, T, O)} \quad (6)$$

By filling out each conditional probability in the diagram above, we are able to calculate each probability. These are highlighted in the above chart.

$$P(D=H,O=W,T=AM)=$$

$$= .2 \times .9 \times .2$$

$$=.036$$

$$P(O=W|D=H) = .036 + .018 / (.036 + .162 + .018 + .002)$$

$$= .054 / .218 = 24.7\%$$

Thus, there is roughly a 25 percent chance of the participant's next trip being home, given that the user is currently at work.

This is the theory behind the Bayesian networks created for the underlying Markov Models.

Although, instead of a simple network including 3 variables and only 2 or 3 states per variable, there are 5 or 6 variables with a possibility of thousands of states per variable. This is explored later in this paper in section 4.4 Matrix Size.

2.6 Literature Review Conclusions

This chapter has shown that the field of short term destination prediction, while new, is flourishing with different forms of probabilistic models. GPS travel surveys have been implemented for over ten years, now being included in the National Household Travel Survey as an add-on, and large regional travel surveys (Schonfelder et al, 2005) as the major data collection endeavor. However,

despite having access to other data sources such as demographics, land use, and point of interest data, they have not been implemented in the research field.

The modeling frameworks for short-term destination prediction also leave open the availability for estimation of different travel behaviors or demographics. This truly opens up entire fields that have not yet been explored in transportation.

The following section will explain what data is needed to accomplish these research goals, and how it was collected for implementation.

CHAPTER 3: Data

The data chapter will be broken down into multiple sections that explain each portion of the overall end data product. These include: GPS Data, Land use and POI data, and Data Processing/integration.

3.1 GPS Data

In this subsection, the GPS survey conducted for this research will be explained. All data used in this research was acquired through the GPS survey spanning from October 2011 to February 2012.

The breakdown of the section will be:

- Pilot Survey
- GPS device selection
- Postcard and Email Campaign
- Website Design
- Survey Design
- Respondent Sampling
- GPS mailing and returns
- Online Travel Diaries
- Data analysis
- Survey Statistics

All forms sent out to participants with instructions on GPS installation and online trip diary example are in the appendix section B.

3.1.1 Pilot Survey

The pilot survey was used to determine the feasibility of a large scale survey (230 participants over 70 days) for variables such as; response rate, data collection, man hours, total time, etc. The framework for the pilot survey is similar to the full scale survey except for some small changes including size of the participant group, 20 participants, and length of the survey, 2 weeks.

3.1.2 GPS Device

The GPS device was selected at the start of the project. In previous work, a Maryland research team had worked with the QSTARZ 1000XT. It has a battery life of 42 hours after two hours of charging with more than acceptable accuracy (<3meters). The cost of each device is approximately \$80. With the device continuously plugged into participants' cars, the device will stay charged permanently. In case of no action, the device will switch into sleep mode which drastically conserves battery life. The GPS records at one minute intervals. This interval accomplishes a number of purposes. The data collected can be used to determine origins and destinations. The maximum stop interval for determining destinations is two minutes; this is based on previous literature review on stop time (Wolf, 2001). The shorter the collection time interval, the more useful the data is in determining vehicle route. Since the GPS survey lasts for a period of two months, the one minute time interval was selected so as not to exceed the data capacity of the device. If the car is driven continuously for two straight months the GPS device will be able to record every data point.

Each GPS device is set to the 1 minute time interval specification and given a unique identifier for each individual. After each device has been manually formatted and charged they are ready to be shipped to participants acquired through the postcard and email campaign.

3.1.3 Postcard and E-mail Campaign

The postcard and E-mail campaign are used to obtain participants for the travel survey. The purpose of this campaign is to get recipients to go to the travel survey website and fill out the initial participation form for full participation. From the participant form, users will be selected for data collection.

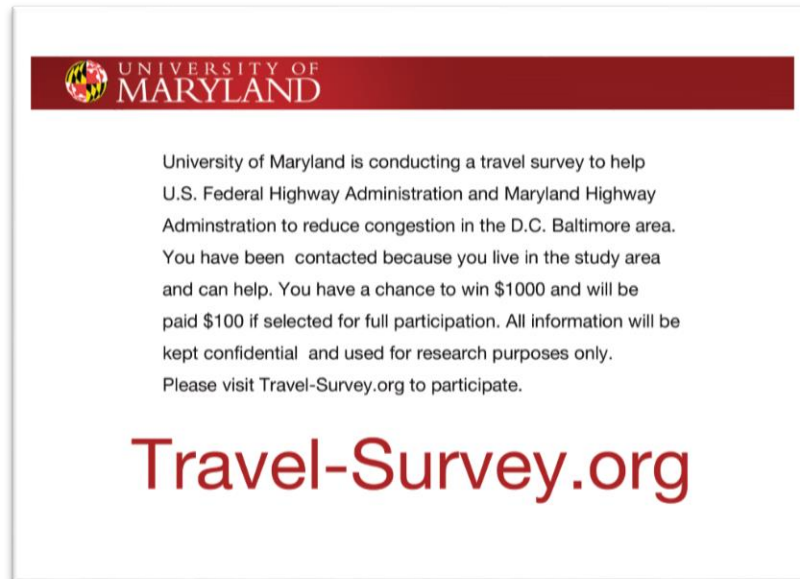


Figure 5: Front side of postcard

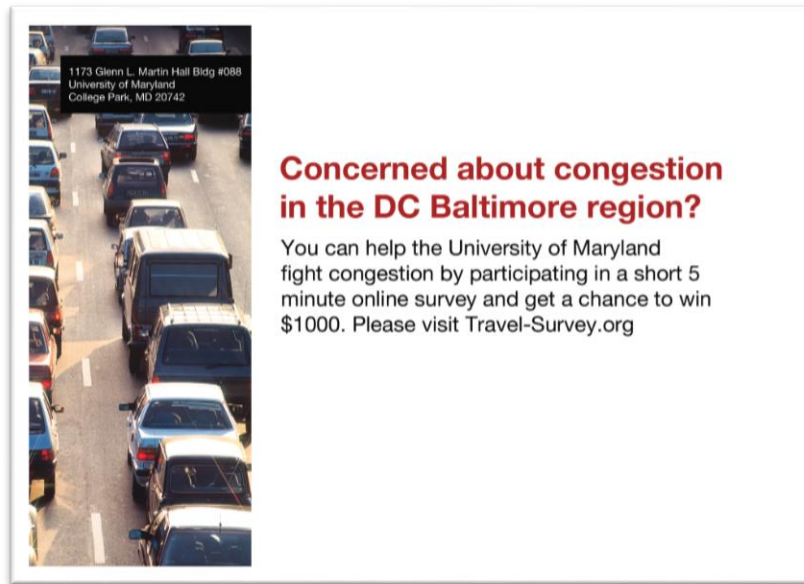


Figure 6: Back side of postcard

The target geographic area is the community encased in the I270-MD200-I495-I95 region. This is in both Montgomery and Prince George’s County. The postcard campaign was sent to 22,000 households. Of the 22,000 postcards sent, there were 893 registrations on the website for a response rate of about 4%

While the email campaign was originally used in the pilot survey, the response rate was so low that it was not used for the full scale study. The response rates for the pilot survey email campaign are as follows:

Table 1: Email campaign tracking report numbers

Sent Messages	1000
---------------	------

Received	895			
Messages				
Total Bounces	69			
Soft Bounces	5			
Hard Bounces	64			
Undelivered	36			
Metric	Total	Total Rate	Unique	Unique Rate
Opens	23	2.57%	22	2.46%
ClickThroughs	13	1.45%	12	1.34%
Unsubscribes	1	0.11%	1	0.11%

Of the 1000 emails sent out, 22 individuals opened the email to read it and only 8 registered to take part in the online survey. The registration rate for the email campaign comes to 0.8% compared to the 4% level of the postcard survey.

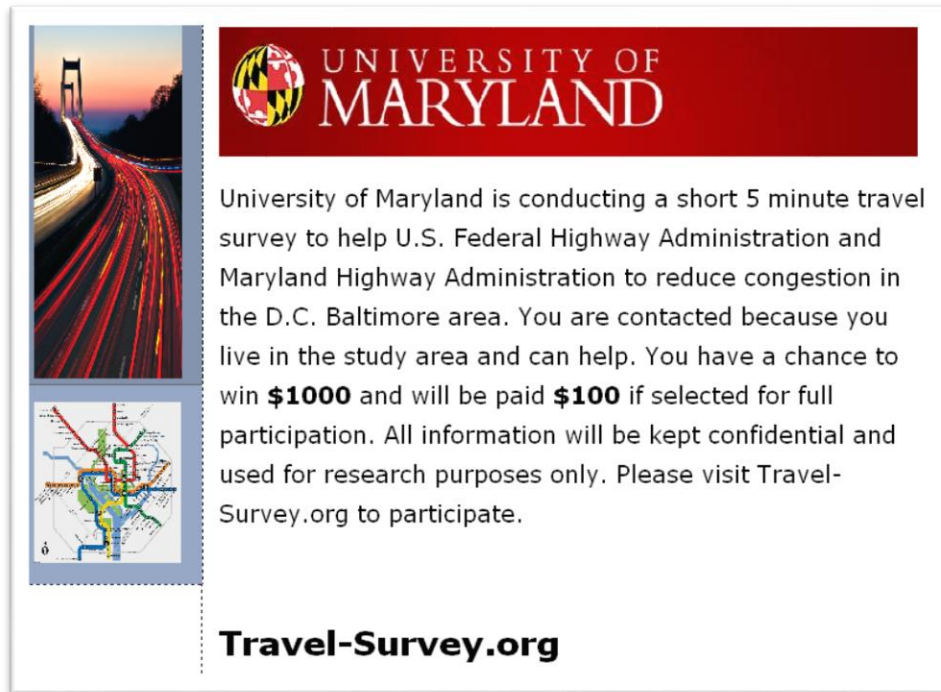


Figure 7: Email sent to individuals in the project area

3.1.4 Website Design

The website is an important aspect of the travel survey because we focus all possible participants on this page. To get people to transition from receiving a postcard to participating in our GPS survey, the website must be easy to operate and understand. In this section a breakdown of each page will be shown and how it is designed to get users to participate. If viewing this document via electronic means, please visit travel-survey.org to view the website.

From the participation page they can fill out the initial form to allow us to contact them. The main page of the website gives a small blurb about why travel surveys are important as well as remind them of the monetary benefit of taking part in the survey (\$100). Also included on every page is a link back to the homepage, a link to the Maryland State Highway Administration (SHA), the

Federal Highway Administration (FHWA), and a link to the University of Maryland (UMD) Website.

A FAQ page and contact information page are also included in case people are confused as to how to participate. These buttons are available on every page except when filling out the participation form.



Figure 8: Front page of the survey website

3.1.5 Survey Design

In this section the initial participation form will be explained. This form is used to get basic demographic and driving behavior information so that we can contact the participants after the respondent sampling portion. The demographic information is used for participant selection, and in later steps of the research is also used in the learning algorithms for the creation of search and decision rules.

This form contains three separate forms: one for people who received the postcard, one for people who received an email from us, and one for those that found our website by another means. The purpose of having these different forms is for gathering slightly different information from these three groups. For example, for those who received our postcard, we already have their mailing address, so it is not necessary to ask that information again. We try to minimize the number of questions to reduce burden on the participant. In this way we attempt to maximize the number of responses we get by making the survey less invasive. Below is the opening participation form screen:



Figure 9: Participation page to start the survey

Once the participants begin the survey, all other links to tech support, financial supports, and participation are dropped so they will not leave this page. The only thing left on the page is the questionnaire and the links to UMD, SHA, and FHWA.

The first page includes a series of demographic questions so that respondents can be selected for full participation and used for later analysis. Also, we gain knowledge about their travel patterns by asking about frequent roads taken.

Survey Participation

Thank you for your participation as it is very important in the improvement of the area's transportation service.

Please input the 9-digit zip code on the postcard you received.

Do you have a valid driver's license?

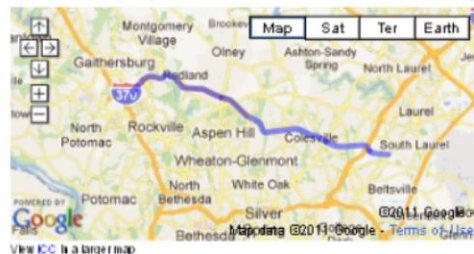
☐ Yes ☐ No

How many miles do you drive on a typical day?

Your top 3 most used roadways:

- ☐ I-95
☐ I-270
☐ I-495
☐ I-66
☐ I-695
☐ I-70
☐ I-370
☐ Route 32
☐ Route 198
☐ Route 295 Baltimore Washington Pkwy
☐ Route 29 Columbia Pike
☐ Route 97 Georgia Avenue
☐ Route 355 Roderick Pike
☐ US 50 John Hanson Highway
☐ Inter County Connector
☐ Other, please specify:

The Intercounty Connector (ICC) is a new roadway that will connect the I-270 and I-95 corridors. Shown below in blue is the location of the ICC. (You can move around and zoom in on the map to get a better idea of its location)



How often do you think you will use the ICC? Open Nov. 11, 2011.

----- Please Select -----

Please tell us a little more about yourself. This will help us get a more representative travel pattern.

Gender

☒ Male ☐ Female

Age

45

Education

Associate Degree

Annual Household Income

\$100,000 - \$125,000

NEXT

Figure 10: Participation form first page

Figure 11: Participation form second page

The final page includes contact information. While users may choose to not fill out certain sections of this page, an error message alerts them to make sure they meant to leave these sections blank. The section for email is mandatory however, as it serves two purposes: it allows us to contact them for the GPS portion of the survey and it is used for an instant follow-up email notifying that their submission was successful.

3.1.6 Respondent Sampling

Once information is obtained on travel behavior and demographics, the participants for the GPS portion can then be chosen. Participants are selected to get a representative sample of the surrounding area. Demographics considered are: Sex, Age, Household Income, and Education level (Appendix). While not a demographic measure, the driver's license status and amount of

average driving per day is checked in the response form. This is to ensure that each participant will be supplying the project with data.

#	<input type="checkbox"/>	Date Submitted	Submitter's IP Address	Language Code	Identifier	License	travel time	Most used roadways	Other Roadways	ICC	Gender	Age	Education	Income	First Name	Last Name	Phone
1	<input type="checkbox"/>	2011-09-11 14:02:08	173.66.196.105	en-GB	20903-1314	Yes	25	I-495, Route 29 Columbia Pike, Other, please specify:	New Hampshire Avenue	Less than once a month	Female	53	Master's Degree	\$50,000 - \$75,000			
2	<input type="checkbox"/>	2011-09-10 14:16:52	208.68.5.251	en-GB	20910	Yes	30	I-495, Route 29 Columbia Pike, Route 97 Georgia Avenue		Less than once a month	Female	60	Doctorate Degree	\$150,000 & up			
3	<input type="checkbox"/>	2011-09-05 17:25:46	70.110.22.103	en-GB	20901-1947	Yes	25	I-495, Route 29 Columbia Pike, Route 97 Georgia Avenue		Once a week	Male	47	Master's Degree	\$100,000 - \$125,000			
4	<input type="checkbox"/>	2011-09-01 16:30:25	71.191.132.217	en-GB	20906-1118	Yes	20	I-95, I-495, Route 198		Several days a week	Male	45	Master's Degree	\$150,000 & up			
5	<input type="checkbox"/>	2011-08-30 16:35:48	68.50.114.31	en-GB	20906-2672	Yes	20	Route 355 Rockville Pike, Inter County Connector		Several days a week	Male	48	Associate Degree	\$50,000 - \$75,000			
6	<input type="checkbox"/>	2011-08-30 10:49:14	216.81.81.82	en-GB	209102246	Yes	40	I-495, Route 295 Baltimore Washington Parkway, Route 97 Georgia Avenue		Less than once a month	Male	46	Bachelor's Degree	\$150,000 & up			

Figure 12: Example of participation form results

A stratified random sampling procedure is employed to create a sample that matches demographics for the area closely. While participants were selected based upon demographic information to give the most representative sample, a bias towards over-education and high income does occur. To view the participation demographics compared to those of the surrounding area, please go to Appendix C.

3.1.7 GPS Mailing

Each package mailed to participants includes the GPS device, instructions on use, and the GPS charger. After the survey period was over (2 months) each participant received a preaddressed prepaid shipping label to be sent back to the University of Maryland. The participant simply puts the GPS device back in the box and drops it off at any FedEx location.

3.1.8 Travel Survey

The travel survey is the only link between the raw data from the GPS device and the actual user travel patterns. The online survey was designed to include all the pertinent travel information while being simple to fill out. On the first page the user inputs their name and the date for which they are filling out the form. This is used to link to the GPS received from the participant. Then, the next page asks for information on a single trip for the day. If the user has more trips for that day, they are prompted to click the continue button, or if they are finished recording all their trips, they can finish by clicking submit. It is simple, yet provides us with time, trip purpose, travel mode, and exact location.

GPS Survey

Thank you for taking part in the University of Maryland's GPS travel survey. Please fill out this form as fully as possible.

If you have any questions or have trouble filling out the survey, please email me at ckrause@umd.edu or call me at 301-852-3392.

Thanks for your help

Cory Krause





First Name (*)

Last Name (*)

Enter date for which you are filling out your travel diary: (*)

September 2011						
Su	Mo	Tu	We	Th	Fr	Sa
28	29	30	31	1	2	3
4	5	6	7	8	9	10
11	12	13	14	15	16	17
18	19	20	21	22	23	24
25	26	27	28	29	30	1
2	3	4	5	6	7	8

Figure 13: Travel Survey First Page

[Home](#) > [GPS Participation](#)

GPS Survey

Each page will be used for an individual trip. For example, if you drove from work to the grocery store, then to the day care then home, that is 3 separate trips (and would use 3 pages of this survey):

Where were you at 3am?

Did you have at least one trip this day (either by car, walking, train, any mode of transportation) (*) ☐ Did not have any trips ☐ I had at least one trip

What time did you leave this location? (hh:mm am/pm)

What mode of transportation did you take on this trip?

What was the purpose of the trip?

What time did you arrive at the destination of this trip? (hh:mm am/pm)

Where is that destination?

Nearest intersection:

Nearest landmark? (Building, park, school, church, etc...)

If you have no more trips to report, please click submit to finish the form.

If you have more trips for the day, please click on next.

Figure 14: Travel Survey Second Page

The daily trip diary created for the agent based modelling project has not been used in the short term destination prediction research.

3.1.9 GPS Data Checking

The final step in the pilot study is to upload and check the data to make sure that each device functions properly. This is done in a two-step process. First, the data is opened in an excel file to make sure the data points are available. Then, using a geospatial tool (using GeoStats' TravTime software) the data points are checked for major error in data location.

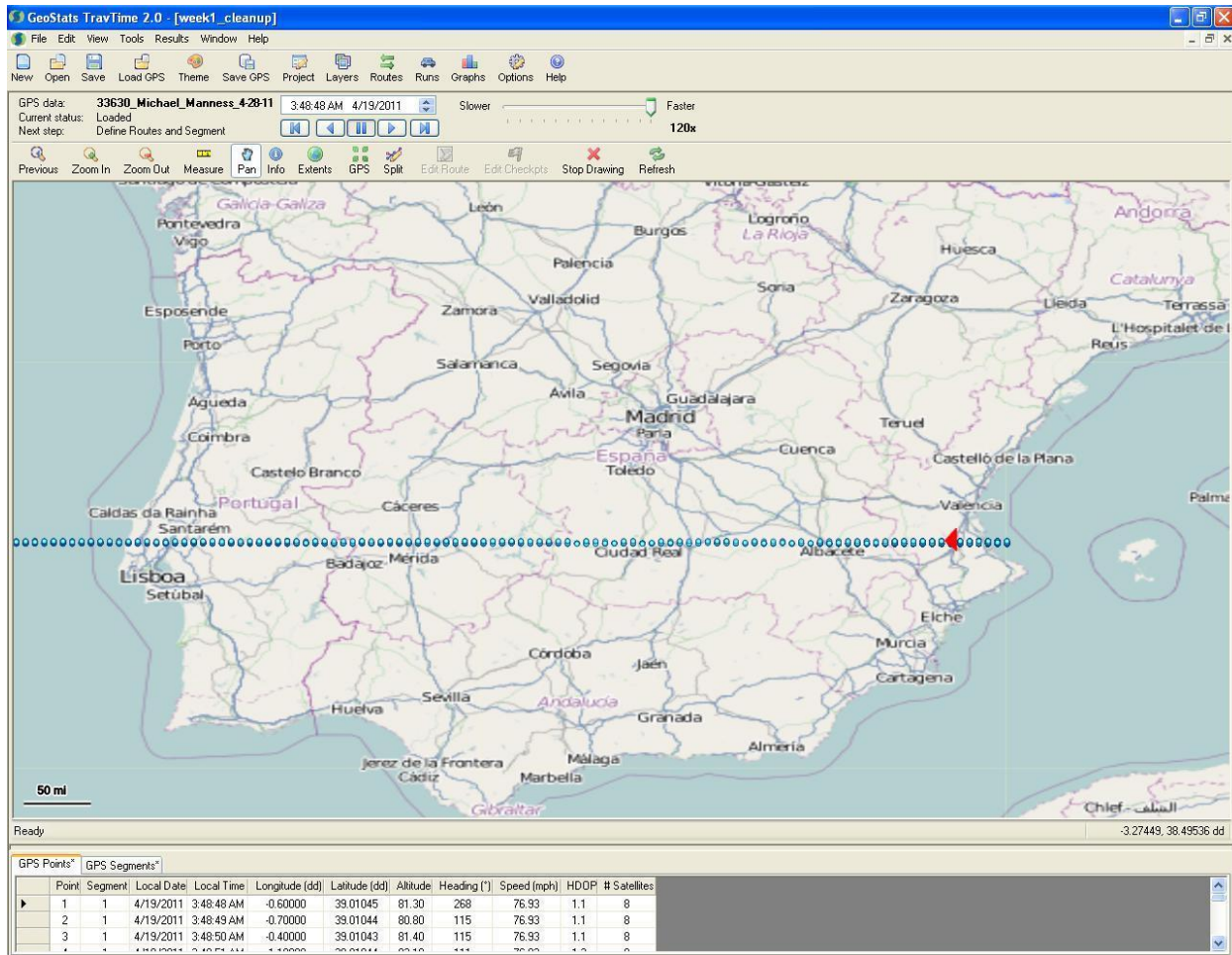


Figure 15: An example of data error.

For example, the data shown above is not valid as some of the trips occur over water. While this has been observed, the failure rate for these devices is about 5% and is bearable for the survey.

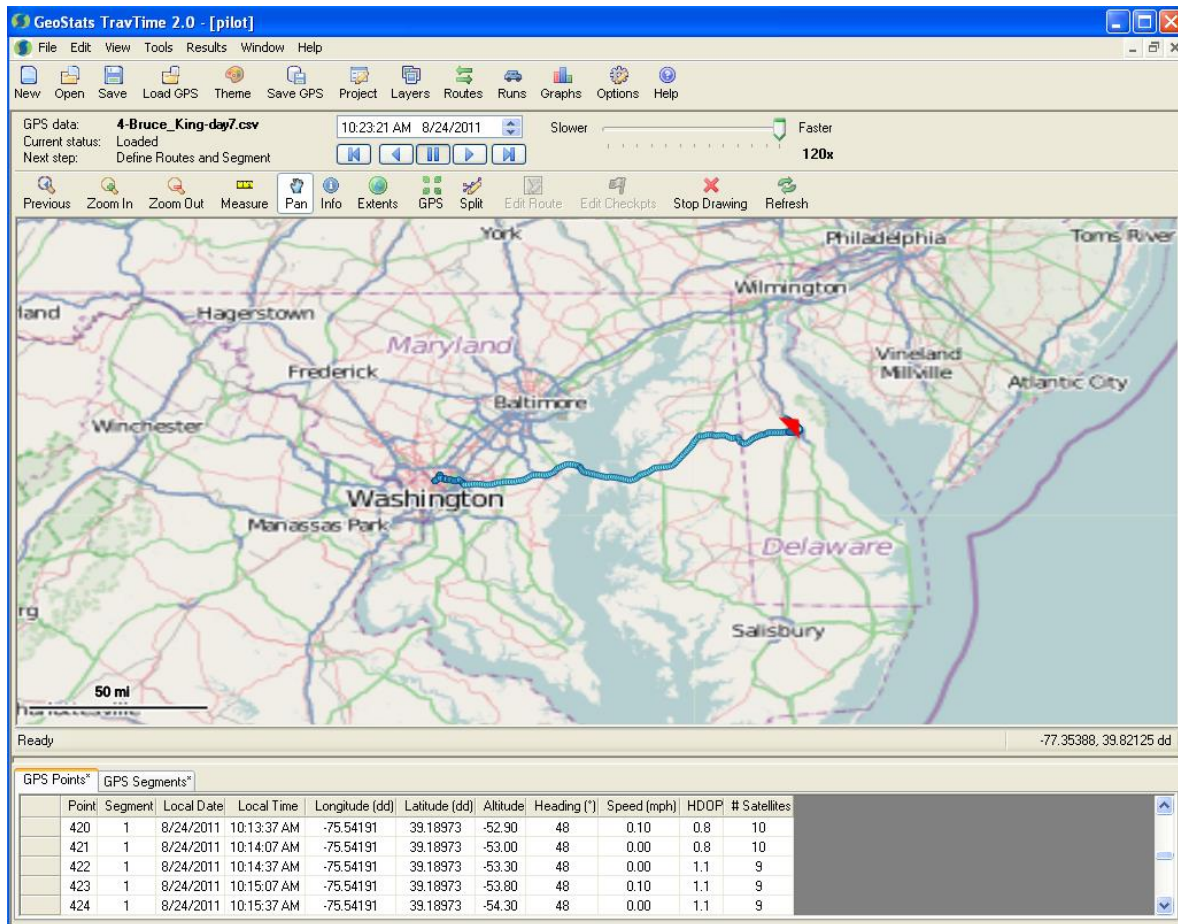


Figure 16: Correct data as recorded by the GPS device.

Above is an example of data when it correctly represents user location. Shown is a relatively long term trip from Maryland to Delaware. When zooming in (below), the exact route taken by the traveler is shown in greater detail.

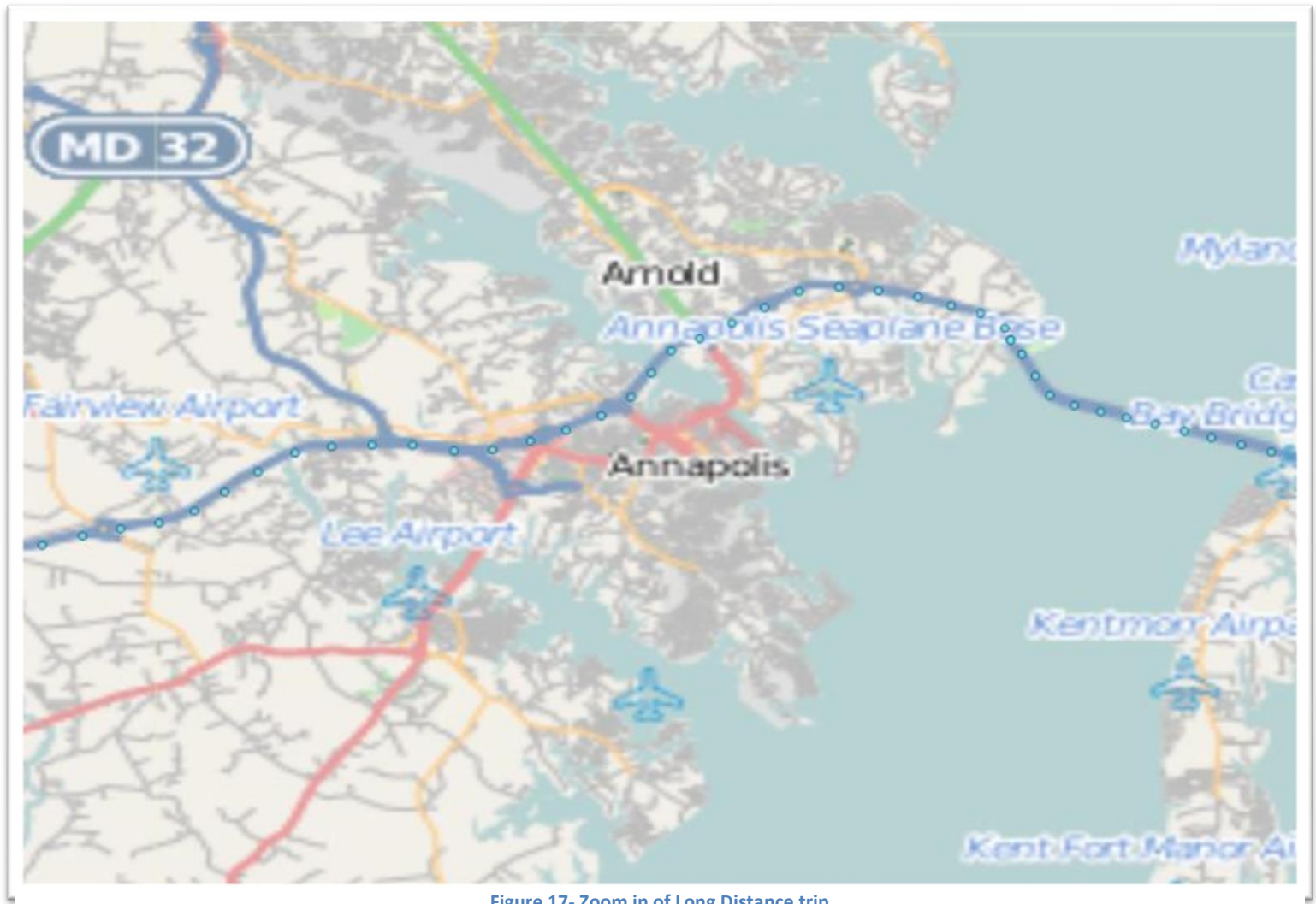


Figure 17- Zoom in of Long Distance trip

Once an individual dataset is checked for accuracy, it is saved to the operating computer and the server for backup. Checking the data accuracy resulted in 218 usable datasets for analysis, of the 263 original.

3.2 Open Street Map

The point of interest data that is used to derive trip purpose for each trip was collected and aggregated from the open crowd sourced data collection organization OpenStreetMap

(<http://www.openstreetmap.org/>). The data is converted from the original XML format into GIS

layer files for linking with the GPS data. The image below shows the intricate point of interest (blue points) and land use (light blue layer) data. In total there are roughly 26,000,000 points of interest in Maryland, DC, and Virginia (the locations where the majority of trips occur in the dataset).

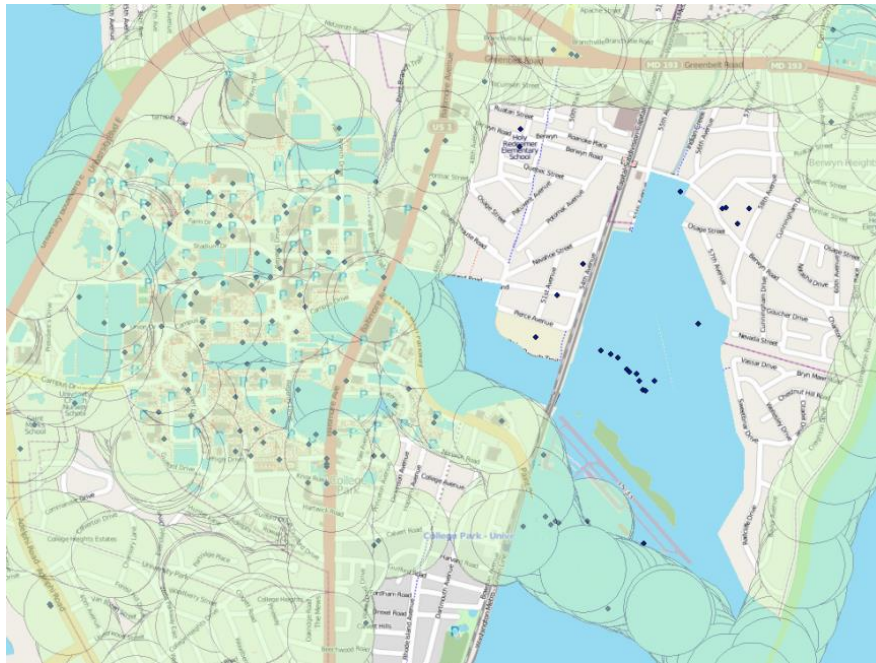


Figure 18- GPS point buffers and the spatial join to important POI data around the University of Maryland

The point of interest (POI) data is linked to the GPS data via a spatial join in the GIS environment. The distance threshold for POI to GPS linking is 300 meters. Due to the computational time for linking POI and GPS points, only the DC/Maryland/Virginia region link to POI points. The long distance trips included in this research do not have land use data allocated to them, thus making it difficult to accurately attribute trip purpose. This may lower the overall model accuracy, but due to the difficulty in identifying long distance destination locations, the overall impact was likely minor.

3.3 Data Processing

3.3.1. Data Excluded from the Analysis

The GPS dataset is an integral component to this research topic. The devices were checked for robust data, and compared to the online trip diary to insure viability. After these steps were complete, a usable dataset of 218 participants was established.

3.3.2 Scripting for GPS

For all data processing efforts, VBA code is developed to take the raw data into the format used for this research's in depth analysis. This is included in the Appendix. This includes VBA code to 1. Convert raw GPS points into a trip format giving basic variables on individual trips, 2. Create a unique O/D and route identifier for each trip that is unique to the user, 3. Convert latitude longitude distances into route distances.

The GPS dataset is collected at a recording interval of 1 minute per data point. The first step is to calculate the trip ends and hence origin destination points for each trip. A trip was determined as any series of points that had a velocity greater than 0 miles per hour, for more than three consecutive minutes. This definition of a trip is consistent with multiple previous studies (Wolf 2000). A number of trip statistics is calculated on the trip level, with travel time and distance amongst the most important for this study.

It is important to explain the way in which origins and destinations are derived and how they are clustered with the same identifier. First, all origins and destinations are defined as locations. Any GPS point is quantified as a location if the speed is 0 for a period of 3 minutes or longer. A 2 minute and 55 second stop to drop someone off, or to go through a fast food restaurant would not be caught and considered a stop, however, if the threshold were lowered, then locations would be designated at interactions or during traffic congestion. With each location designation, a trip is

determined as all points either with or without a velocity between the two location endpoints. It is important to include the trips for which there is no velocity because these are not necessarily stops (a GPS point with zero velocity) at locations, but possibly stop signs, traffic signals, etc. With each trip and origin/destination that is determined, the locations should be given some identifier so that similar locations can be grouped together. This way it will be easier to estimate the location when the locations can be grouped together. A destination is considered a unique destination if it is not within 300 meters of any other location for an individual user. Of the 20,651 locations used for this research, a total of 4,832 locations were not within 300 meters of any other trip end. The remainder of the trips is given location ID's that matched the rest of the trips that it lays within 300 meters of. If it lies within 300 meters of multiple points, it is given a matching id with the location that was visited most recently to that location. This identification number will be used to estimate locations and added to arrays which will be explained later. 300 meters was chosen as a distance for which a normal parking lot would encompass, yet is not so large that it would grab locations from neighboring destinations. This value can be changed at the discretion of the researcher. Later in the paper, the predictive accuracy of the model will be shown at differing distance thresholds.

3.3.3 Calculating Trip Characteristics

The distances from each origin/destination pair needs to be calculated in order for trip purpose rules to be derived. This is useful in deriving distance to home/work from the current location and setting trip purpose based on this distance. The Haversine formulation derives the distance from all origins to destinations, as well as all origins to home and to work.

Haversine Formula:

$$\text{Distance} = R \cdot 2 \cdot \text{atan2}(\sqrt{(\sin^2(\Delta\phi/2) + \cos\phi_1 \cdot \cos\phi_2 \cdot \sin^2(\Delta\lambda/2))}, \sqrt{1(\sin^2(\Delta\phi/2) + \cos\phi_1 \cdot \cos\phi_2 \cdot \sin^2(\Delta\lambda/2))}) \quad (7)$$

Where:

ϕ is latitude, λ is longitude, R is earth's radius (mean radius = 6,371km)

3.4 Data Conclusions

The end result of data collection is a database of over 20,000 trips, and nearly 5,000 locations.

Each string of GPS points signifying a trip is linked to the surrounding area's land use and point of interest data. From the Travel Survey filled out by participants, demographic information is also collected and linked to each trip in the GPS file. The next chapter will use the data collected for implementation in phase one of this dissertation's short term destination research: the Tiered Time- Origin Model.

CHAPTER 4: Tiered Time Origin Markov Model

4.1 Introduction:

Chapter 4 of this dissertation will explain the tiered model framework that has been developed to make predictions more quickly and be able to add trip purpose classification when it is available, and ignore it when it is not. This baseline model is compared to the current state of the art Markov model, and how improvements have been made to take advantage of multiple models.

4.2 Model:

The model is designed to make estimates of a trip's destination, and then learn about the trip after it has been taken. In this way, the next trip will have more information in five major categories that are used as reference points for the trip. These include: origin, time of day, day of week, user, and trip number.

4.2.1 Data Elements

- **Origin:** Each origin is given a numerical string that is used to associate location to other origins. If this location is within 300 meters (984 feet) of any other location, they will have the same numerical string.
- **Time of day:** The time when the car is started to perform a trip. (00,01,02,03,...,23) There are 24 elements possible in this array. If a trip is taken at 11:59pm, it is given the label of 23, whereas midnight would be indicated by 00.
- **Day of week:** The day of the week the trip is taken. (M,Tu,W,Th,F,Sa,Su) There are 7 elements possible in this array.
- **User:** Each participant is given a number to differentiate their travel patterns with others in the data arrays. (1-300)

- **Trip number:** Trips are counted to keep them in order and base future trip predictions on previous trip numbers. Some participants have over 500 trips, while others have fewer than 100.

Below is a graphical breakdown of how the model works, followed by an explanation of each step, concluded with a mathematical formulation.

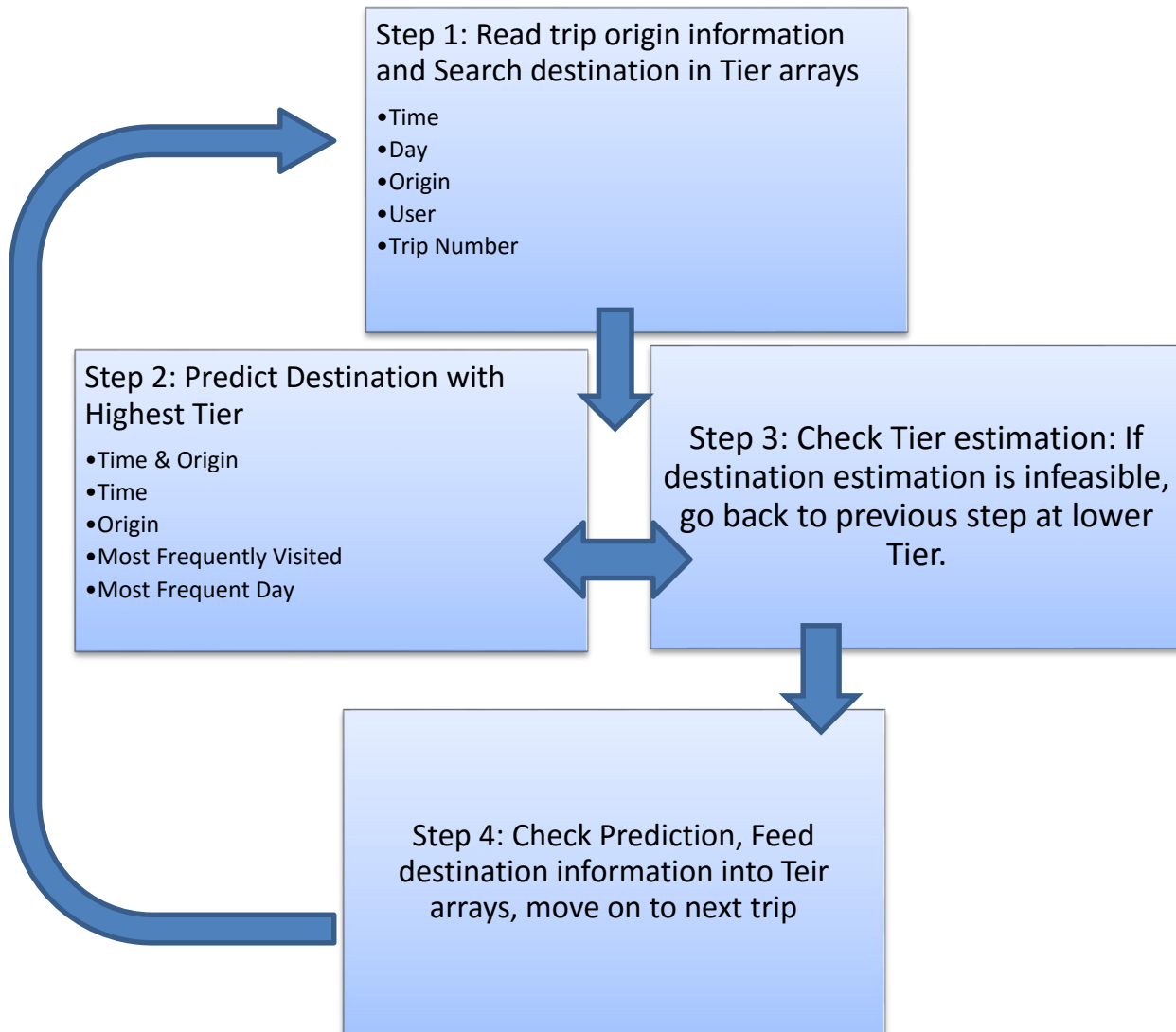


Figure 19- Graphical Representation of Tiered Time Origin Model

In Step 1, the prediction model reads the trip information from the user. This includes reading the geographic latitude/Longitude information of the origin of the trip, the identification tag of the user, and the start time. The location information is then assigned a location identification that corresponds to other's in the same area. The identification tag of the user is a simple number between 1 & 300. Time is broken down into hour segments to give 24 possible values for the

variable. Also, day of the week is collected, for 7 possible variable values. Once this data is read, the script moves onto step 2.

During step 2, the script searches through each tier to see what data is available to it to make a prediction. A tier is available if there is enough pre-existing information such that the probability of one destination yields a higher result than any of the others. Also, the destination chosen must exist at least twice in the tier array. The reasoning behind this is to make sure destinations aren't chosen just because the array lacks other information. It is better to move down to a lower tier than to select a destination based on limited information. After a tier is chosen, the script calculates the destination prediction, and then delivers the prediction result to the next step.

Step 3 performs a check on the calculation in two important ways: 1) if the destination prediction being made from the selected tier equation is equal to the origin of the trip, this location is deemed infeasible¹. 2) if the previous destination is the same as the currently predicted destination, the location is deemed infeasible. If either of these checks shows infeasibility, the script goes back to Step 2 at the next lowest tier. Once each tier has been exhausted, the script will provide the user's most common destination location as its estimation. If step 3 deems this prediction infeasible, step 2 will predict the 2nd most common destination, then 3rd, etc...

Figure 18 shows a fifth tier in step 2 of the model. While the model has this functionality, it has been removed due to low accuracy. Even though it requires the least amount of information and is almost always available, selection of destination based on day of the week has a very low accuracy.

¹ While it may be feasible that a trip originates from the destination location, it is rather rare. It is far more common for this type of prediction to be incorrect. Later in the paper we see that trip purpose type "Driving" is rather low, and this may be due to this rule's implementation. However, most driving occurs to drop off a passenger at a location, and not simply to drive for leisure. The author believes this rule increases the overall accuracy of the model, particularly in the first few trips of a participant when limited information is available to the model.

It has been kept in this dissertation for possible uses in future steps, even though it is not currently in use.

Step 4 is then performed to check the accuracy of the estimation then feed destination information into the arrays that are used in Step 1. A single trip's calculation has been performed and the script now moves onto the next trip starting at Step 1.

4.3 Formulation

This section will set forth the formulation for the Tiered Time Origin model that is the baseline for the additional trip purpose module. Each of the tiers is in itself a Markov model which derives probability for arrival at each visited location. Based on the data that is available to each tier, the algorithm is able to move throughout the tiers and creates the best solution for the case at hand.

4.3.1 Variables

U = set of all users in the survey

i = Trip number (i signifies the current trip)

P = trip purpose category (Home, Work, Other, Driving, Social/Recreation, Shopping, School/Daycare)

T = set of all time steps (24 total time steps)

V = set of all visited locations

l = location to be identified as visited location

4.3.2 Probability of location destination by tier

Each Tier probability is defined as the probability that a trip will end at a certain destination, given all the other information currently in the model. The probability for each visited location is reported to the next step in the algorithm for destination selection.

4.3.3 Mathematical Formulas

Time & Origin:

Highest probability visited location by user, time of day, and origin.

$$P^{T\&O} (v_i = l | u_i = u, t_i = T, v_{i-1} = l_k) = \frac{\sum \{v_r | v_r \in V, u_i \in U, t_i = t_k, v_{r-1} = l_k\}}{\sum \{v_r, u_i | v_r \in V, u_i \in U, t_i = t_k, v_{r-1} = l_k\}} \quad (8)$$

Where: $V = \{v_1, v_2, \dots, v_n\}$ is the set of all visited locations

$U = \{u_1, u_2, \dots, u_n\}$ is the set of all users

$T = \{t_1, t_2, \dots, t_{24}\}$ is the set of all time intervals

l_k is the previous location

v_r is the next visited location

u_i is the current user

Time:

Highest probability visited location by user and the time of day.

$$P^T (v_i = l | u_i = u, t_i = T) = \frac{\sum \{v_r | v_r \in V, u_i \in U, t_i = t_k\}}{\sum \{v_r, u_i | v_r \in V, u_i \in U, t_i \in T\}} \quad (9)$$

Where: $V = \{v_1, v_2, \dots, v_n\}$ is the set of all visited locations

$U = \{u_1, u_2, \dots, u_n\}$ is the set of all users

$T = \{t_1, t_2, \dots, t_{24}\}$ is the set of all time intervals

v_r is the next visited location

t_k is the previous time period

t_i is the current time period

Origin:

Highest probability visited location by user and the current location they are at.

$$P^o(v_i = l | u_i = u, v_{i-1} = l_k) = \frac{\sum\{v_r | v_r \in V, u_i \in U, v_r = l, v_{r-1} = l_k\}}{\sum\{v_r, u_i | v_r \in V, u_i \in U, v_{r-1} = l_k\}} \quad (10)$$

Where: $V = \{v_1, v_2, \dots, v_n\}$ is the set of all visited locations

$U = \{u_1, u_2, \dots, u_n\}$ is the set of all users

$T = \{t_1, t_2, \dots, t_{24}\}$ is the set of all time intervals

l_k is the previous location

v_r is the next visited location

u_i is the current user

Most Frequently Visited:

Highest probability visited location by user and the current location they are at.

$$P^{MFV}(v_i = l | u_i = u, v_{i-1} = l_k) = \frac{\sum\{v_r | v_r \in V, u_i \in U, v_r = l\}}{\sum\{v_r, u_i | v_r \in V, u_i \in U\}} \quad (11)$$

Where: $V = \{v_1, v_2, \dots, v_n\}$ is the set of all visited locations

$U = \{u_1, u_2, \dots, u_n\}$ is the set of all users

v_r is the next visited location

Most Frequent Day:

Highest probability visited location by user and day of the week.

$$P^D(v_i = l | u_i = u, d_i = D) = \frac{\sum\{v_r | v_r \in V, u_i \in U, d_i \in D\}}{\sum\{v_r, u_i | v_r \in V, u_i \in U, d_i \in D\}} \quad (12)$$

Where: $V = \{v_1, v_2, \dots, v_n\}$ is the set of all visited locations

$U = \{u_1, u_2, \dots, u_n\}$ is the set of all users

$D = \{d_1, d_2, \dots, d_7\}$ is the set of all days

d_i is the current day

4.4 Matrix size

The average participant visits 172 locations throughout the travel survey. Of those 172 instances of visiting a location, there are, on average, 78 known places visited two or more times per

person. These locations can be identified and revisited. If a probability matrix of the Bayesian network was created using each of the variables included in the Tiered Origin Model, the structure would be massively large filled almost entirely with zero. The size would come to: 78 Origins x 78 Destinations x 24 hours per day x 7 days per week x 7 purpose types (for future steps) = 7,154,784 cells. Gambs, Killijian, Cortez (2012) used a similar approach with only 3 locations (Home, Work, and Other) instead of the average of 172 locations found in this survey. The paper used origin, destination, and time of day (3 origins x 3 destinations x 24 hours per day = 216 cells), and had mixed accuracy results ranging from 65% to 77%. This research chose to use the new tiered Markov approach to maximize the predictions that are able to be made. The overall accuracy may be lower, but it allows us to make predictions in fine detail; predicting that a person is going to location 'Other' is not very helpful.

4.5 Array filling and algorithm example

This section will show an example of a user taking his first trips as a survey participant, and how the algorithm responds to his trip behavior.

The first trip taken by the user is a Monday morning trip to work. It is taken at 6:30 am, and there is no prediction made, because it is the user's first trip.

Array 1: Empty

User	Time	Day	Origin	Time&Day
-	-	-	-	-

Trip 1:

User	Day	Time	Origin	O-Lat	O-Long	Trip Number	Destination	D-Lat	D-Long	Location Prediction	Correct?
Tom Smith	Monday	0630	Home	38.99492	-76.9061	1	Work	38.97077	-76.9251	-	0

On the first trip that the user took, all of the available arrays are empty for the user. The state is completely unknown to the model, and has no basis for future trips. For this reason, the location prediction is empty, and a check is done after the prediction is made to see if it matches. There is no match between the Location Prediction column and the Destination Column (the destination and destination lat and long column cannot be seen until after the prediction is made), so a 0 is placed in the "Correct?" column. Now that the prediction has been made, the arrays are filled for the next trip.

Location Identifier:

Location ID	Lat	Long
1	38.994922	-76.906102
2	38.970771	-76.925089

End of Iteration 1.

Array 2:

User	Time	Day	Origin	Time & Origin
Tom Smith- 2	0600 - (2)	Monday (2)	Home (1) - Work (2)	0600 (1) - (2)

The first value in each array is the searching variable, whereas the second value is the array output.

Now some array information has been added to each of the 5 arrays. Based on the structure of the model, the algorithm will now search from the top most tier for the information that matches the current trip. Below is the second trip that is taken.

Trip 2:

User	Day	Time	Origin	O-Lat	O-Long	Trip Number	Destination	D-Lat	D-Long	Location Prediction	Correct?
Tom Smith	Monday	0630	Home (1)	38.99492	76.9061	1	Work (2)	38.97077	76.9251	-	0
Tom Smith	Monday	1200	Work (2)	38.97077	76.9251	2	Restaurant (3)	38.99375	76.9055	Work (2)	0

Tier 1 estimation: Time & Origin

The algorithm searches in the Tier 1 array: Time & Origin, for the time 1200, and origin (1), since there is no information that matches the time and origin, the next step is to move onto time & day. Again, the time is 1200, and the day is Monday, since both do not match,

the next tier is taken, which is array type time. There is no array information that starts at time 1200, so the algorithm moves to array: origin. The origin location identifier is (2), which is not found in the origin array, so the final step is to move to the lowest tier array and search by User. The user matches up because it is the same person, so the algorithm estimates that the user is going to location 2 even though the user has just come from location 2. The estimation is incorrect, and the algorithm moves onto the next step.

End of Iteration 2.

Array 3:

User	Time	Day	Origin	Time & Origin
Tom Smith- 2	0600 - (2)	Monday (2)	Home (1) - Work (2)	0600 (1) - (2)
Tom Smith- 3	1200 - (3)	Monday (3)	Work (2) - Restaurant (3)	1200 (2) - (3)

The information from the previous trip is now available to the user. The trip from work to the restaurant can now be looked up for trip

3.

Trip 3:

User	Day	Time	Origin	O-Lat	O-Long	Trip Number	Destination	D-Lat	D-Long	Location Prediction	Correct?
Tom Smith	Monday	0630	Home (1)	38.99492	- 76.9061	1	Work (2)	38.97077	- 76.9251	-	0
Tom Smith	Monday	1200	Work (2)	38.97077	- 76.9251	2	Restaurant (3)	38.99375	- 76.9055	Work (2)	0
Tom Smith	Monday	1400	Restaurant (3)	38.99375	- 76.9055	3	Home (1)	38.99492	- 76.9061	Work (2)	0

The algorithm does the same searches as before, time& origin produces no findings, time& day produces no findings, and Origin, and Time produces no findings. Again, the user category is used, and since there is no most likely (most common location), the first value is used (2). Again, the estimation is incorrect. Now moving onto the next trip, the additional array information begins to pay off. Using this method it is easy to see why the accuracy before the fifth trip is below 40%.

Array at start of trip 4:

User	Time	Day	Origin	Time & Origin
Tom Smith- 2	0600 - (2)	Monday (2)	Home (1) - Work (2)	0600 (1) - (2)
Tom Smith- 3	1200 - (3)	Monday (3)	Work (2) - Restaurant (3)	1200 (2) - (3)
Tom Smith- 3	1400 - (1)	Monday (1)	Restaurant (3) Home (1)	1400 (3) - (1)

Trip 4:

User	Day	Time	Origin	O-Lat	O-Long	Trip Number	Destination	D-Lat	D-Long	Location Prediction	Correct?
Tom Smith	Monday	0630	Home (1)	38.99492	-76.9061	1	Work (2)	38.97077	-76.9251	-	0
Tom Smith	Monday	1200	Work (2)	38.97077	-76.9251	2	Restaurant (3)	38.99375	-76.9055	Work (2)	0
Tom Smith	Monday	1400	Restaurant (3)	38.99375	-76.9055	3	Home (1)	38.99492	-76.9061	Work (2)	0
Tom Smith	Tuesday	0637	Home (1)	38.99492	-76.9061	4	Work (2)	38.97077	-76.9251	Work(2)	1

This is the first correct prediction made. The next morning, the user drives to work at roughly the same time of day. So, after the algorithm searches the top tier: Time& Origin, the array shows 0600 (1) – (2). The time is at the 6am hour from origin Home (1), to destination (2). This estimation is made and the next trip is taken.

This process is repeated for the duration of the travel survey. The higher tier approach is taken first, if there is not a higher probability destination, the next tier array is studied. While this example has only shown the first 4 trips taken by the user, the algorithm is able to learn significantly more information, as the average user takes 172 trips during the survey period. The next section will show accuracy of the model in terms of Tiers.

4.6 Results:

Tier 2: Time of Day and Origin

This is the most accurate of all the prediction methods not including the derived purpose of the trip. If there has been a previous trip by the same user from the same origin, at the same hour of the day, then given the most common of the destination outcome will be correct 59.5% of the time.

Tier 3: Time of Day

The accuracy surprisingly only decreases to 56% for this prediction approach. The additional origin information is not overwhelmingly useful, as these results tend to show us that if a person tends to go to the same location at the same time of day, the origin of the trip is not of great importance.

Tier 4: Origin

Tier 3 is the first tier estimation procedure that drops below the average overall accuracy of 47%, at 36%. This tier makes the bulk of predictions at over 10,000 instances.

Tier 5: Most Common Location

For nearly all of the trips a 'most common location' prediction occurs. For each individual, this estimation becomes their home location, since this is where the user spends most of their time. However, if the previous location was their home location, this estimation will turn into their second most common location. This tier is, as expected, the lowest accuracy of the prediction methods at 30%.

Location by differing accuracy definitions:

In the previous section results are given at the 300 meter threshold; if the estimated location is less than 300 meters from the actual trip end, then the estimation is deemed correct. In this section, the

differing levels of accuracy by thresholds will be shown. Tables 1 and 2 shows the accuracy by 100 meters, 200 meters, 300 meters, 500 meters, 750 meters, 1km, and 2 km's.

With the differing accuracy thresholds, the accuracy changes from about 42% at 100 meters, to about 49% at 1000 meters. While the threshold can continue to be raised to increase the perceived accuracy of the algorithm, many of the incorrect estimations are not simply close locations to incorrect guesses, but rather completely incorrect estimations from many kilometers away (such as estimating 'work' when the trip is going to 'school' 2 km away).

True accuracy:

Due to the nature of location destination prediction, it is not possible to make an estimation of a location until that location has been previously visited. If you combined these occurrences with all the occurrences for which it is the first time that the user goes to that location, the true accuracy of the model is actually much better. Of the 20,651 trips only 15,819 are estimable by this, or any true prediction method. Below, the graphs are broken down into results with previously visited and unique location destinations and differing accuracy definitions.

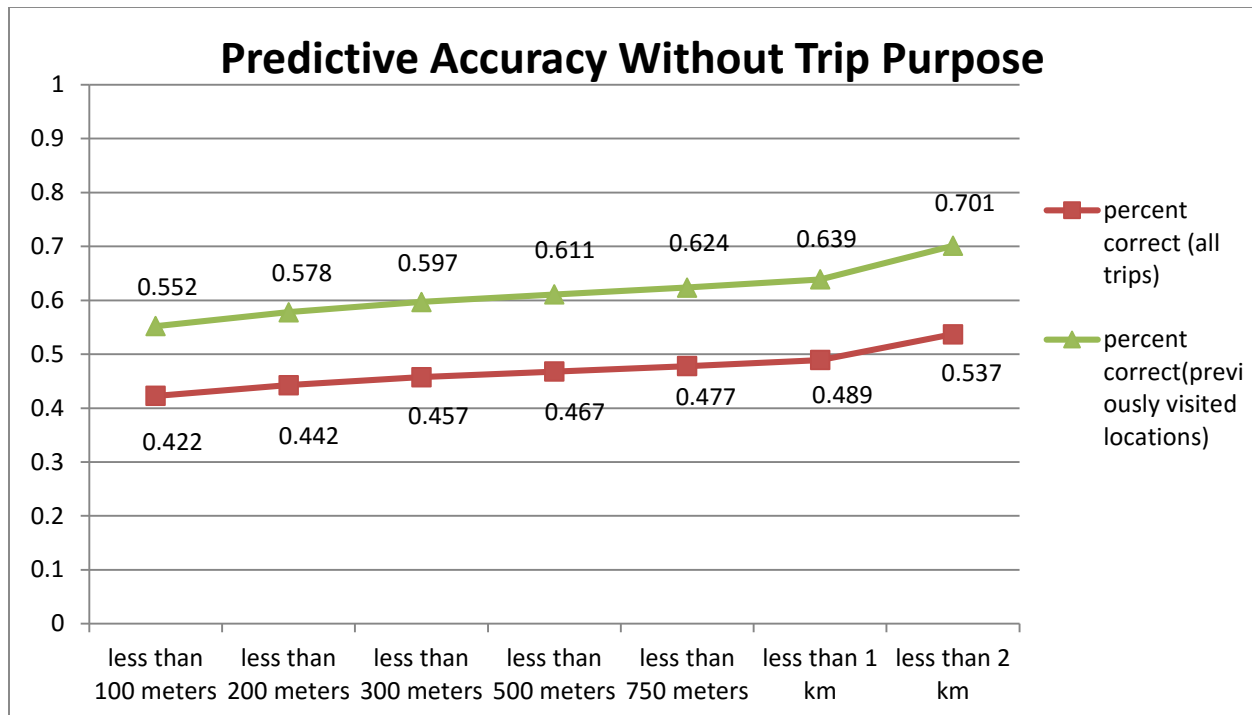


Figure 20- Model Accuracy by Distance threshold

4.7 Tiered Time-Origin Model Conclusions

While the estimations can be made much faster than previous models (Gao 2012, requires 315 days of learning before model output), the resulting estimations are not as accurate. The location classification needs to be set to 2 km in order for the accuracy of the Tiered Time Origin Model to exceed the best currently available. Even though the location clustering algorithm in the Gao 2012 is not explored in depth, it is likely lower than 2 km. What is achieved through this chapter is not great accuracy, instead a structure that can employ high accuracy modules like those shown in the next two chapters. The Tiered Time Origin model will show value in its versatility.

CHAPTER 5: Markov Model with Future Trip Purpose Information

5.1 Introduction

This chapter explains how the addition of land use data is added to create an effective trip purpose estimation that will increase the overall accuracy of destination prediction. The methodology for deriving the purpose for each individual trip will be explored followed by the inclusion of the trip purpose module into the existing Tiered Time-Origin Model framework. Finally, results will show a great improvement in the accuracy as a whole.

5.2 Methodology

WEKA (Waikato Environment for Knowledge Analysis) machine learning software was used for this portion of the research. The rule set for designated trip purpose that was used (Lu 2013) was altered to take advantage of land use characteristics available in OpenStreetMap; the land use characteristics from OpenStreetMap were matched to those in the Lu et al. dataset and run through the decision tree system. For example, the Lu dataset includes land use types Single Family and Multifamily, whereas in OpenStreetMap, types house, dormitory, hotel, farm, and apartment are recorded. While these small differences exist, the same matching methodology and end trip purposes is used.

The accuracy of the approach cannot be tested because each trip's true purpose is not collected during the survey. It is believed that the accuracy approaches 80%, but since some land use characteristics have changed and the data was collected in different regions of the U.S. (Midwest versus east coast), there could be a small decrease in accuracy. For the full list of land use

variables in this step, please see the appendix (there are 26 categories totaling well over 500 variables).

It is important to note this trip purpose uses advanced information that could not be known to the system at the time. While the rules derived in previous literature are useful, for realistic implementation into the algorithm, the future destination land use information must be removed and new rule sets derived. This will be done in chapter 6.

5.3 Trip Purpose Determination

The structure for determining each trip's purpose is below:

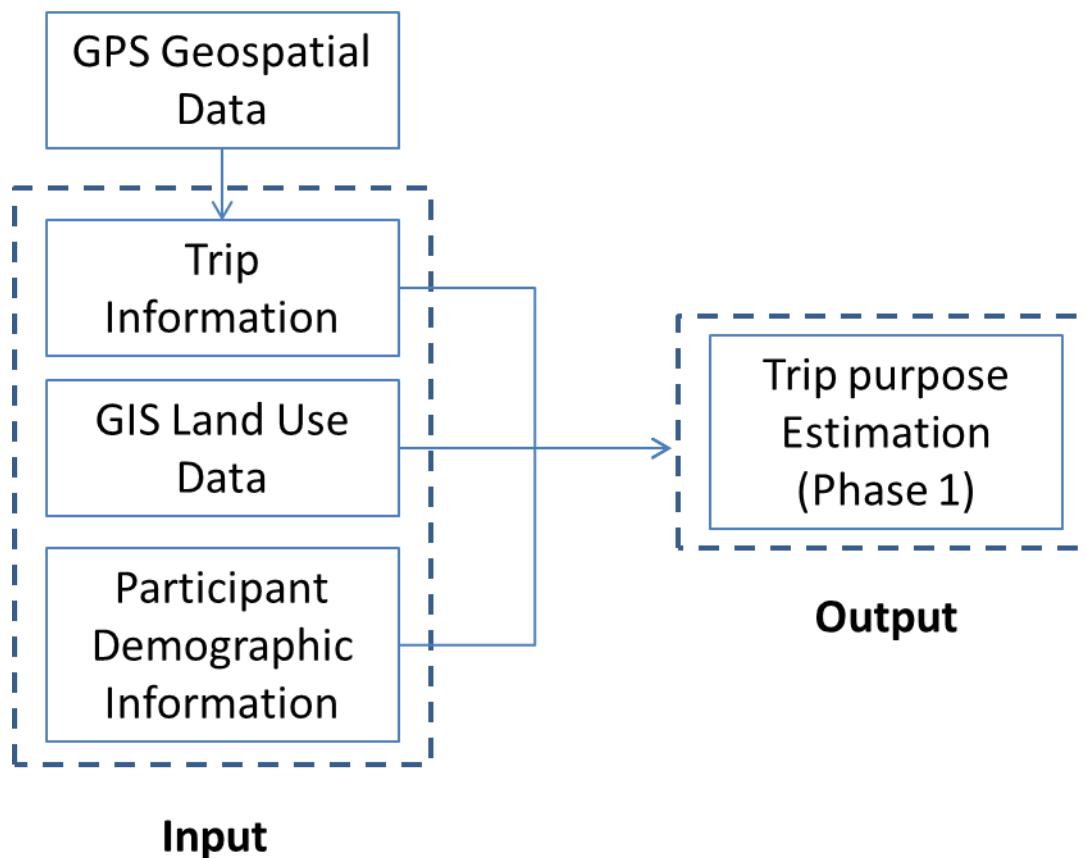


Figure 21- Trip Purpose Estimation based on GPS Data, GIS Land Use and Participant Demographics

The exact inputs for the model are important as well. Below is the breakdown of what constitutes Trip Information, GIS Land Use Data, and Participant Demographic Information.

Table 2- Input Variables for Trip Purpose Estimation (phase 1)

Category	Variables
GPS Geospatial Data	<ul style="list-style-type: none"> • Trip Start Time • Trip Duration Time • Day of Week • Previous Trip End Time • Soak Time • Previous Activity Duration
Participant Demographic Information	<ul style="list-style-type: none"> • Income • Education Level • Age
GIS Land Use Data	<ul style="list-style-type: none"> • Land Use Type of Current location (start of trip) • Land Use Type of Destination (end of trip) • Trip Destination Type (Home, Work, or Other)

Table 3- Participant Income Range

Participant Trip Data: The 500 plus variables are categorized into the location types below.

- Business Unit
- Commercial
- GOV/PUB/Service
- Residential
- Restaurant
- Mixed_Use
- Institutional
- Industrial

- Leisure
- Recreation Site
- Shops

As with trip purpose detection, the algorithm uses a 300m threshold from the GPS location to estimate the trip origin location. In the event that multiple land use types are picked up by the matching process, the nearest location is used. Not all trip locations could be estimated because either the location has not been identified yet in OpenStreetMap (such as a new building) or the system did not recognize any nearby location (no location existing within 300 meters).

The next section will take the newly derived trip purpose from this approach and use it as a new variable in the Tiered Time Origin Model to create the first Trip Purpose based destination prediction.

5.3.1 Trip Purpose Model Formulation

Highest probability visited location by Trip purpose classification

$$P^p(v_i = l | u_i = u, p_i = p_k) = \frac{\sum\{v_r | v_r \in V, u_i \in U, p_i = p_k\}}{\sum\{v_r, u_i | v_r \in V, u_i \in U, p_i = P\}} \quad (13)$$

Where: $V = \{v_1, v_2, \dots, v_n\}$ is the set of all visited locations

$U = \{u_1, u_2, \dots, u_n\}$ is the set of all users

$P = \{p_1, p_2, \dots, p_7\}$ is the set of all purposes (Home, Work, Other, Driving, School, Social, Shopping)

5.4 Results:

Similar to the previous section, the results are given by differing distance thresholds and in two categories: all trips, and only previously visited locations. First, the accuracy for the individual tier is explored.

5.4.1 Tier 1: Trip Purpose

Trip purpose information has not been previously used in short term destination prediction. This may be due to the fact that it is not included in raw GPS data, and the derivation of trip purpose requires a working knowledge of travel surveys. The implementation of trip purpose has the largest impact on the accuracy of the model as a whole. With estimations made on 15,336 trips, over 75% of all trips can be estimated in this approach. It is particularly accurate at determining home and work locations due to the designation of these trip purpose types. Once a single location has been identified as home based purpose, the algorithm estimates the end location at 94% accuracy, with the correct estimation of work based purpose at 89% accuracy. Other trip purposes lower the accuracy to an average of 76%.

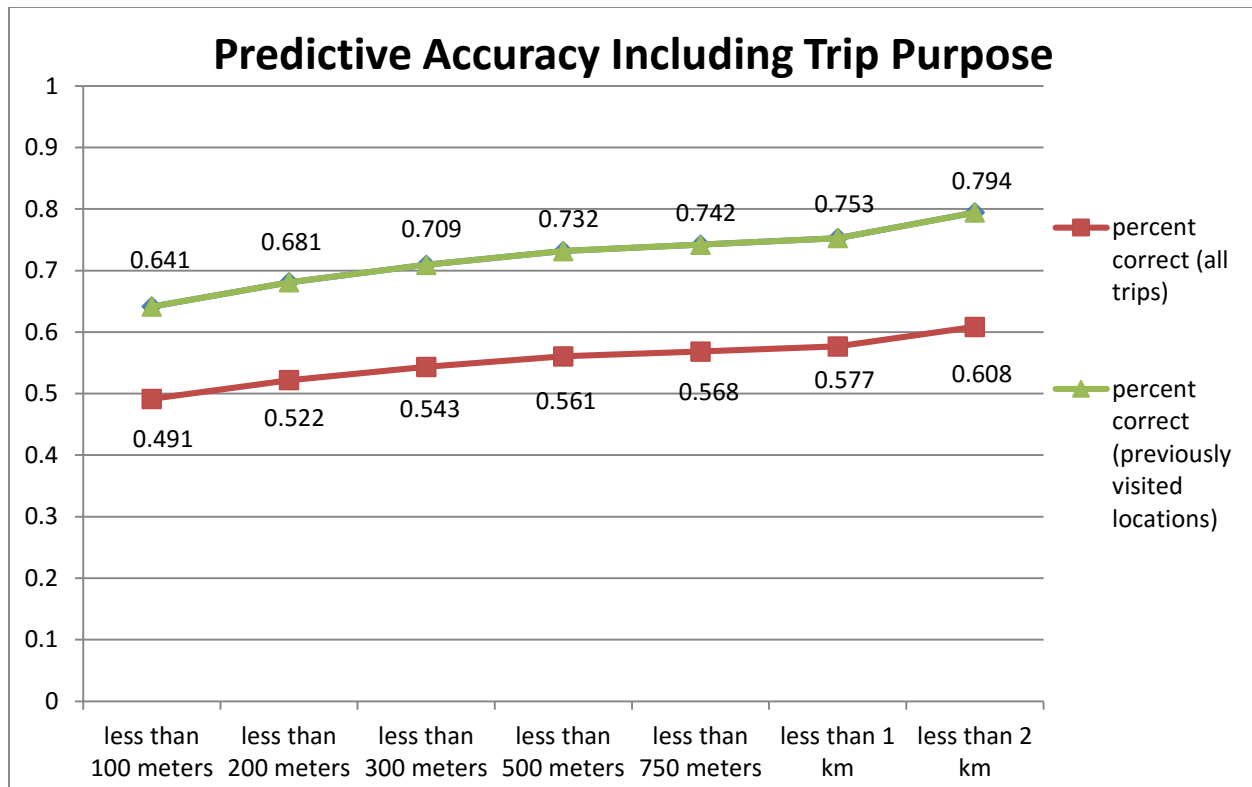


Figure 22-Model Accuracy by distance threshold & location availability

The above graph shows a major improvement in accuracy over the previous chapter (9%). Most importantly, it shows an improvement over the best model found in the literature review (6%). There is clearly an advantage in using trip purpose to predict end trip location.

The accuracy of the model is also heavily dependent on the individual for which it is predicting. Different users have a more stable and easy to predict travel pattern. Below is the graph of the algorithm's prediction accuracy for different users. The accuracy ranges from 11% to 82%. One could imagine an adaption to the algorithm based on the nature of the user's travel pattern. People with a steady job for which they travel on a regular basis tend to travel to their work location more regularly. The algorithm could weight work predictions more heavily. On the contrary, for an individual who is unemployed the algorithm could clear its arrays frequently to better adapt to

changing travel patterns. The opportunities for improvement depend on the understanding of underlying travel behavior and participant characteristics.

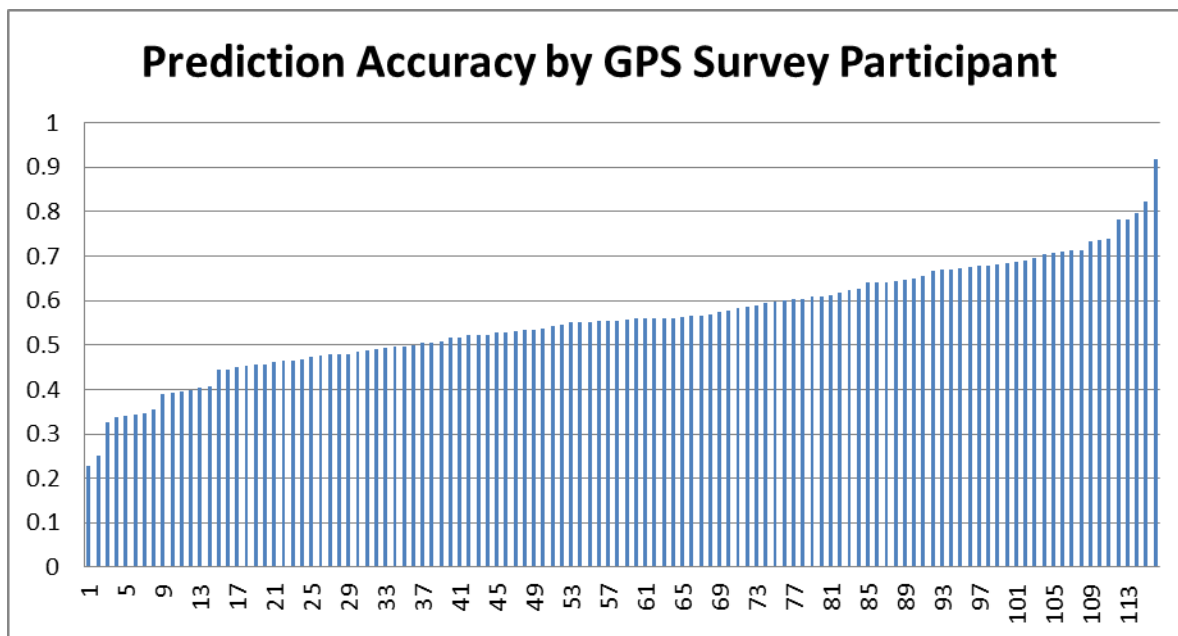


Figure 23- Percent of accurately identified trip destinations by GPS participant

5.4.2 Accuracy as algorithm learns trip behavior

The reason for this hierarchical type model is that while the higher tiers are more accurate predictors, they are also less likely to have information set into the arrays already. As the model goes into lower Tiers, the accuracy decreases, but it is able to at least make some sort of prediction. As seen in the graph below, as more trips are made, more information is fed into the arrays and higher tiers are able to be utilized. The longer the survey goes, and the more trips the users take, the accuracy increases.

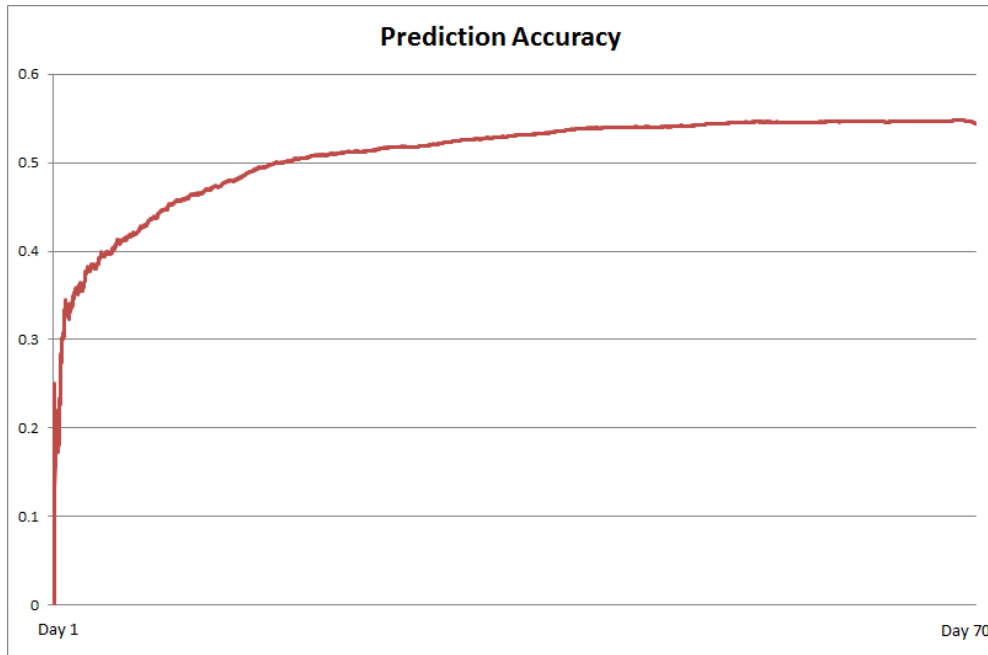


Figure 24- Percent of correctly estimated trip locations as Survey continues

The prediction accuracy starts out at 0% then over time, the model increases to an accuracy of 56.1%. The algorithm is learning from past experiences and is able to predict future trips as trips are made. The model has no back knowledge on day one and cannot make a reasonable estimation as to where the trip will end. After this first trip however, there is at least one destination in the knowledge base of the user. The algorithm then has very slightly more knowledge as before. Due to this fast learning approach, by about the third day, the algorithm has probably learned the home and work location for each user, and can therefore have reasonable estimations (with 37% percent accuracy). After this initial learning phase, the algorithm gains accuracy slowly as time goes on and trips occur. The model does not reach its maximum predictive power until day 70 when it reaches its peak accuracy of 56.1% for all trips (previously visited locations and non-previously visited).

5.4.3 Accuracy by day of week

Another interesting result from this research is the predictive accuracy by day of the week. It's clear that during the week, it is easier to predict an individual's location due to a more scheduled day.

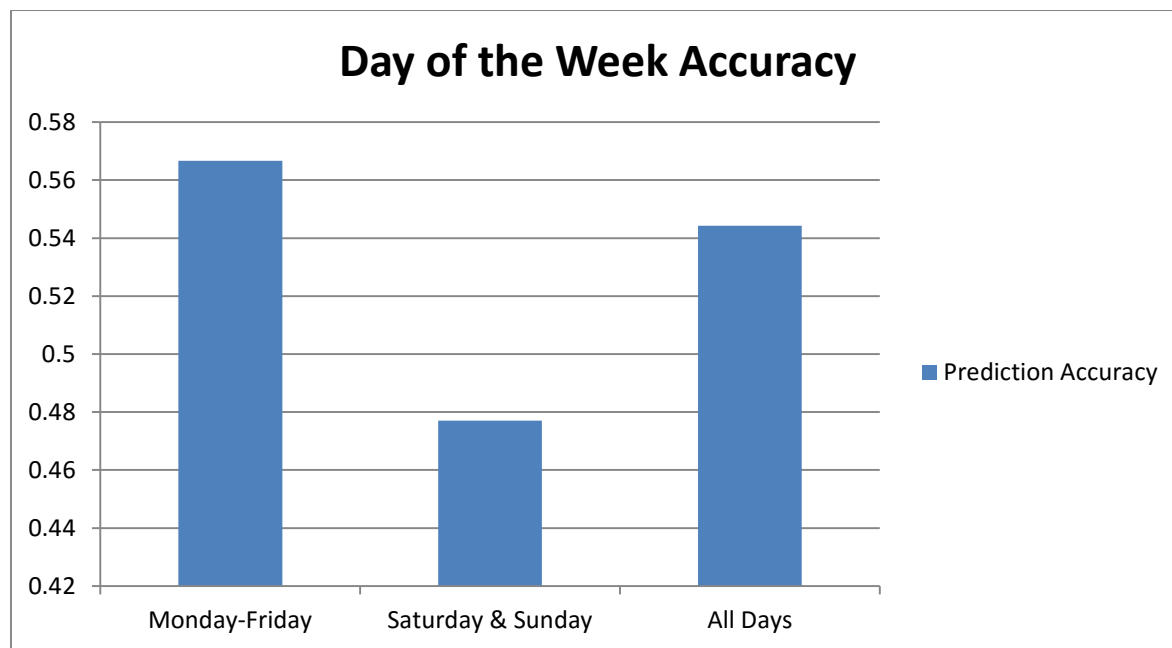


Figure 25- Accuracy by day of the week

This table includes all trips at an accuracy level of 300 meters. To get an idea of why it is much easier to predict the trips that occur during the week, it is easiest to look at the unique trips. While Saturday and Sunday make up only 28% of the days of the week, they account for 33% of the unique trip locations. On the weekend people tend to travel to locations that they have never been before. Saturday has 29% of all its trips going to unique locations, followed by Sunday at 24%, then 21 or 22% for the remainder of the week. The large discrepancy is due to a standard work week. Seeing as work trips are the easiest to estimate, it further adds to the accuracy difference. In the dataset used for this work, 92% of work trips were taken from Monday to Friday.

5.5 Applications

After viewing the results of the algorithm, possible applications are made clearer. With a model accuracy of 71% for previously visited locations at accuracy of 300 meters, route choice models and shortest paths are easy to derive. These estimations can be sent to centralized locations to optimize the transportation system via intelligent intersections by shifting light phases for upcoming volume changes. Congestion pricing becomes more feasible if the network is aware of a higher volume of users as they approach the roadway, not after they have already gotten on the system. As a whole, this type of location destination prediction leads to a smarter system that would be able to adapt to future changes in the network.

So far, only the public benefit to users via changes to the network has been explained in reference to possible applications of this research; however, one can imagine the immense benefit to private companies and personal users as well. If a person's regular vehicle had estimation for your destination as you got into your vehicle and started a trip, the device would be able to tell you information about the upcoming congestion towards the location. The device would not only be able to make suggestions for places of interest around your area, but also where it believes your trip will end, and places along the way. All of this can be done with roughly 71% accuracy without the user needing to interact with the GPS units at all. The user does not need to type in destination location information. They would only need to get in their vehicle, start driving, and receive information pertinent to the driver's destination. This is particularly useful to current GPS users who do not want to hassle to put in destination location information for places they know how to get to, such as home and work locations. The algorithm can learn your home and work locations after only a few days of normal travel, then give the user real time knowledge when it believes that this trips are being taken. This can lead to safer roads for those users who type in the destination of their trip while driving, or a faster commute for those who take the time to manually enter a

destination at the start of their trip. The applications for smarter personal GPS units are evident and exciting for users and developers.

5.6 Conclusions:

The prediction model is able to accurately predict 56% of a user's destination location before they start their travel, 71% when considering only previously visited locations. For some users the accuracy is as high as 91% while other users' accuracy is as low as 21%. The accuracy during the weekdays is also higher as compared to weekends due to the predictable routine of many users. The addition of trip purpose increases the overall accuracy from 45 to 56% for all trips, and from 60 to 71% for trip locations that have been previously visited. While these results are promising, they include end of trip land use data, which could not be implemented in real world application of this system. To move forward with this approach, further techniques are applied in Chapter 6 to accurately predict trip purpose without destination land use data.

CHAPTER 6: Markov Model with Derived Trip Purpose and Training Sets

6.1 Introduction

The significant contribution of this chapter is the inclusion of the first trip purpose destination prediction algorithm that does not rely on end of trip point of interest data. The algorithm is able to estimate where the participant travels for a particular purpose using only the land use characteristics from the trip origin. This is done by using the previous chapter's trip purpose and deriving a new trip purpose based on a decision tree and a certain number of trip's learning sets (5, 15, and 30 trips). The model is able to learn user trip behavior and make more accurate predictions than previous research efforts.

6.2 Markov Model (Baseline)

The model estimates the trip's destination then learns about the trip after it has been undertaken. The next trip will have more information in three major categories that are used as reference points for the trip. These include: origin, time of day, day of week, and trip purpose information/classification (used only for the trip purpose module). Below is a graphical breakdown of how the model works, followed by an explanation of each step.

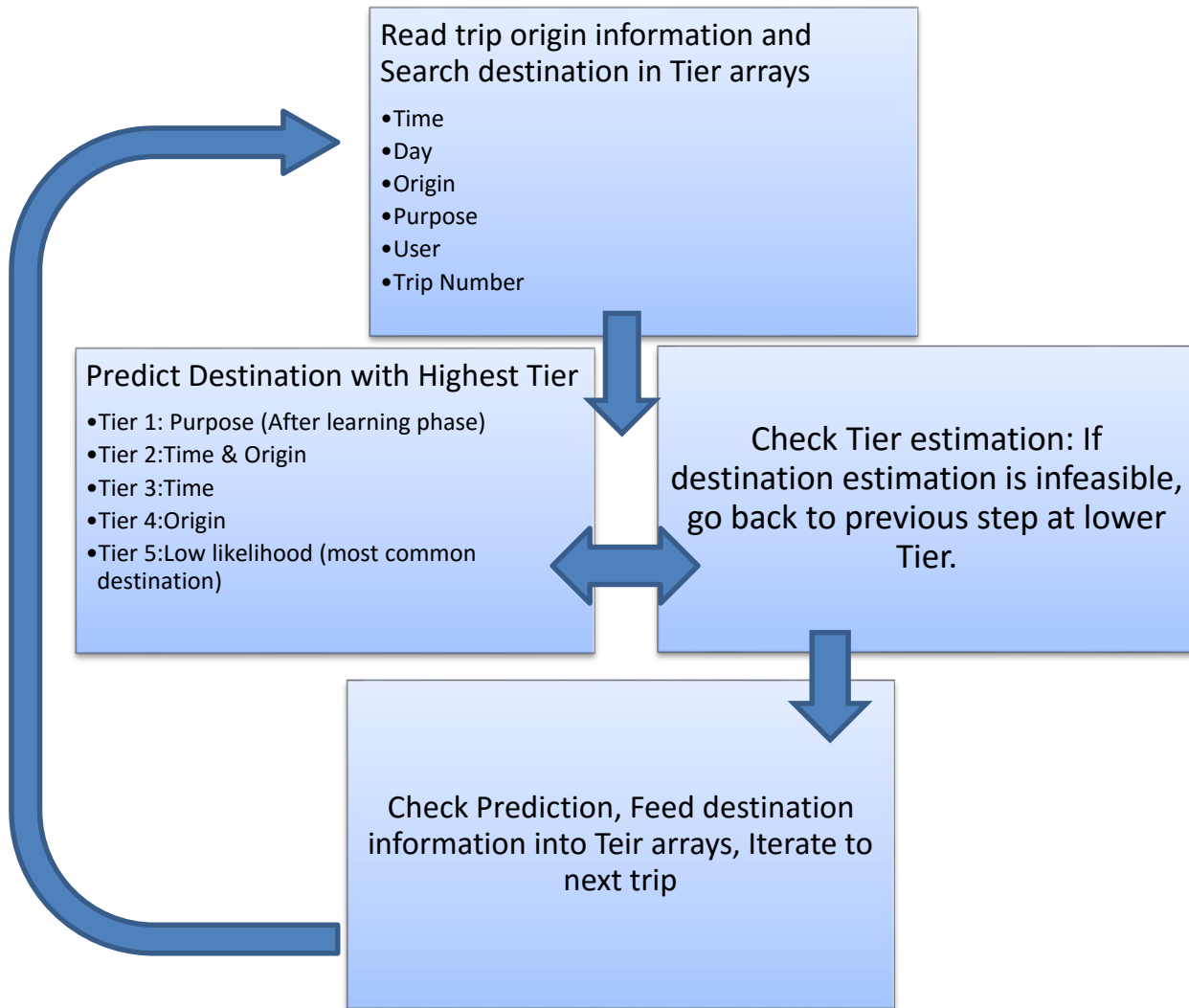


Figure 26- Graphical Representation of Tiered Trip Purpose Model

The prediction model starts out by reading the trip information from the user. This includes the latitude longitude information of the origin of the trip, the identification tag of the user, and at what time the trip has started. Only the starting information of the trips is used. The entire process of destination prediction is done before the vehicle is in motion. The script searches for previous trips like this one in arrays for time, origin, user, demographic information, etc... Based on what the script finds in these arrays, it will then make predictions. If this is the first trip of the GPS survey for the user, then there will be no information saved up in the arrays, as information is only loaded

into each array after the trip has occurred. With its previously learned information, the model starts at Tier 1. A search is made for purpose information from the same user, and pulls the most common destination location that occurred from that trip purpose. Since this has the highest accuracy of all the prediction methods, it is done first. The algorithm selects the most likely destination based on percentage of trips whose trip ended at that destination. If no estimation can be made, either due to lack of learned information from previous trips, or no one destination has a higher likelihood of occurring than any other destination, the model moves onto the next tier.

Tiers 2 through 5 work identically to previous chapters, but are used less due to Tier 1's availability. At no point can information from future trips be used, and past trip estimations are not changed after the algorithm has moved onto the next step.

6.3 Trip Purpose

The literature review shows that vehicle destination prediction is a relatively new field that is advancing quickly due to the availability of accurate moving point data via GPS loggers, actively transmitting GPS systems, and smart phone applications. Modeling has become increasingly accurate in regards to correct estimations from the mid-point to the end of the trip. Also, the estimation of trip purpose has been a well-established field for many years. But, the only applications of trip purpose in the destination prediction field are by applying a purpose after the estimation is made. By applying trip purpose estimation before destination is predicted, a more accurate destination location can be made by giving a set of possible destinations that would achieve the same trip purpose. Trip purpose can be useful in a few ways:

If the user tends to take a certain trip purpose at a certain time, or from a certain origin, the model will be able to narrow down the possible alternative locations. Searching for locations by purpose

yields better results than searching by time, day, or trajectory. For example, if a person goes shopping every Saturday morning, the subset of available destination location estimations should be only those whose land uses type support shopping. If a person deviates from their route, it would be most beneficial to first search other shopping locations instead of choosing the next most probably location.

For each participant's first 15 trips, the trip characteristics are recorded into a separate database for machine learning. Once the first 15 trips are taken, a pre-established rule system is run to estimate the purpose for each of those trips (Chapter 5). This trip purpose is based on an end of trip location, which must be removed. Using the first 15 trips and a J48 machine learning algorithm, a rule set is created that determines trip purpose using only start of trip information. The final decision tree is shown in the appendix, along with model specifications. This rule set is applied to all future trips. The approach derives a trip purpose estimation based on previous trips and requires no information after the start of the trip. A step by step explanation is shown in the remainder of the section.

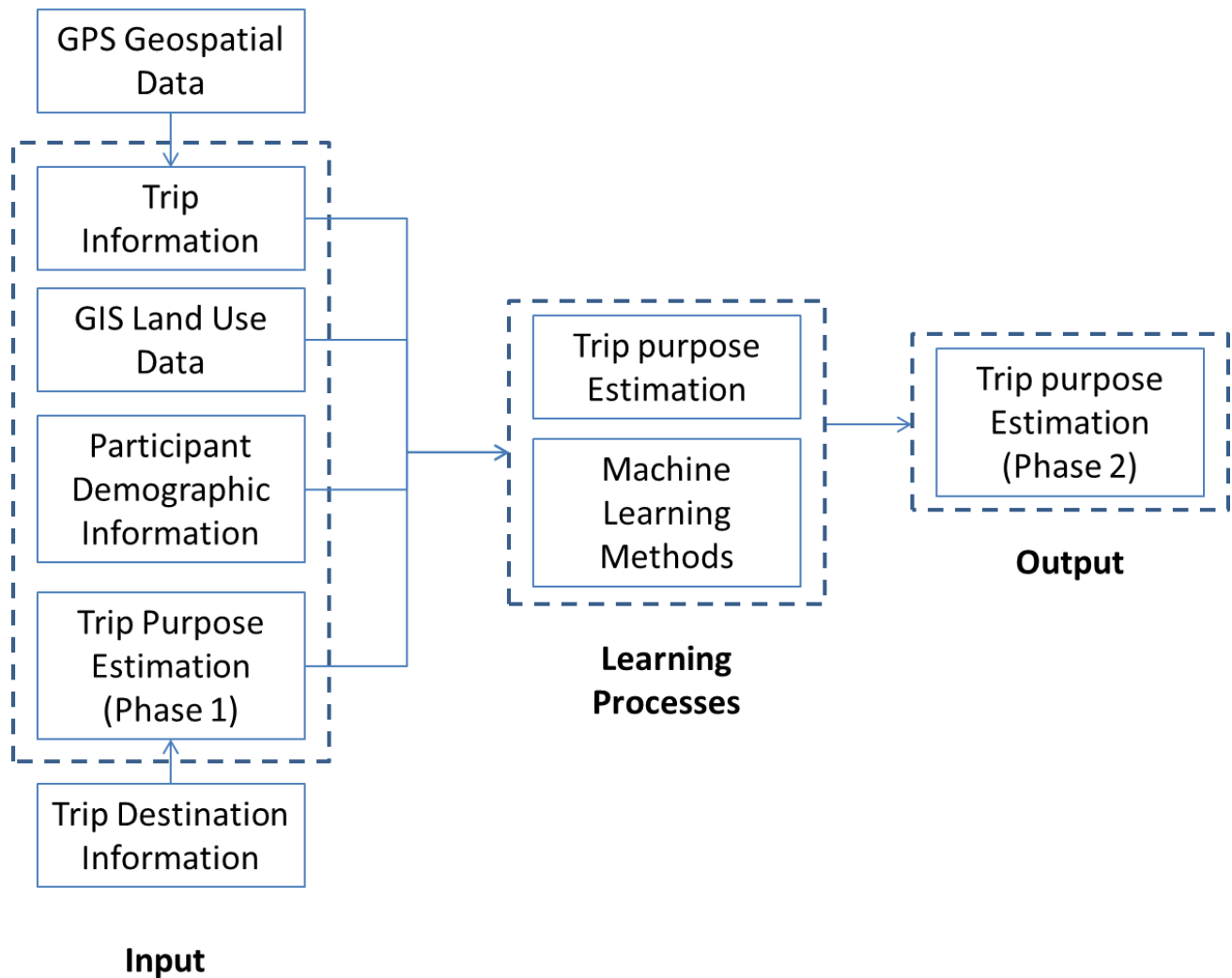


Figure 27- Trip Purpose Estimation Based on GPS, GIS, and Machine Learning Methods

Figure 27 shows the input for predicting the phase 2 Trip purpose. The Trip Destination Information is kept to learn off of, but is removed for the final estimation. Below is the variable list for machine learning.

Table 4- Input Variables for Trip Purpose Estimation (Phase 2)

Category	Variables for Trip
GPS Travel Attributes	<ul style="list-style-type: none"> • Trip Start Time • Day of Week • Previous Trip End Time

	<ul style="list-style-type: none"> • Soak Time • Previous Activity Duration • Latitude and Longitude • Distance to Home • Distance to Work
Participant's Characteristics	<ul style="list-style-type: none"> • Income • Education Level • Age
Land Use Data	<ul style="list-style-type: none"> • Land Use Type of Current location (start of trip) • Trip Origin Type (Home, Work, or Other)

Figure 28 shows the WEKA machine learning explorer. The dataset is transferred into an .arff file to load into the software. From there, the variables are selected and explored for impact on the predicted variable (purpose). The histogram in the bottom right corner shows the distribution of trip purposes as derived in Chapter 5. The algorithm then optimizes the decision tree for correct estimations of trip purposes using the variables available to it, then outputs the rule set.

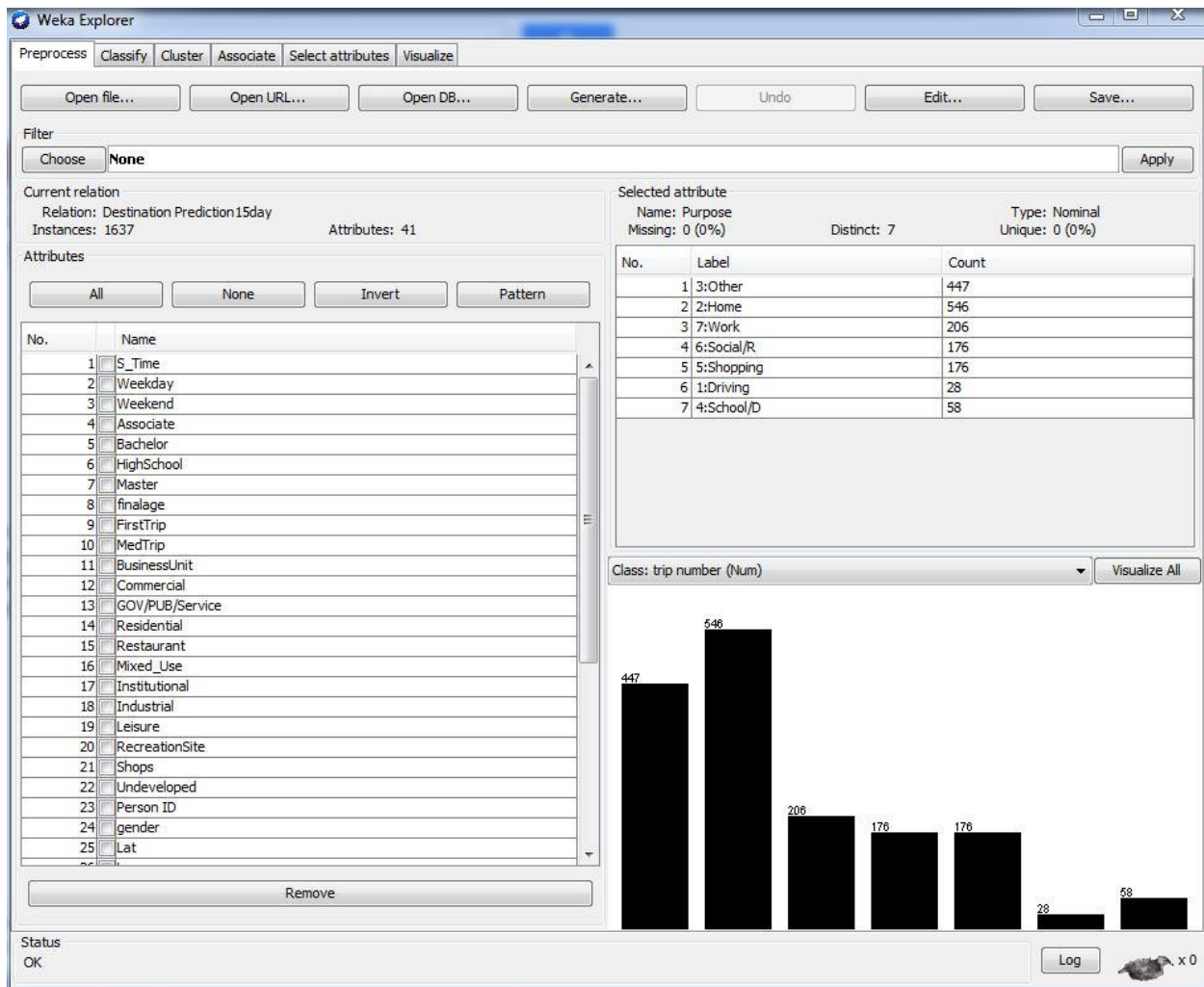


Figure 28- Setting the variable list and predicting variable (purpose) in WEKA for the J48 Algorithm

This rule set is then applied to the remainder of trips in the machine learning software by setting it is a new rule system (instead of learning algorithm such as the J48 or C4.5). Take as an example, one² such rule that was derived through this process and included in the 15-trip purpose rule set for allocating School type trips is:

If Distance to home < .4 miles

And

² The entire decision tree can be found in appendix D.

If time since last trip < 14 hours

And

If Driver's income < \$25,000

And

If it is the first trip of the day,

Then

Trip Purpose = School.

For all trips that meet this qualifier, the training set will allocate the trip purpose type "School" to the trip. The algorithm framework runs normally and searches for trip purpose first as the top Tier in step 2 (Figure 26), and maximizes the probability of previous locations that have been allocated as type "School" and predicts that location. If the user begins visiting a new location that is type school, the model will begin to estimate the more frequently used new location. The definitions for trip purpose cannot be changed after the 15 trip purpose learning period is complete, but the location predictions made by the trip purposes are updated based on travel patterns even after the 15 trip learning period. The results section shows a doubling of accuracy for School type trips when using the trip purpose model compared to the Tiered Time Origin Model.

6.4 Results

The results are given in two categories: with and without the trip purpose module. When the trip purpose module of the algorithm makes a prediction, then that prediction is used. If it does not have a prediction based on the user's purpose, then the lower tiers of the baseline model is used. Below, the accuracy of the two models is compared.

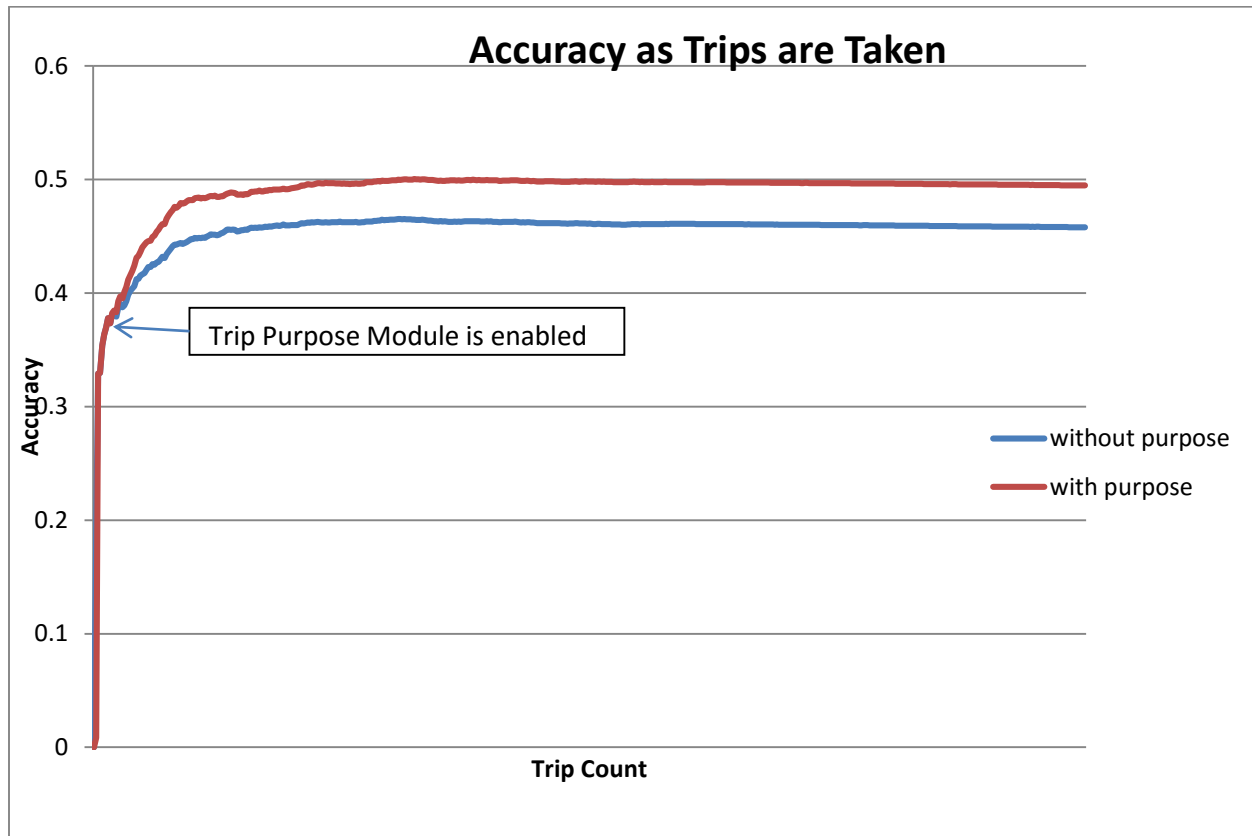


Figure 29- Algorithm Accuracy: With and Without Trip Purpose Module

Improvements are seen as soon as the trip purpose model is turned on at trip 16. By categorizing the previous 15 trips and searching only those locations that match the estimated trip purpose, a sudden increase in accuracy is shown. This chart shows the cumulative accuracy over the entire survey period. For instance, on trip 16, the graph shows the accuracy for trips 1-16. The true accuracy for trip 16 is 56.0%, and the cumulative accuracy shown is 40.5%, due to the very low accuracy when the model is accumulating knowledge and making poor estimations.

There is a slight drop off in accuracy as time goes on. This has not been previously studied in destination prediction. It is likely due to the model gaining a vast amount of learned data that is becoming older and unreliable. The decrease in accuracy is marginal (0.02%), but it is interesting that the model has a tipping point between amassing useful trip information and having too much information that is no longer helping overall accuracy. Future work could develop an algorithm that discounts older information when no longer accurately depicting current travel. This may marginally increase overall model accuracy.

6.4.1 Prediction Accuracy by Trip Purpose

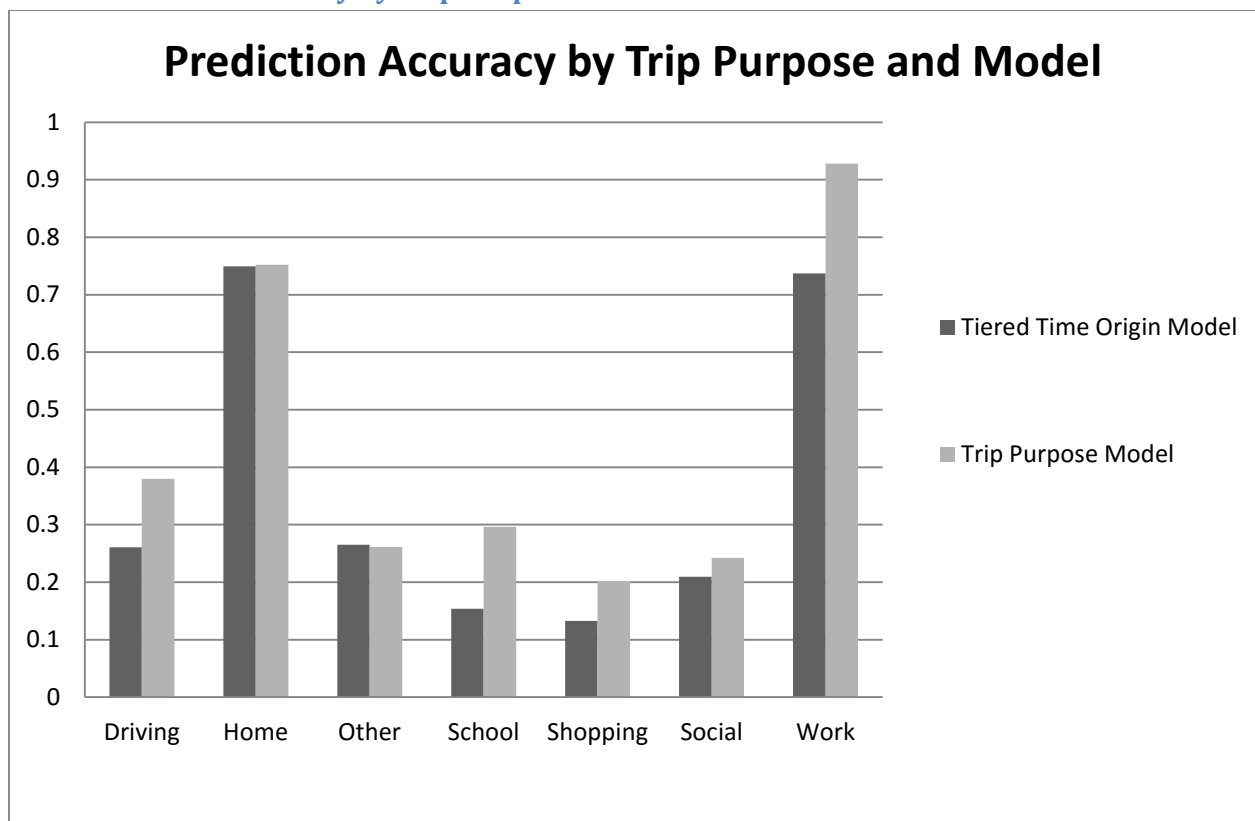


Figure 30- Prediction Accuracy by Trip Purpose

The increased accuracy of the purpose model is made by five trip purposes: Work, Social, School, Driving, and Shopping. Home and Other show either a slight decrease in accuracy or very

marginal improvement. This can be largely explained via the land use characteristics that help to identify these zones giving increased information about the area. The land use characteristics that identify these locations are: BusinessUnit, Commercial, Government/Public/Service, Residential, Restaurant, Mixed Use, Institutional, Industrial, Leisure, RecreationSite, Shops, and Undeveloped. School type trips see a doubling in accuracy due to matching similar trips that use the land use type Government/Public/Service. Driving is making a major improvement as well, although this trip type makes up a very small percentage of overall trips, not making a large impact on the model as a whole. Understandably, ‘Shopping’ type trips remain the hardest to accurately predict since there are many shopping locations an individual can visit compared to the relatively smaller number of Work, Home, School, etc. locations. Even with this difficulty, there is about a 5% improvement over the Tiered Time Origin baseline model.

6.4.2 Training sets of differing length

To show the value in the trip purpose model, the same methodology was used, except with a 5-trip and 30-trip learning set. Like in the 15 trip learning set, if the addition of the trip purpose module is benefitting the system as a whole, then we should see a noticeable uptick in accuracy after implementation. This leads to another interesting research question: what size learning model is best to maximize the overall accuracy of the model? Is it best to sacrifice the first 30 trip’s accuracy in order to more greatly increase the accuracy of trips 31-500, or does a small learning set of 5 trips suffice in increasing the model accuracy for after the learning set is enacted (without sacrificing accuracy for trips 6-30)? Below the overall accuracy with similar graphs is shown.

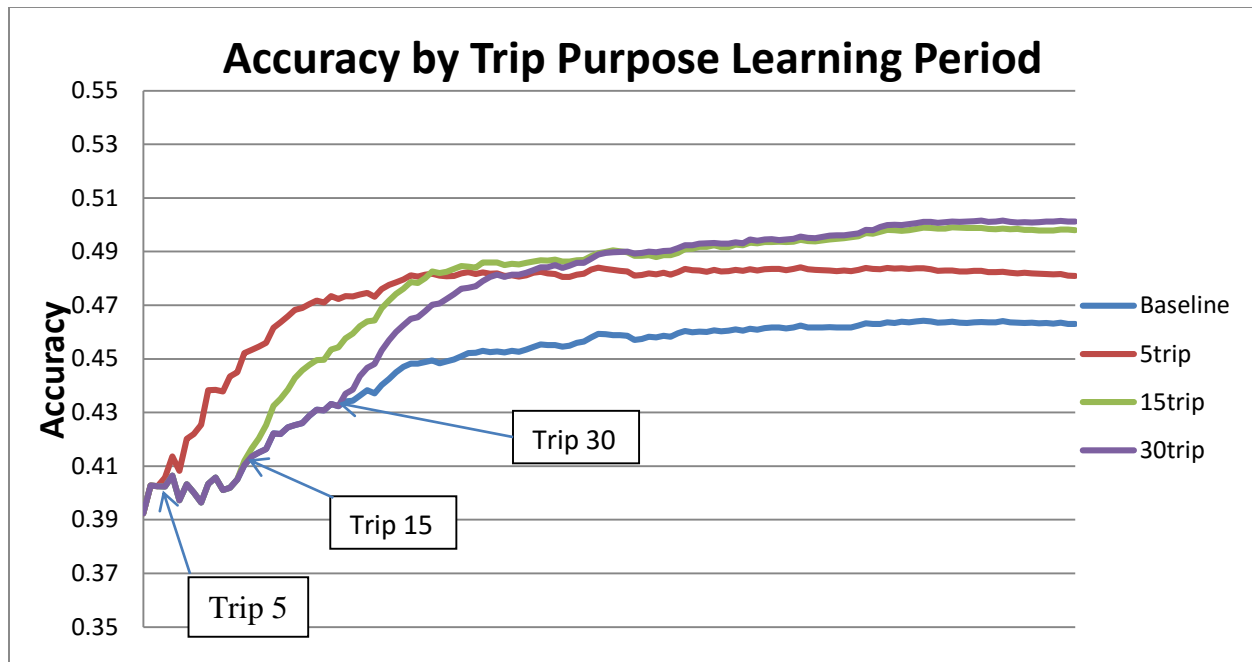


Figure 31-Accuracy by Trip Purpose Learning Period starting at trip 3 increasing to trip 130.

The 5-trip learning model has an accuracy advantage early on in the prediction process; with the highest accuracy levels until trip 42. The accuracy plateaus and eventually has the worst accuracy amongst trip purpose models. Again, the 15 trip set has an advantage over the 30 day trip set until trip 83, where the 30 trip learning set overtakes it. Clearly, the models benefit from more time learning, but it depends on the amount of time the survey period lasts, and whether the user is willing to accept lower accuracy predictions for their first trips to get more accurate trip predictions later on.

It would be plausible to run a new learning model after each trip is taken, updating the purpose definitions for all subsequent trips. This would not be difficult to accomplish for pseudo real time calculations, but in a real world environment, when trips may stop and start again in as little as 3 minutes, running a new learning set between trips may require significant computing power. To gain an idea of the computing power necessary, The 8 gig of ram quad core computer used in this

analysis often took over 20 minutes to learn from previous trips to predict the future trip purpose; the amount of data needed to run through the WEKA program is quite large, particularly with 10 folds.

6.4.3 Prediction Variation

The trip purpose model increases the prediction accuracy of the Tiered Time Origin Model by 5%, but it is also important to note the nature of the predictions. Since this is a start of trip prediction model, the prediction may be altered as time passes and in-route algorithms take over in future applications (Chapter 7). It is not only important to get the prediction correct, but if it is not correct, give a feasible subset for possible locations. This was studied by looking at the prediction types in the Tiered Time Origin and Trip Purpose models.

Table 5- Model Accuracy Compared to Trip Purpose Prediction

Model	Percent of Destinations Estimated as Home or Work Location	Overall Model Accuracy	Model accuracy after learning period
Tiered Time Origin (Baseline)	90.5%	45.7%	N/A
Trip Purpose Learning (30 Trip)	77.3%	50.03%	51.16%
Trip Purpose Learning (15 Trip)	82.0%	49.4%	50.11%
Trip Purpose Learning (5 Trip)	89.6%	47.1%	47.22%

Combined 5-15-30	62.4%	52.34%	52.74%
Trip Purpose			

6.4.4 Confusion Matrix

Confusion matrices are shown below that explains how well each trip purpose prediction matches up with the prediction that is made in Chapter 5 (i.e. the original prediction with roughly 81% accuracy). For each of the three learning periods, the accuracy by purpose type is compared.

Table 6- 5-trip Purpose Confusion Matrix

Count of Correct?	Column Labels							Grand Total	Accuracy
Row Labels	1:Drivin	2:Home	3:Other	4:School	5:Shoppi	6:Social	7:Work		
1:Driving	73	44	73	8	41	29	0	268	27%
2:Home	97	5053	637	28	411	462	20	6708	75%
3:Other	53	975	2651	2	1211	700	200	5792	46%
4:School/D	16	48	190	308	112	18	11	703	48%
5:Shopping	27	364	733	11	939	62	2	2138	44%
6:Social/R	23	738	370	19	152	836	0	2138	39%
7:Work	0	22	71	1	1	4	2216	2315	96%
Grand Total	289	7244	4725	377	2867	2111	2449	20062	60.2%

Table 7- 15-trip Purpose Confusion Matrix

Count of correct?	Column Labels							Grand Total	Accuracy
Row Labels	Driving	Home	Other	School	Shopping	Social	Work		
Driving	65	27	89	17	27	24	0	249	26%
Home	84	4763	899	56	257	289	14	6362	75%
Other	68	775	3137	106	712	610	96	5504	57%
School	26	37	107	318	114	32	28	662	48%
Shopping	27	222	904	64	725	85	0	2027	36%
Social	60	506	384	26	70	975	0	2021	48%
Work	0	47	172	1	1	4	1964	2189	90%
Grand Total	330	6377	5692	588	1906	2019		19014	62.8%

Table 8- 30-trip Purpose Confusion Matrix

Count of correct?	Column Labels							Grand Total	Accuracy
Row Labels	1:Drivin	2:Home	3:Other	4:School	5:Shoppi	6:Social	7:Work		
1:Driving	55	49	75	4	25	21	0	229	24%
2:Home	65	4650	507	42	135	344	12	5755	81%
3:Other	59	534	3288	78	523	493	32	5007	66%
4:School/D	8	38	101	359	58	41	0	605	59%
5:Shopping	26	222	810	33	709	51	0	1851	38%
6:Social/R	27	362	386	20	69	987	0	1851	53%
7:Work	0	9	54	1	1	3	1917	1985	97%
Grand Total	240	5864	5221	537	1520	1940	1961	17283	69.2%

Table 5- Estimated Trip Origin Land Use Distribution

Trip Origin Land Use	Estimated Trips	Percentage
Residential	4723	33%
Recreation/Leisure	2474	17%
Business/Commercial/Industrial	1047	7%
Mixed_Use/Shops/Restaurants	2731	19%
GOV/PUB/Service/Institutional	3542	24%
Land Use Total	14517	
Total Trips	20651	
% of Total Trips	70%	

The Tiered Time Origin model estimates either home or work location over ninety percent of the time. At such high levels of home and work prediction, the model accuracy cannot be high. At the start of a trip, the most likely location by time of day, day of week, and origin is almost always either home or work. The results show the trip purpose model gains accuracy by shifting many of the trips that were being estimated as either home or work to other trip types. By not over-estimating work trips, that trip purpose type increased by 20%. Also, destinations which can be signified by land use type such as shopping, social, and school, all saw an increase in accuracy. This is likely due to the shifting away from over-estimating work type trips. Also, a major difference in the 5, 15, and 30 trip purpose learning model can be seen. Additional time to learn from user's trip purpose activity does improve the model despite only marginal increase in

accuracy. When this model is applied to in-route destination prediction, a major increase in accuracy may be seen.

6.4.5 Combination 5-15-30 trip learning models

To get the most accurate model possible, the 5, 15, and 30 trip training models are combined. This allows the model to continue to learn about trip purpose classification while trips are being undertaken. After the fifth trip is taken, the 5 trip model with learned trip purpose behavior of the previous 5 trips is turned on. For the next 10 trips, this learning mechanism is used. After the fifteenth trip is taken, the 15 trip learning model is turned on, and the 5 trip learning model is turned off. Since information has been learned throughout the past 15 trips, the model is more accurate, and it better able to estimate trip purpose leading to better destination predictions. The same events occur for the 30 trip model: for trips 15-30, the 15 trip model is used, and then the 30 trip model is turned on. This results in another uptick in performance. There is one drawback to this method; in order to stay at such a high accuracy level, a new learning model will need to be implemented every 15 trips, without a 45,60,75,90, etc... trip learning model, the model accuracy begins to taper off. The model reaches its' maximum accuracy at trip #144. At this point the behavior of the drivers tend to change and the models for trip purpose start to become outdated. With these outdated trip purpose definitions, the model accuracy decreases from 53.5% to 52.7%. This is still a major improvement over the single stage learning models, and the baseline tiered time origin model.

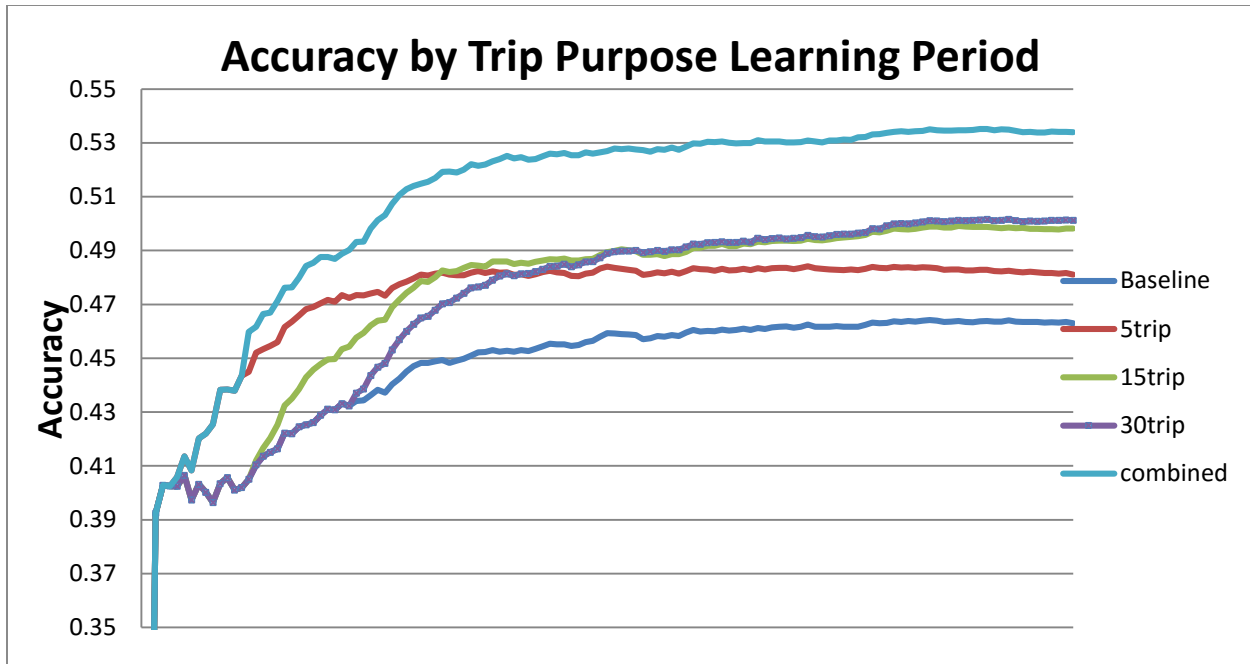


Figure 32- Model Accuracy with combined learning method

6.4.6 Comparison with state-of-the-art models

The literature gives the best destination prediction model in Gao et al.'s (2012) HPHD with an accuracy of 50.05%. The dataset includes 3,373 locations and a 315 day training set. The Tiered Time Origin model in this paper has both most frequent hour and day categorization, and produces a model accuracy of 45.7% with the GPS data collected: 20,652 locations and no training period. It is difficult to compare such models, as the data is drastically different. The data used in the HPHD model is cell phone trace data, which is on the user much more often than the inside the vehicle data used for this approach. This may lead to higher accuracy due to capturing similar locations using multiple modes of transportation, and more importantly, the ability to define a location based on speed of the location trace. Using cell phone data, the end point of a trip is more defined, whereas a vehicle may need to park hundreds of meters away from its intended destination, which leads to a higher number of end locations.

The purpose model, with the combination 5-15-30 trip training period and 20,652 locations has an accuracy of 52.74%, improving the best accuracy model (HPHD) in the literature by 2.69%. This

is without any training period, starting with an incorrect guess at the start of every survey period. Taking into account the network size (including long distance trips in 22 states) and about eight times the number of locations, the benefit of the newly generated model may be larger than the accuracy improvements suggest. What should not be lost is that the improvement over the Tiered Time Origin Model is 7.04%, which is considerable improvement over the baseline model.

6.5 Regression Analysis

The following tables present the regression statistics for the trip destination prediction model without derived trip purpose and with derived trip purpose:

Table 6- Regression Statistics- TTOM

Tiered Time Origin Model Without Derived Trip Purpose	
Regression Statistics	
Multiple R	0.565
R Square	0.319
Adjusted R Square	0.318
Standard Error	0.411
Observations	20651

Table 7- Regressions Statistics- Trip Purpose Model

Tiered Time Origin Model With Derived Trip Purpose	
Regression Statistics	

Multiple R	0.581
R Square	0.338
Adjusted R Square	0.337
Standard Error	0.407
Observations	20651

Human behavior is not easy to predict, thus the relatively low R squared values received from the model. What we want to study is the impact of each variable on the overall accuracy of the model. It is more important in the early stages of trip purpose destination prediction algorithms to figure out what impacts the overall accuracy of the model for future algorithms to adapt to these findings. For example, in finding that Social type trips have the lowest accuracy, and social type land use characteristics impact that heavily, changing prediction methodologies for those trips that give social as the trip type may increase accuracy greatly.

The regression analysis is done using 16 independent variables and one dependent variable. Dependent variable (y axis) is prediction accuracy, and the independent variables are: trip number, income, soak time, day (week or weekday), trip purpose, trip time of day and the trip origin land use. The following tables show the regression statistics from each of these variables in the Tiered Time Origin model without and with derived purpose respectively:

Table 8- Regression Model- without Trip Purpose

<i>Variables</i>	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	0.2004	0.0139	14.3724	0.0000

Trip number	0.0001	0.0000	3.1747	0.0015
Income	0.0000	0.0000	6.7575	0.0000
Soak time	0.0000	0.0000	22.3105	0.0000
Weekday	0.0570	0.0067	8.4861	0.00000
Time of day	-0.1081	0.0104	-10.3945	0.0000
Driving	-0.0215	0.0234	-0.9197	0.3578
Home	0.3777	0.0080	47.0458	0.0000
School/D	-0.1100	0.0175	-6.3026	0.0000
Shopping	-0.0888	0.0106	-8.4179	0.0000
Social/R	-0.1235	0.0108	-11.3902	0.0000
Work	0.3778	0.0107	35.4468	0.0000
Residential	0.1390	0.0076	18.3071	0.0000
Recreation/Leisure	0.0081	0.0095	0.8531	0.3936
Business/Commercial/				
Industrial	0.0631	0.0134	4.7053	0.0000
Mixed_Use/Shops/				
Restaurants	-0.1514	0.0101	-14.9390	0.0000
GOV/PUB/Service/				
Institutional	0.0300	0.0086	3.4933	0.0005

From the coefficients of Table 15, it can be concluded that the variables which have the most effect on the accuracy of the model are Work and Home. Work and Home each increases accuracy by 37.8% over the baseline purpose of Other. Since Work and Home are the most common trip purposes (and each are usually at a constant location), they are easiest to predict. The next value

that adds the most percentage on accuracy is the trip origin land use: Residential. The trip origin location land use Mixed_Use/Shops/Restaurants contributes with the highest decrease in the model's accuracy over the baseline characteristic of None. A possible explanation might be that it is hard to predict a trip that originates from Mixed_Use, Shops or Restaurants because there are a lot of possibilities for trip purpose. Looking at the P-value results, it can be concluded that 14 variables have meaningful value in the model (are significant), since their P-values <0.05 . Only Driving purpose and Recreation/Leisure origin land use were not significant to the model. The model which uses trip purpose classification can be found in table 16.

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>
Intercept	0.1646	0.0138	11.9343	0.0000
Trip number	0.0002	0.0000	4.8582	0.0000
Income	0.0000	0.0000	7.3886	0.0000
Soak Time	0.0000	0.0000	19.7496	0.0000
Weekday	0.0335	0.0066	5.0363	0.0000
Time of day	-0.0569	0.0103	-5.5251	0.0000
Driving	0.0715	0.0232	3.0882	0.0020
Home	0.3930	0.0079	49.4689	0.0000
School/D	0.0314	0.0173	1.8201	0.0688
Shopping	-0.0193	0.0104	-1.8495	0.0644
Social/R	-0.0881	0.0107	-8.2093	0.0000
Work	0.5938	0.0105	56.2976	0.0000
Residential	0.1570	0.0075	20.8884	0.0000
Recreation/Leisure	0.0294	0.0094	3.1203	0.0018
Business/Commercial/				
Industrial	0.0380	0.0133	2.8606	0.0042
Mixed_Use/Shops/				
Restaurants	-0.1222	0.0100	-12.1821	0.0000
GOV/PUB/Service/				
Institutional	0.0209	0.0085	2.4529	0.0142

Table 9- Regression Model- with Trip Purpose Tier

The variables which have the most effect on accuracy are again Work and Home. When compared to the contributions made by the Work and Home on the Tiered Time Origin Model, Work contribution to model's accuracy increases by 21.6%, while Home contribution increases by 1.5%. The variable which causes the highest accuracy decrease is Mixed_Use/Shops/Restaurants origin land use. Even though it still decreases the accuracy; it still has a 2.9% improvement on the effect when compared to TTOM. The other variables that decrease the accuracy are: Social purpose and time of day with. As explained before, Social purpose is hard to predict because people's taste toward social activities tend to change and also there are a lot of variability and choices in social activities. In terms of the time of day, it can be seen that as time of day increases, the accuracy decreases. Morning trips are more predictable than evening and night because trips at the morning are mostly from home to work, while trips at evening or night can be from work to a social location, restaurant, shop, etc. The P-value results show that all variables are significant in the model, except School and Shopping purpose.

As the overall model results show, every variable when using trip purpose methodology increases its positive impact on accuracy of the model. Home purpose, Work purpose, and Residential origin land use have the highest accuracy. Most importantly, the trip purpose methodology turns Mixed_Use/Shops/Restaurants, Driving and School variables from decreasing the overall accuracy of the model, to increasing it. The trip purpose module has a true positive benefit over the accuracy of the model.

6.6 Conclusion

This is the first research to use trip purpose for predicting destination location. Literature review shows the need for increasing start of trip prediction accuracy, with little advancements made in this area. Using the approaches defined in this paper, a small amount of GPS data is used to

estimate an accurate start of trip destination. By incorporating demographic and land use data, trip purpose is derived, which improves the accuracy of this start of trip model. Overall accuracy of a baseline spatial temporal Markov model is improved by approximately 5% by the end of the survey period, but improvements can be seen as early as trip number 6, which shows a fast improvement with little drawback. The model was shown to be over 90% accurate at predicting work trips. By giving users advanced route-specific travel information before they enter the network, the way people commute on a daily basis could be significantly impacted. This new methodology can lead to significant advances by applying it to in-route destination prediction algorithms.

CHAPTER 7: In-Route Destination Prediction

7.1 Introduction

As of yet in the dissertation, the prediction algorithm has been activated and enacted at the start of each trip. The trip purpose module developed in chapter 6 proves that a narrowed group of destinations is predicted. These locations include only those that are supported by particular trip purposes. Table 8 (of Chapter 6) shows that the 5-15-30 combined trip training model resorts to choosing non home and work locations 37.6% of the time. The chart below shows that the Trip Purpose Model (combined 5-15-30 learning sets) is able to estimate varying trip purpose types correctly. This is noted by the varying colors of the top left corner of the chart (estimations correctly made by the trip purpose model, but incorrectly made by the Tiered Time Origin model). The bottom right corner shows the opposite (correct estimation for the Tiered Time Origin model, but incorrect for the purpose model). The mostly red color shows the propensity to over predict Home based trips. The bottom left shows that both models still have difficult with ‘Other’ type trips.

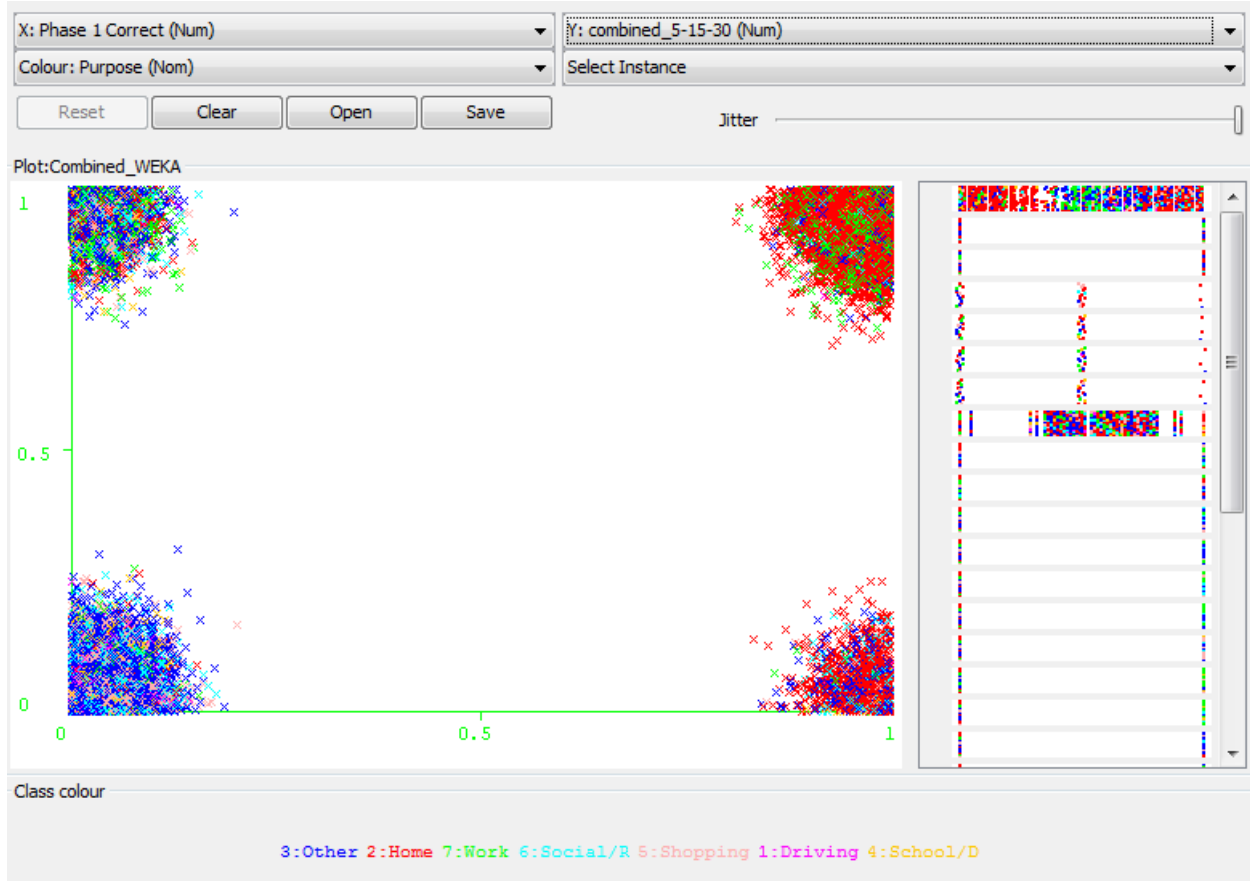


Figure 33- Comparison of accuracy by trip purpose for Tiered Time-Origin versus Trip Purpose Model

Using this specified group of locations, it is hypothesized that not only is the destination prediction more accurate, but the estimation options are more accurate as well. The possible number of locations is likely rather low for a particular trip purpose.

There are current in-route prediction models that re-predicts with each new GPS point. This works by giving more weight to a destination that it is getting nearer to. By the end of the trip, the model is more accurate, often times getting near 85% accuracy. The models use all available trip destinations that have been previously visited, then after each step apply the statistical model to give new chances to each of the available destinations.

The work in this chapter will use the trip purpose module explained in chapter 6 and use it as the input for an in-route prediction model. The in-route model will use the start of trip model as the first possible estimation. At the 25%, 50%, 75%, and 90% trip completion marks the algorithm will run to match the current route taken with the database of trips that has already been taken. The database is kept in 3 different bins: routes with the same estimated trip purpose as past trips, routes with the same trip origin as the last trip, and routes with no trip purpose match and no trip origin match. This allows the model to, just like the start of trip model, use the best model that is available to it in real time.

The model maximizes the accuracy of destination prediction by selecting the destination that matches the highest value of route links. The question remains: is a 50% route match on a trip with the same trip origin more likely to have a destination match than a 40% route match on a trip with the same trip purpose? The results portion will provide tables on the accuracy of each of the 3 bins, with varying route match percents on various route completion percents. Using these tables a final model is constructed that uses the different bins when they are available, and uses the one with the highest accuracy level available to it at all times.

The decision of when to use which of the 3 bins of route information is an important one to make. Often times there are multiple routes available for matching from each of the 3 bins (there are always more trips available in the 3rd bin, since it requires no specification). From the results provided, a final model will be formulated which uses the best model at all times. The model is fairly simple: maximize the route match from the 3 bins at all times, selected the route that matches the current one with the highest accuracy related to previous trips (based on final destination

accuracy), then output the destination prediction. The model reruns with every new GPS point and roadway link, then outputs a new prediction at 0%, 25%, 50%, 75%, and finally 90% of the completion of the trip. The destination prediction may change many times during the trip, and can even switch back and forth depending on the route that the individual is taking during the current trip.

The remainder of the chapter will cover a literature review of in-route prediction models, followed by an example run-through of the algorithm with regards to the 3 bins, a model formulation of the in-route model, the raw results output from the models, and finally a creation of a singular model that incorporates all three bin models, with the highest accuracy model being compared to the one's covered in the literature review.

7.2 Lit review – In-route prediction

There have been several in-route destination prediction algorithms. This literature will cover three models that are comparable and state of the art. First, the in-route models began in 2008 with Krumm, and Horowitz, where the GPS locations were set into a grid network and future destination is predicted based on trajectory and a Markov Chain system. The overall accuracy of the model is highly dependent upon the shape and features of the network.

The most similar and comparable model with the highest end accuracy in the literature is shown by

Alvarez-Garcia, Juan Antonio, et al. "Trip destination prediction based on past GPS log using a hidden markov model." *Expert Systems with Applications* 37.12 (2010): 8166-8171. The model maximizes the route match in a markov model that considers important pivot points where turns and differences in routes are often made. By considering the turning movements at important

junction points, the likelihood at going to a certain destination can be determined with fair accuracy. The model increases in accuracy considerably as the trip progresses. The final model will be compared to the Alvarez and Antonio model due to the high accuracy end of trip prediction accuracy. Only locations that have been visited 3 prior times were considered in the results portion of this paper.

The model to most closely compare the proposed trip-purpose in route model is the Alvarez-Garcia (2010) model of in-route prediction with a hidden Markov model. The comparison table below shows the total number of locations per participant, and visits per location

Table 10- Literature Review model data comparison

	Trip Purpose (Krause)	Hidden Markov (Alvarez-Garcia)
Locations per participant	7.5	6.8
Visits per location	17.0	15.7

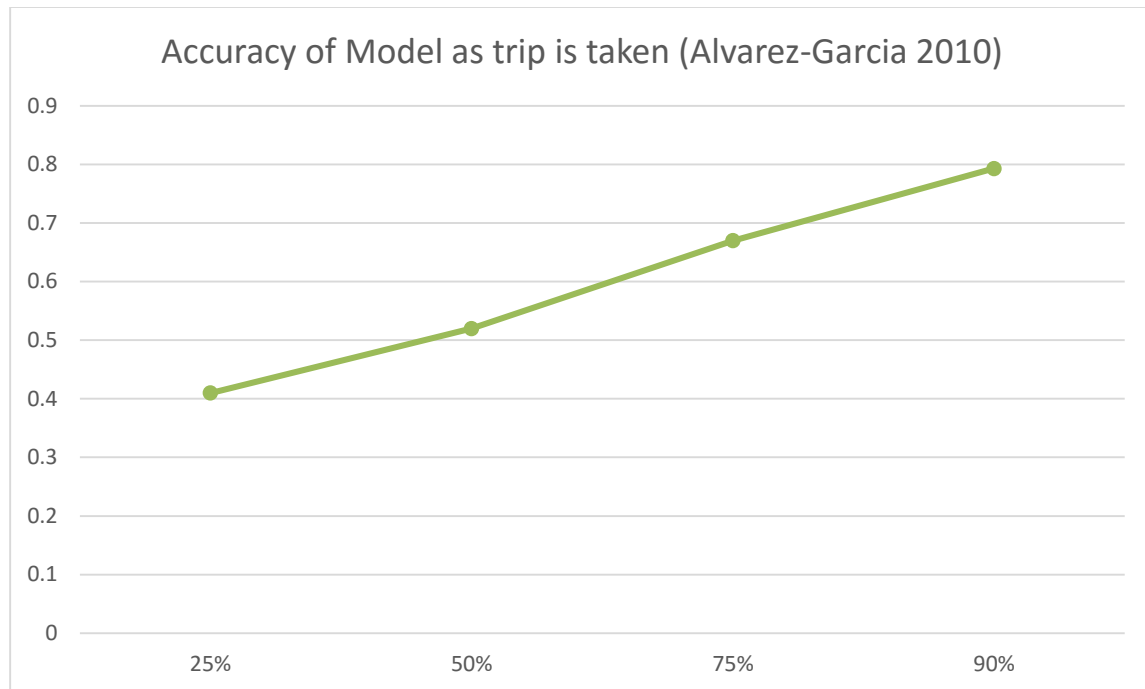


Figure 34- Alvarez-Garcia model participant averaged

Horowitz and Krumm have been leaders in the research field with multiple papers (2006, 2008, 2010). Starting in 2006 Krumm tested the performance of a route prediction algorithm that gave a median destination error of two kilometers at 50% trip progression. Krumm's 2008 paper: A markov model for driver turn prediction uses a simple markov model to make probabilistic predictions by looking at a driver's most recent path along the trip. Similar to this work, the model is trained from their long term history and predictions are made based on their historic choices. In one participant selection, the model was able to predict with 90% accuracy the direction of the next turn along all routes. Krumm's 2008 "Route Prediction from Trip Observations a simple markov model is used to predict the destination of trip using another simple markov model on learned behavior. The end results show about a 40% end of trip accuracy for all locations, and a 40% accuracy at midway point on destination locations that have previously been visited. For non-repeat trips, the start of trip model actually performs better using trip purpose. Compared to the repeat trips model, the blue line indicates the single prediction being made is accurate, purple

meaning that one of the top 10 prediction is accurate. When comparing to the model explained in this dissertation, the Krumm 2008 performs worse at every point (except at 100%, where the trip purpose model is not run). Running the model at 100% completion is a rather trivial task realistically. There is not much point in telling people where they are going when they have already arrived, and accuracy can easily be maximized at 100% trip completion with small algorithm tweaks. These are the main reasons why accuracy is not measured at 100% trip completion for the trip purpose model.

In the results portion of this paper, the papers reviewed here will be revisited for a full accuracy comparison.

7.3 Methodology

The first step in creating an in-route trip prediction model from the start of trip model is to link the roadway network to the GPS trajectories. The section below describes how the GPS points collected from the travel survey was used to allocate roadway segments to the individual trip.

7.3.1 Roadway Link Allocation

The first step in creating an in-route trip prediction model from the start of trip model is to link the roadway network to the GPS trajectories. The GPS travel points between the origin and destination are connected to the links of the roadway required to travel between nodes. The nature of the GPS points makes this a rather difficult task; points were only taken at a frequency of once per minute. The roadway network has many more links per trip than there are points per trip. Therefore, for each trip, a script was used to connect the links that would be required to take a trip based on the points along the network, based on the shortest path between those points. While it cannot be guaranteed that the final links were actually taken by the real-world driver, the likelihood of taking those links for each trip is rather high.

On a large scale GPS travel survey, it is understandable that the recording frequency is rather low. For a higher recording frequency, the budget for the survey would need to be very high. Based on these constraints and data not being used to develop a commercial application, the route allocation for GPS travel is sufficient. For real world deployment, such as the Ann Arbor Safety Pilot Deployment, a 10Hz GPS device was used, which allows for much more accurate roadway layer linkage to GPS points. The images below give an example of a GPS trip with 3 points, and the links that are allocated to it between points 1 and 2.

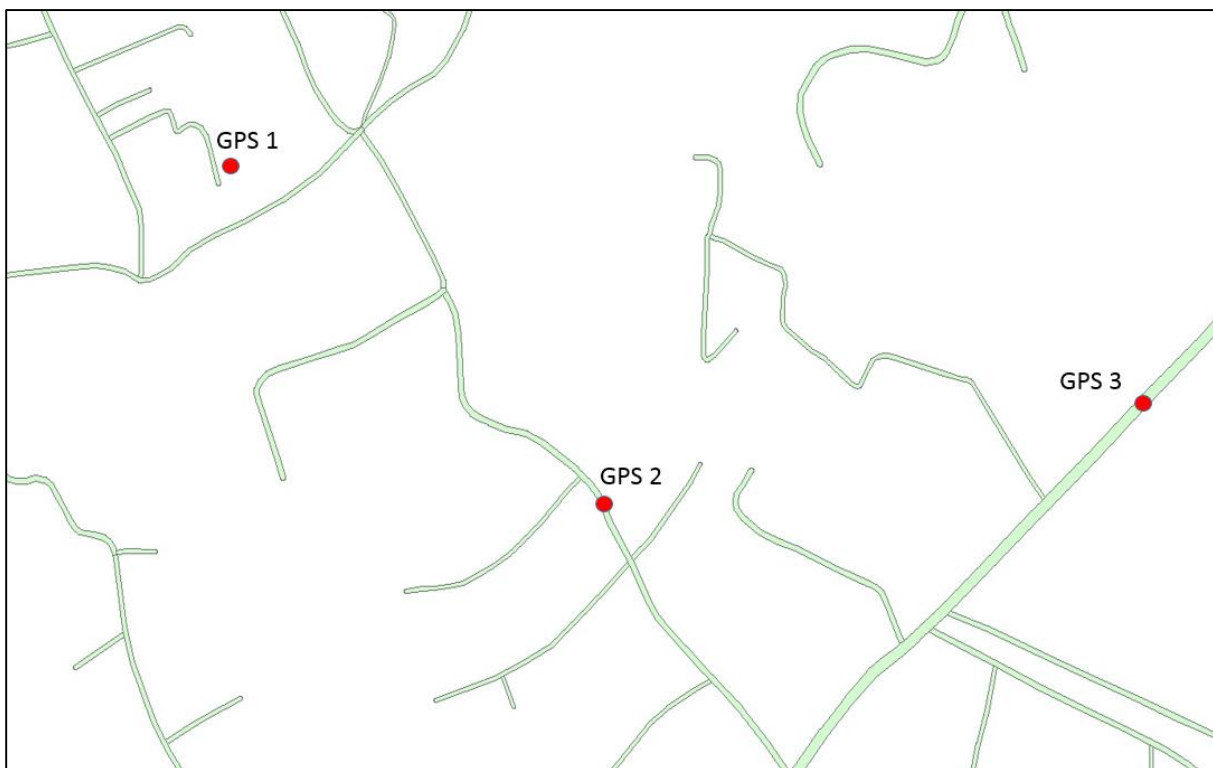


Figure 35: GPS points in roadway layer for link allocation

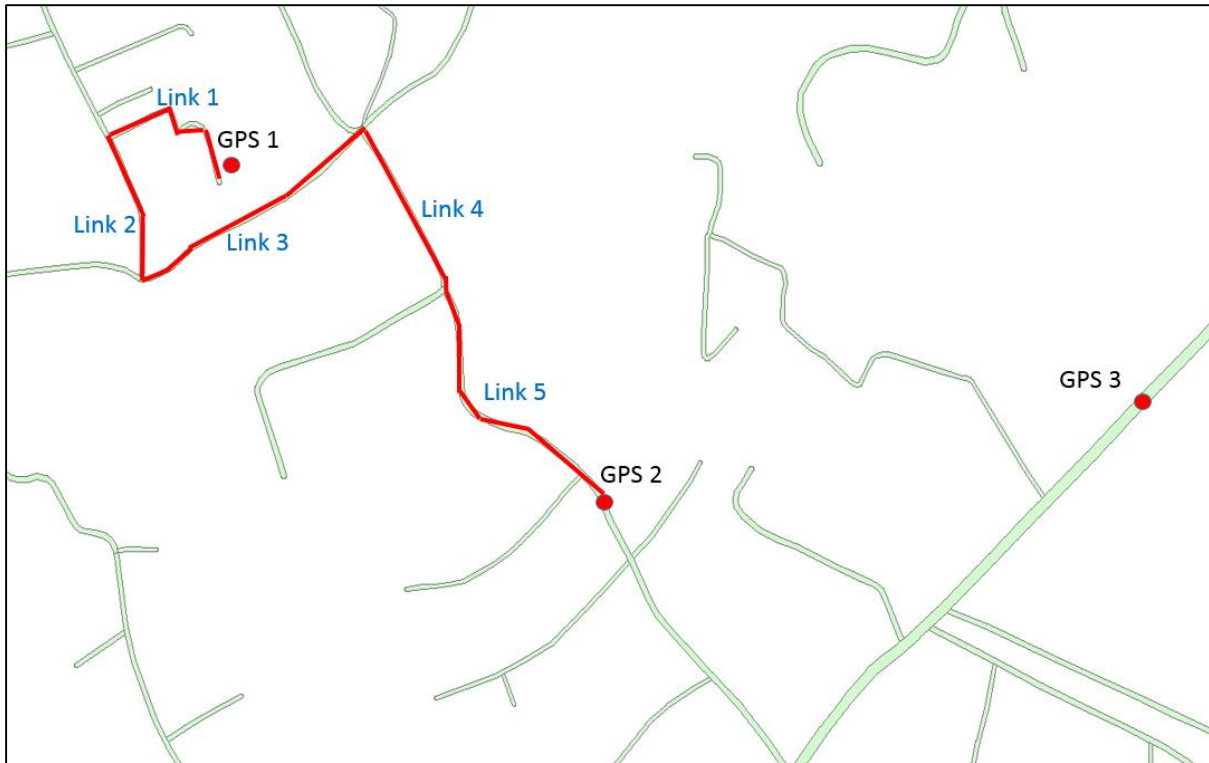


Figure 36: Links matched to GPS points based on shortest path between points

7.3.2 Destination Prediction

The process for selecting a destination is rather simple under this methodology. The algorithm developed starts from the first point of the trip (origin), and searches the road link it is assigned to. This is a 10 digit number in the roadway file. The algorithm then searches for all trips in its memory bank that also includes this 7 digit roadway identifier. For each trip that has this link ID in the trip, it receives one point. Then, this process is repeated for all links in the trip route. For example: given a 60 minute trip, there will be 60 GPS points. Since a link often occurs more often than once per minute, the links are split up to the closest minute rounded down. The results table is broken down into percent of trip taken. Under the 25% column, in our example, 15 minutes of GPS trips will be taken, and given there are a total of 100 links, 25 of those links will be matched with the first 25% of the links in each trip of the user's memory bank. The trip that it matches the most number of links along the route will be assigned as the estimated route to the predicted destination.

As a factor of confidence, the percentage match is recorded. A final selection can be made based upon percentage match of the route. For instance, if a route is matched 100% to a trip in the user's memory bank, the chances are high that those two trips are going to arrive at the same destination. This becomes even more likely as the algorithm uses 50%, 75% or up to 90% of the available trip.

7.4 Model Formulation

The formulation is similar to that in the start of trip model, in that it maximizes the likelihood of choosing a destination based on highest percentage match on the route, with all previous routes taken by the individual. By tracking the location of each GPS point, its location on the roadway network, and matching those network pieces, are we able to find a percent match to previous trips take, how that percentage matches up to the amount of the trip that is taken, then choose the most likely destination based on the route match. Each of the 3 sections below show the formulation for trip purpose based route selection, origin-based route selection, and the route selection with no other qualifiers. The route selection with no other qualifiers will always have the highest possible percentage match, but results will show that the overall accuracy of the model is significantly lower.

A graphical representation of the model is shown below. First, the trip purpose specification is used when available (having the highest probable level of accuracy). The next best case is the availability of Origin based routes for destination prediction, followed by no qualification in route based selection. Finally, if there is no route match along the in-route trip progression, the framework falls back to the start of trip model explored in previous chapters of this research.

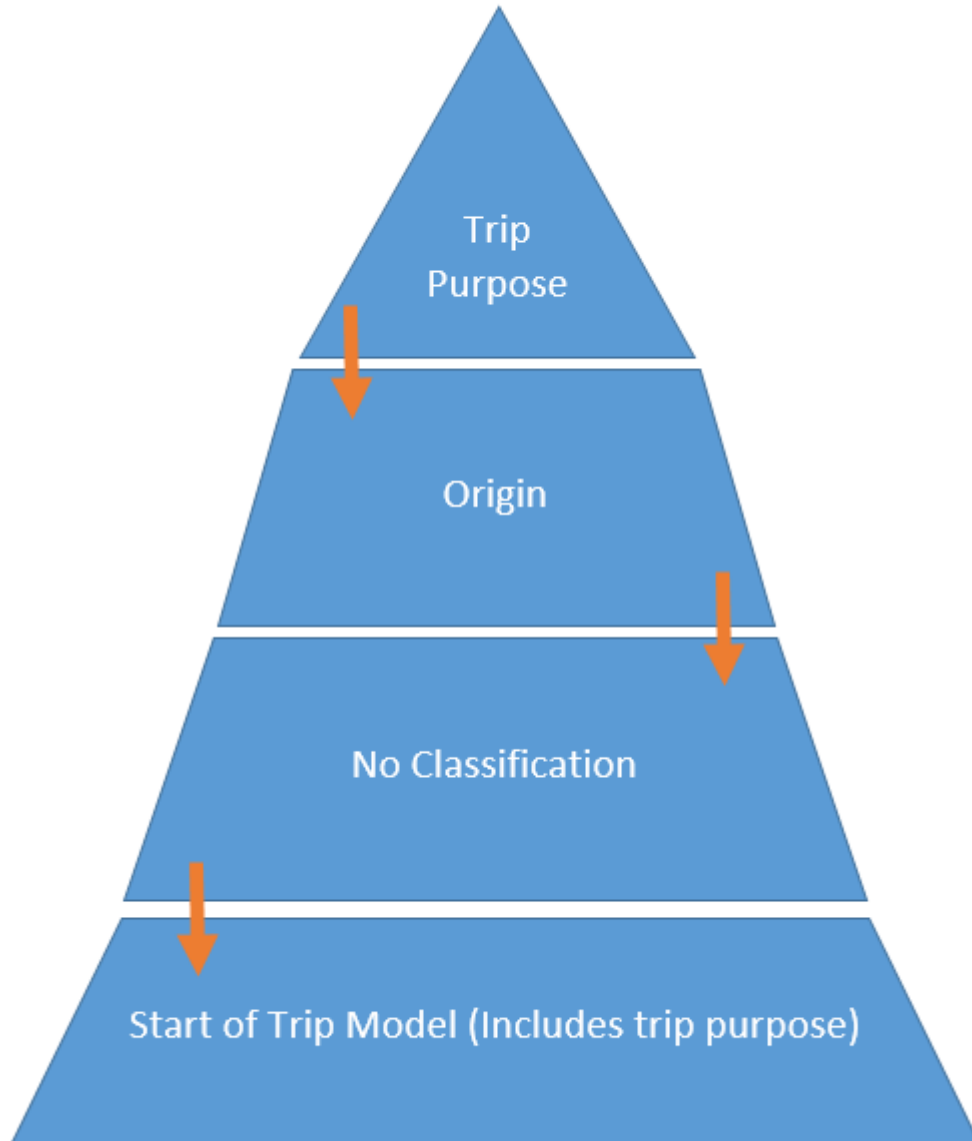


Figure 37- Model Formulation graphical representation.

Only up to a point is this used, for this trip purpose model, even if another location has a higher percentage route match, it only chooses those locations that have the same trip purpose. This may seem counterintuitive, but unlike many other in-route models, this one takes into more than just route information, trip purpose being a major contributor.

7.4.1 Trip Purpose:

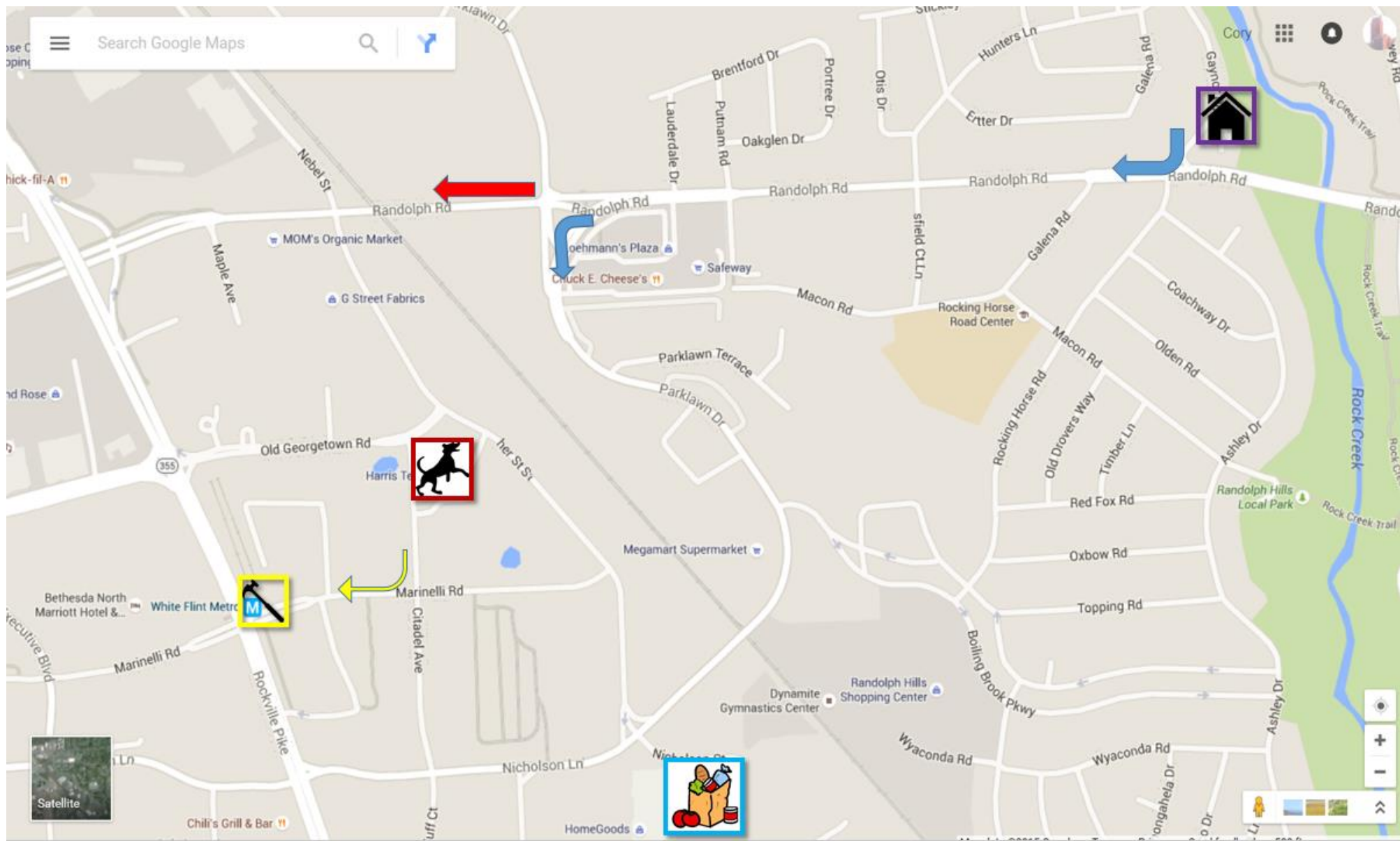
The model formulation is below:

$$P(v_i = l | u_i = u, t_i = T, v_{i-1} = l_k) = \frac{\sum \{N_c, u_i | N_c \in N, u_i \in U, N_c = N_r, p_i = p_k\}}{\sum \{N_r, u_i | N_r \in N, p_i = p_k\}}$$

The formulation maximizes the destination prediction by selected the route that most highly matches the current route being taken. But, only those routes that match the current trip purpose is selected.

Example of trip purpose prediction method.

The below image is an example of how the trip purpose prediction works, and when the trip purpose model can perform better, and when it would perform worse than a simple maximization by route links and location. In the example below, the trip starts off in the home location (purple) and drives west. From the start of the trip, the start of trip model determines the trip purpose, then only loads the routes that have already been taken with the same trip purpose. The trip purpose is determined by time of day, origin of trip, along with many other variables. From the start of trip, and the first two links of the roadway network, the in-route model matches the first two roadway links with other trips of the same purpose with the first two links. The destination is predicted with the highest percentage of link matches with the same trip purpose. Every minute, as another GPS point loads into the system, the roadway network travel links are generated, then the percentage roadway match is recalculated, and the destination is re-predicted with this updated information. In the example below, there is a leisure/social activity (dog park), shopping activity (grocery store), and work (work location). If the trip purpose model accurately predicts the trip purpose (model accuracy around 70%), the chances are very high that a correct prediction will be made in the first minute of the trip. If however, the trip purpose model is incorrect, the chances become very low that a correct prediction will be made at all, since only destinations that match the correct purpose are considered. No matter how accurately the route matches previous one's in the memory bank, it will not predict that location. This leads to a unique phenomenon: the accuracy is very high at the beginning of the trip (compared to previous models), but will not grow significantly as the trip continues. The next example shows how destination prediction based on route maximization functions.



7.4.2 Origin:

Using the same map as above, the Origin model maximizes the routes available, but only loads those that start from the same origin as the current trip. This may have a higher accuracy than the trip purpose model for a few reasons:

- 1) The trip purpose model may inaccurately estimate the trip purpose, leading to a significantly lower chance of a correct destination prediction,
- 2) There may be multiple locations in the same area that would satisfy the same trip purpose. IF that is the case, it relies simply on the accuracy of GPS points on the roadway network and if the roadway links match up perfectly. If they do, the selection will be made on which location was most recently visited. The Origin model selects the location based on where the trip started and can better choose a location based on the origin location.

$$P(v_i = l | u_i = u, t_i = T, v_{i-1} = l_k) = \frac{\sum \{N_c, u_i | N_c \in N, u_i \in U, N_c = N_r, v_{r-1} = l_k\}}{\sum \{N_r, u_i | v_{r-1} = l_k\}}$$

Using the same example as above, the model will load all previous trip routes that were previously taken from the origin location (home). Based on the current route, it will select the route which most matches it from the training trips. A low accuracy location can be selected if there are a small number of trips from a given location. If there is only one trip from an origin, then it will always select the only origin matching that origin. This type of model requires a very large dataset of learned trips.

7.4.3 None:

This is the simplest case. The model will always select the destination based on only the route being taken, and selects the route which most matches the current one. This is beneficially due to almost always selected a destination, and having many options available to it The overall

accuracy is likely lower, since many times a single route (such as taking I-495 in the Washington DC area) can cause an over selection of destination which require taking high volume roads. The inclusion of this model allows for a shorter learning period all be it with lower accuracy levels.

Model formulation is below.

$$P(v_i = l | u_i = u, t_i = T, v_{i-1} = l_k) = \frac{\sum \{N_c, u_i | N_c \in N, u_i \in U, N_c = N_r\}}{\sum \{N_r, u_i\}}$$

7.5 Results

For the full breakdown of results on all trip progression, model types, and route match thresholds, please visit Appendix G.

7.5.1 Raw Results

The results for this chapter are broken down into 3 categories: The algorithm running with the trip purpose classification, trip origin location, and in-route prediction considering all trips and no classification. These three runs are then broken down by percentage of the trip that can be considered before a prediction is made. The baseline start of trip model basically is a 0% in-route trip run. In addition, the model is run with 25%, 50%, 75%, and 90% of the trip completed. Keep in mind that just because 100% of the trip is complete, the location may not be accurately predicted for a number of reasons: 1) the algorithm does not know the trip is complete, and may be predicting a location further away along the same route. And 2) the person may have never visited that location before thus not having the destination saved in the database. The model will never predict a location correctly the first time the user arrives at a location.

Below is the accuracy of each of the three prediction methods for 25, 50, 75, and 90% of the trip duration. The x-axis shows the amount of the trip that has progressed, followed by the amount of the links that need to match in order for a selection to be made. For instance, “0.25-.1” has the meaning: a destination was predicted 25% of the way through the trip, with a minimum threshold of destination prediction at 10% route match. The word “All” in the second portion of the variable name denotes that all routes are able to be selected, and no minimum threshold is necessary. The reason why the accuracy tends to be lower for these predictions is due to many

destinations being selected even though there is a very small route match to a destination that has a low chance of being selected.

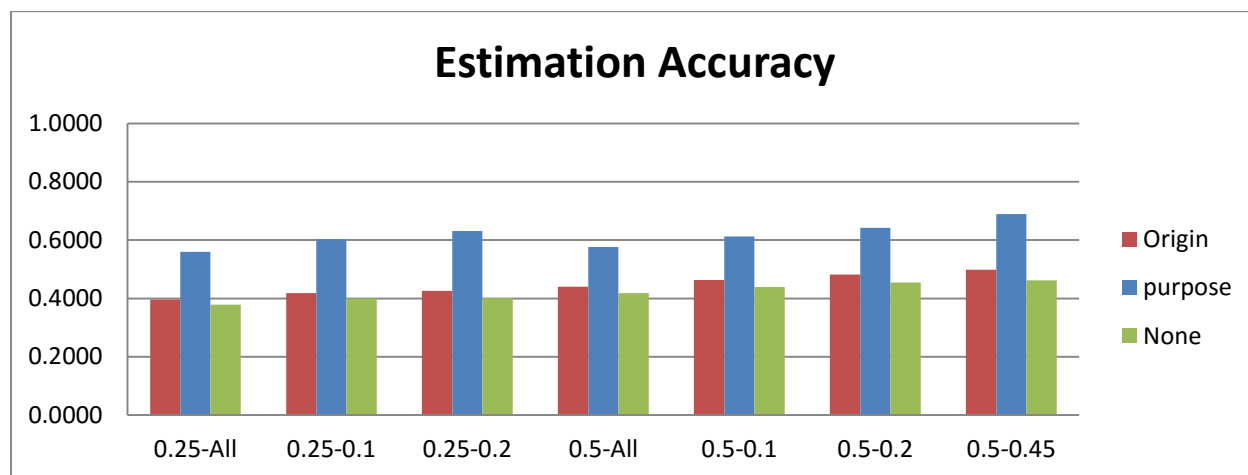


Figure 38- Estimation Accuracy 25, and 50% trip progression

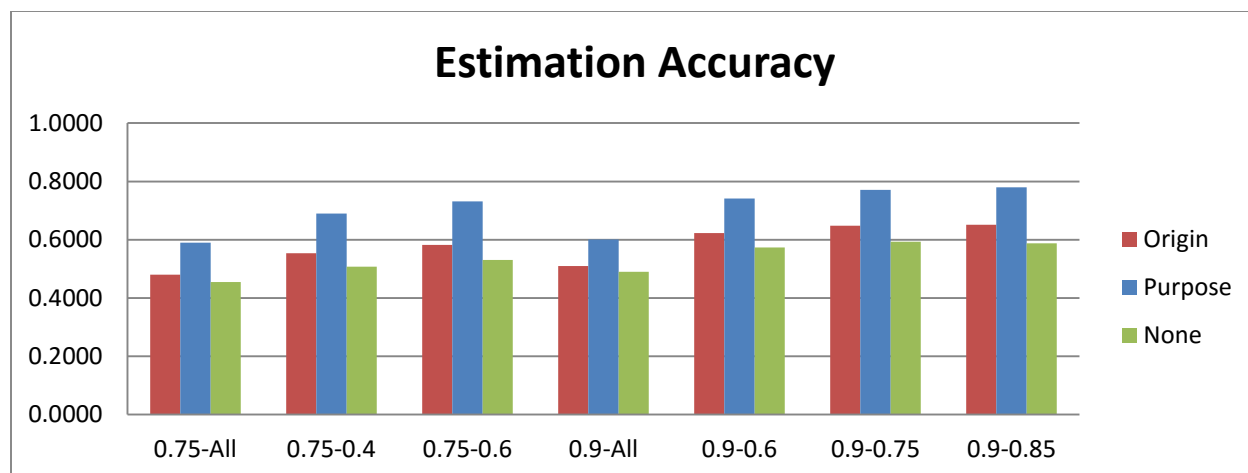


Figure 39- Estimation Accuracy, 75, and 90% trip progression

The accuracy of these methods are both interesting and expected. The highest accuracy model is the one with trip purpose classification. At over 75% accuracy, this method seems to be a feasible real world application. Comparing this accuracy to that with origin classification, it is only 53% accurate when calculated at 50% of a trip. Comparing this result to the start of trip model, they are nearly identical. Since route information is more unique and varied than OD information, more variety is needed for a learning database.

From this increased accuracy, there are negative consequences to the use of the trip purpose and origin classifications.

Below is the number of estimations made by trip progression and route match threshold. An understandable trend occurs: the higher the trip progression, the more estimations are being made. While a higher match threshold is necessary to make a destination estimation, the increased accuracy pays the price of fewer predictions.

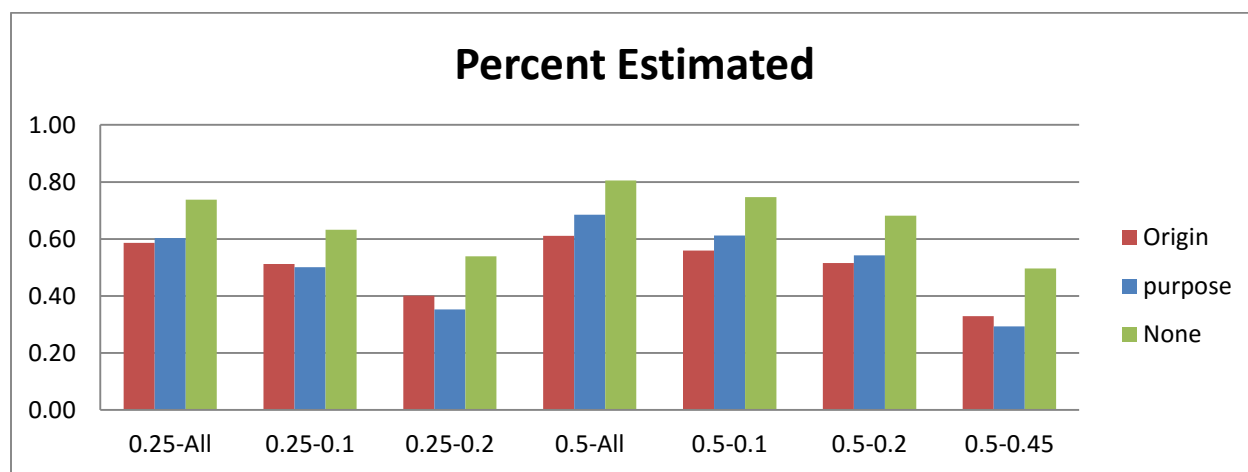


Figure 40- Trips estimated; 25 and 50%

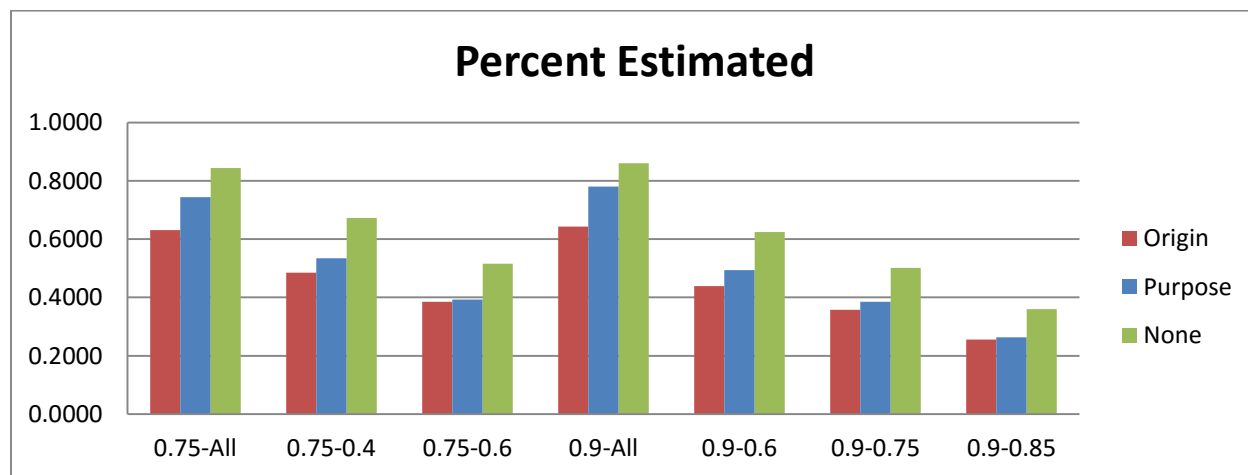


Figure 41- Trips estimated; 75 and 90%

The highest accuracies come from the model when the percent of trip match is at its highest 85%+, but since so few trips have this match level, the percent of estimations drops to below 40%. For the final model, the algorithm will have to use the high accuracy predictions where possible, and less accurate route matches when they are not. The next section will show the overall accuracy of the in route model under two conditions: with trip purpose, and without. The model will also be compared to previous best case models found in the literature review.

7.5.2 Best Model Creation

The best model was created by taking those aspects of each of the 3 model classifications shown in the previous results section: Trip Purpose, Origin, and “None”. When available, the selection will be made based upon the trip purpose, if not routes meet the selection criteria, the model moves onto Origin, if still no routes meet the criteria, then onto “None”. Finally, if there are no routes in the entire system that meet the first 3 selection criteria, the model selects a destination based on the start of trip model explored in Chapter 7 of this report.

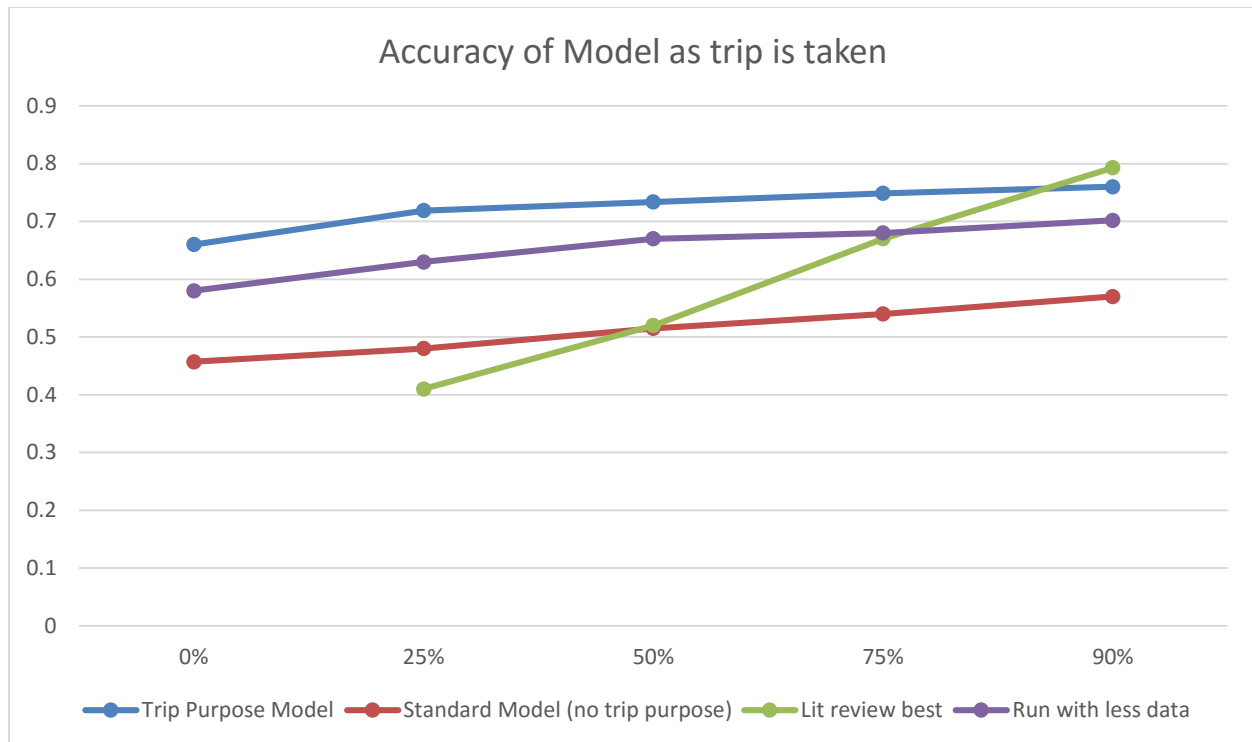


Figure 42- Model accuracy comparison with Alvarez-Garcia model

Another interesting result is how good the start of trip model does in comparison to the in-route model that does not classify route by trip purpose. Without trip purpose, the in-route model does not surpass the start of trip model in chapter 6 until 75% of the trip has been undertaken. This is both interesting, and a testament to the accuracy of the final model of chapter 6. At 50% of the trip, even by considering the most similar trip route, selecting the destination at the start of the trip based on time and trip purpose is even more accurate.

In answering the question: “why does the model accuracy not increase at such a rapid rate”, there are two possible answers:

- 1) The start of trip model accuracy is already rather high. There isn’t much place to go, and the total accuracy maximizes at about 80% because some trips (about 20) seem to be too difficult to predict: like which store at the mall do you plan to go to? Which house are

you stopping at in a neighborhood? The accuracy of the model may not be able to get much higher than 80%.

- 2) Because the predictions are being made by trip purpose, and trip purpose is estimated at the start of the trip, the purpose is most useful at the beginning of the trip. To incorporate in-route trip purpose prediction would require a more advanced model that runs every minute, then ties back into the location prediction model with every GPS point. This needs to be studied further for application. Currently the software used for this dissertation (WEKA) does not seem to have this capability.

7.5.3 Decreased Accuracy over time

Similar to the results shown in chapter 6, there is a non-insignificant drop in accuracy as the model continues. This is proof that the learning procedure is growing old, and needs to emphasize more learned new information and forget old information in classifying trip purpose.

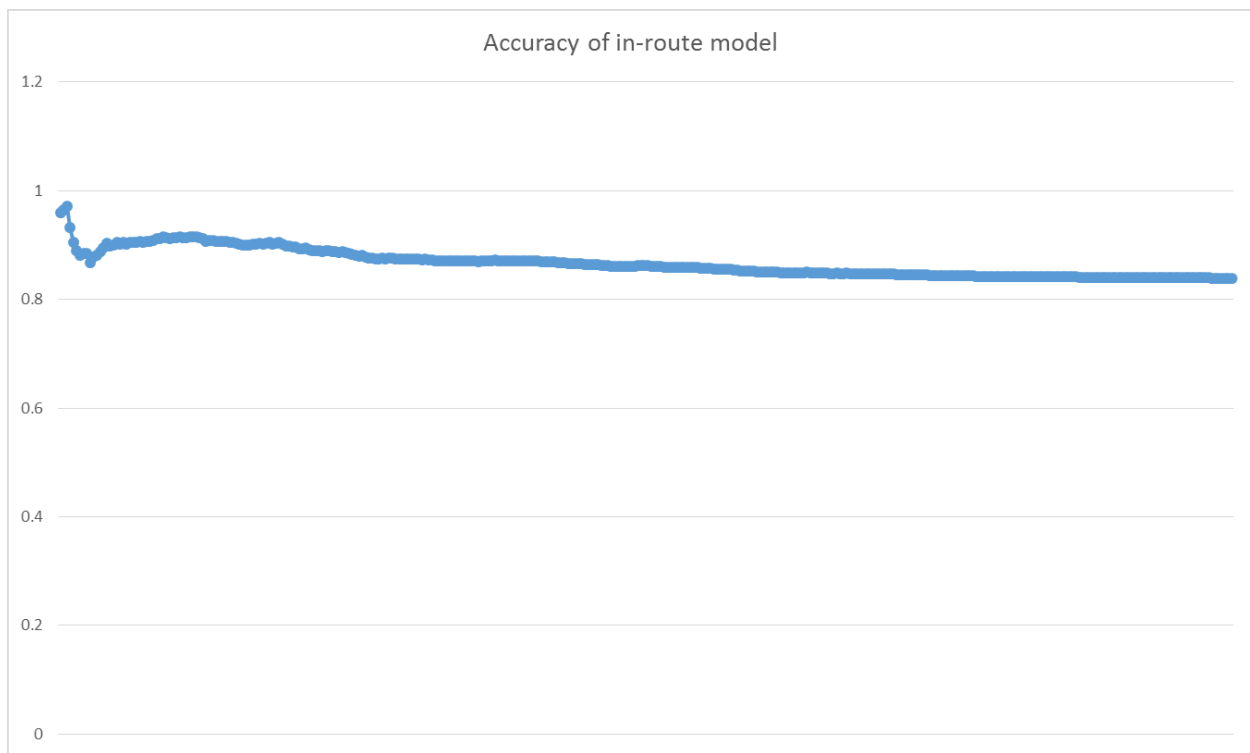


Figure 43- Accuracy of in-route model 90% trip completion, .85 route match

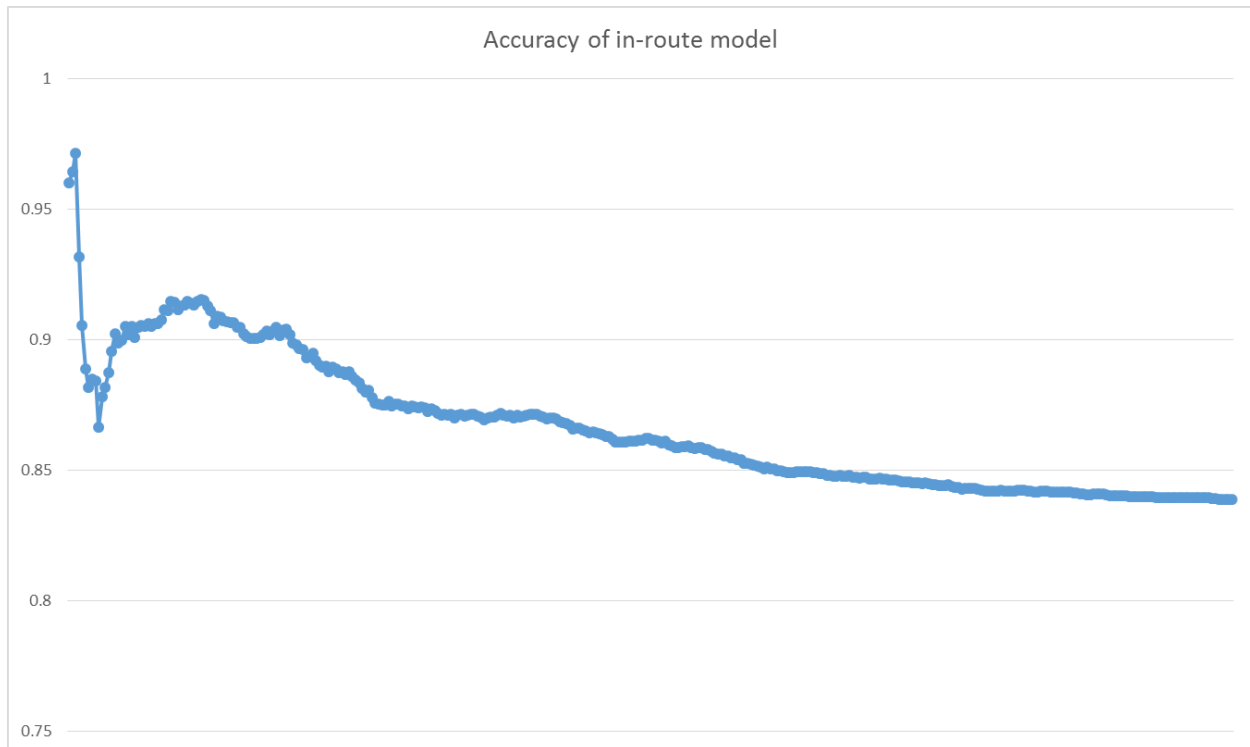


Figure 44- Accuracy of in-route model 90% trip completion, .85 route match (trips 1-30)

7.5.4 Accuracy by Trip Purpose

The accuracy of the in-route model by trip purpose is shown below. Similar to the start of trip model, the work and home based trips are the most accurate. The accuracy of the Work based trips is now nearly 100%. This is largely to do with work based trips being routine, and generally non-unique. With similar routes being taken on multiple occurrences, the destination becomes easier to estimate. In all, the accuracy of all trip types increases from the in-route model. Even lower trip purpose types such as *Shopping* and *Social* reach about a 50% destination accuracy.

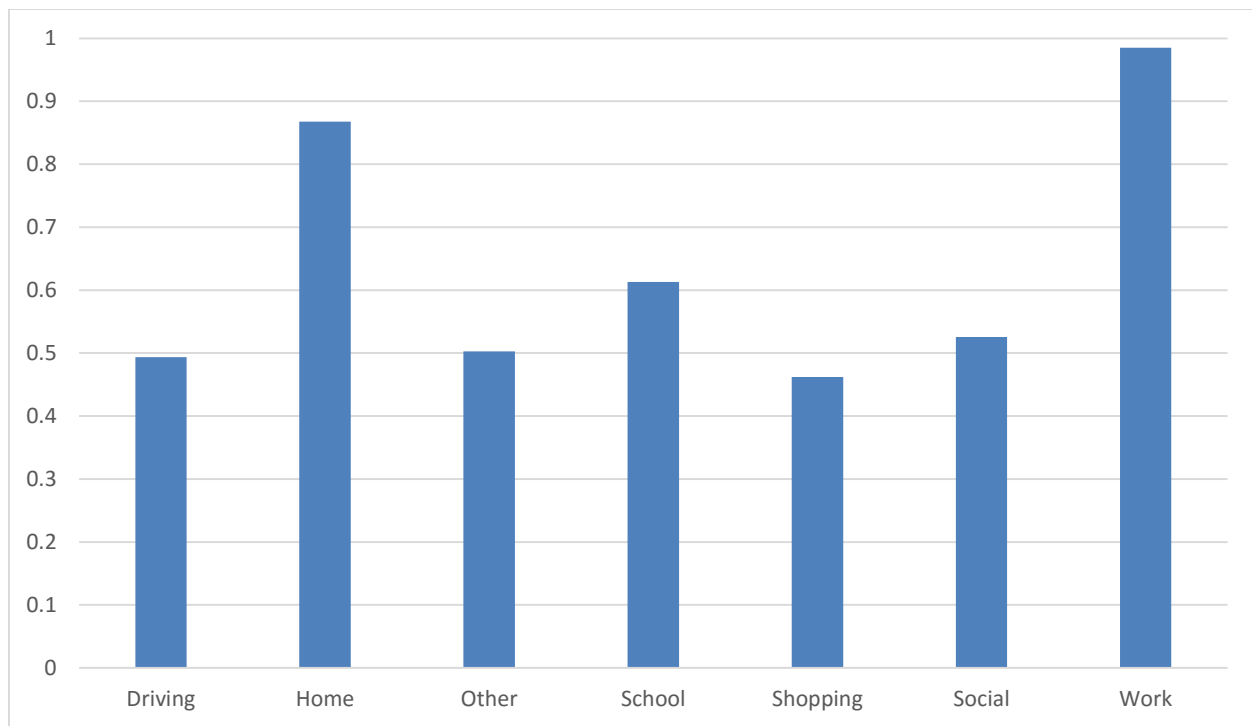


Figure 45- Accuracy by Trip Purpose (in-route model)

Table 11- Accuracy of Trip Purpose (exact values)

Driving	0.493
Home	0.867
Other	0.503
School	0.613
Shopping	0.462
Social	0.526
Work	0.985

7.6 Prediction Validation

For this idea to be fully implemented into a connected vehicle, or in vehicle navigation unit, more reliable estimations need to be given. While 52% prediction accuracy at start of trip, followed by 60% accuracy $\frac{1}{4}$ of the way through the trip, 71% accuracy half way through the trip sounds like good results, we do not want our GPS getting our destinations wrong 30% of the time. This is why prediction validation can be useful. Using the process developed in chapter 7, it is easier to see when multiple destinations have a relatively similar chance of occurring. For instance, if the start of trip model gives a prediction that is becoming less likely to occur based on the results of the in-route model, a confidence level can be displayed, saying that there is conflicting information from the two sets of the model. Conversely, if the start of trip model and the in-route model deliver the same prediction, the chances the prediction is correct is likely to be very high. Some confidence level may be displayed when the route links up to the trip purpose, and origin. If the route from the origin, purpose of the trip and time of day all match up, it could be very likely that a single trip confidence level would be above 95%. This type of work can be done in future papers.

7.7 Conclusions

The inclusion of trip purpose in in-route destination prediction has a clear benefit in terms of accuracy. By simply removing those trips from the database that do not satisfy the same trip purpose, the accuracy of the model increases by 15%. This is a rather remarkable result, as the percentage trip route match is higher for the trips not matching the same trip purpose. Although some trips originate from the same location and take the same route, it is still a better option by 15% accuracy to select based on the purpose of the trip and most accurate route. While the model formulation used for predicting destination may not be the most advanced in terms of overall

accuracy (possibly best for quickness of reasonable predictions), the addition of trip purpose in existing modeling frameworks of in-route models could increase accuracy by 15% near the start of trip (0-50% of trip progression). This addition can make the field of destination prediction feasible and implementable.

CHAPTER 8: Socio-Economic Prediction through GPS Travel Data

8.1 Introduction

Thus far all demographic and economic information has been used as an input to predict trip purpose and trip destinations, but I believe it would be feasible to flip the modelling framework to predict socio-economic factors and use location as the input variables. Application of this research branch may have far reaching applications from transportation engineering. Take for instance the case of collecting census data simply by reviewing travel patterns; trips from home to school could be used to compute driver age, while visiting different type of commercial land could impact income. The applications are wide ranging and deserve study. Thorough methodologies will be presented in future work.

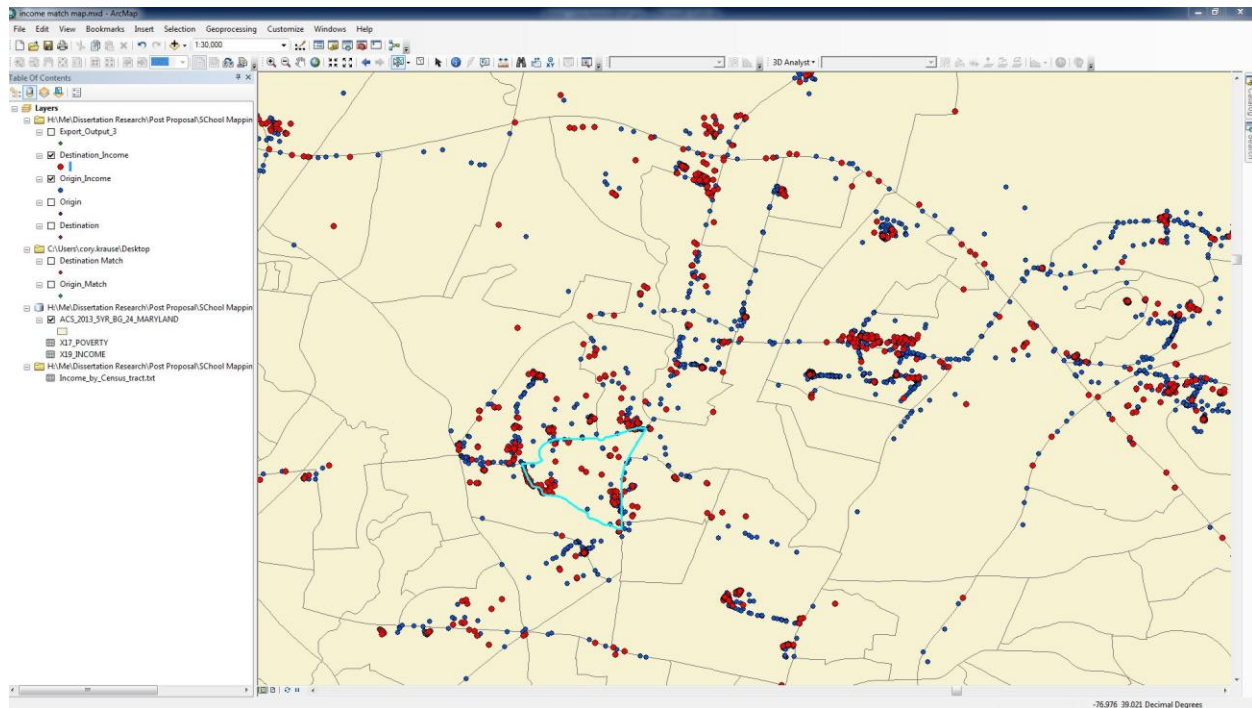
8.2 Data

The GPS and land-use data from previous sections was used and combined with Census data to get accurate socio-economic variables of the surrounding variables as trips were being recorded. The first step was to test the Census data in relation to each census tract to make sure that the data was reasonable and error free over the entire data set. Below a covariance was conducted over the first 1000 geolocations of the state of Maryland. Again, similarly to the land use data, only the locations inside the state of Maryland were used. This may lead to a difficulty in identifying socioeconomics for individuals who often travel outside of the state, but since all of the individual's home location is in the state of Maryland, the effects should be rather minor.

The data was combined via a geographical match of census tract to each gps point in ArcGIS.

With each census tract ID, the variables were linked via an Access Database. The full variable

list can be found on the Census Site: <http://www2.census.gov/geo/docs/maps-data/data/tiger/prejoined/ACSMetadata2011.txt>



8.3 Methodology

8.3.1 Variables

The first step was to determine what variables to include in the modelling of socio-economics.

With data received from the US Census, and the geomapping to trip locations, all trips are now associated to the area of the start or end of the trip. To determine which variables to keep, first a scaled-zero lag covariance was conducted on 100 geolocations. It was necessary to first make sure that the data made sense, and that I would not include variables in the final model that has a high covariance with another variable.

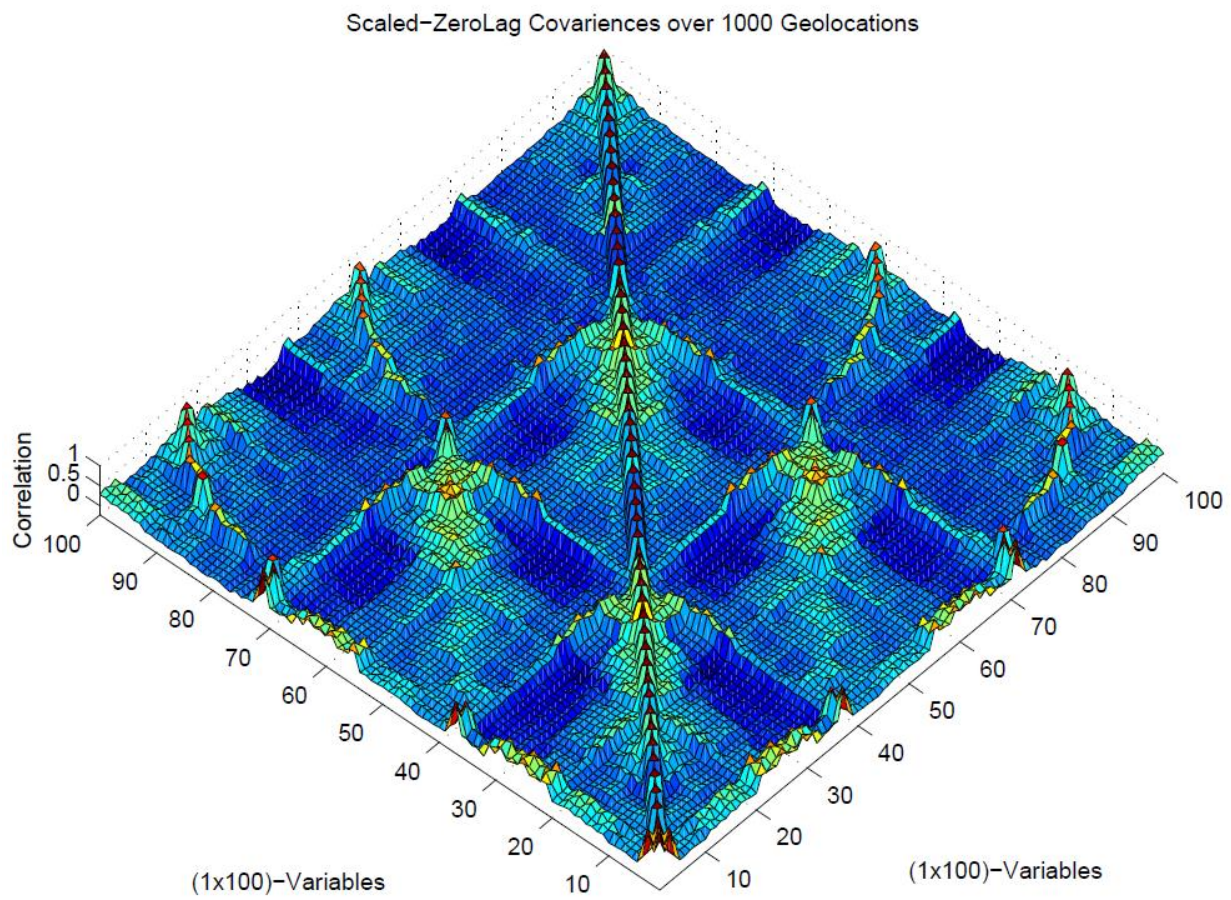


Figure 46; Scaled ZeroLag Covariances over census tract locations. 100 income variables

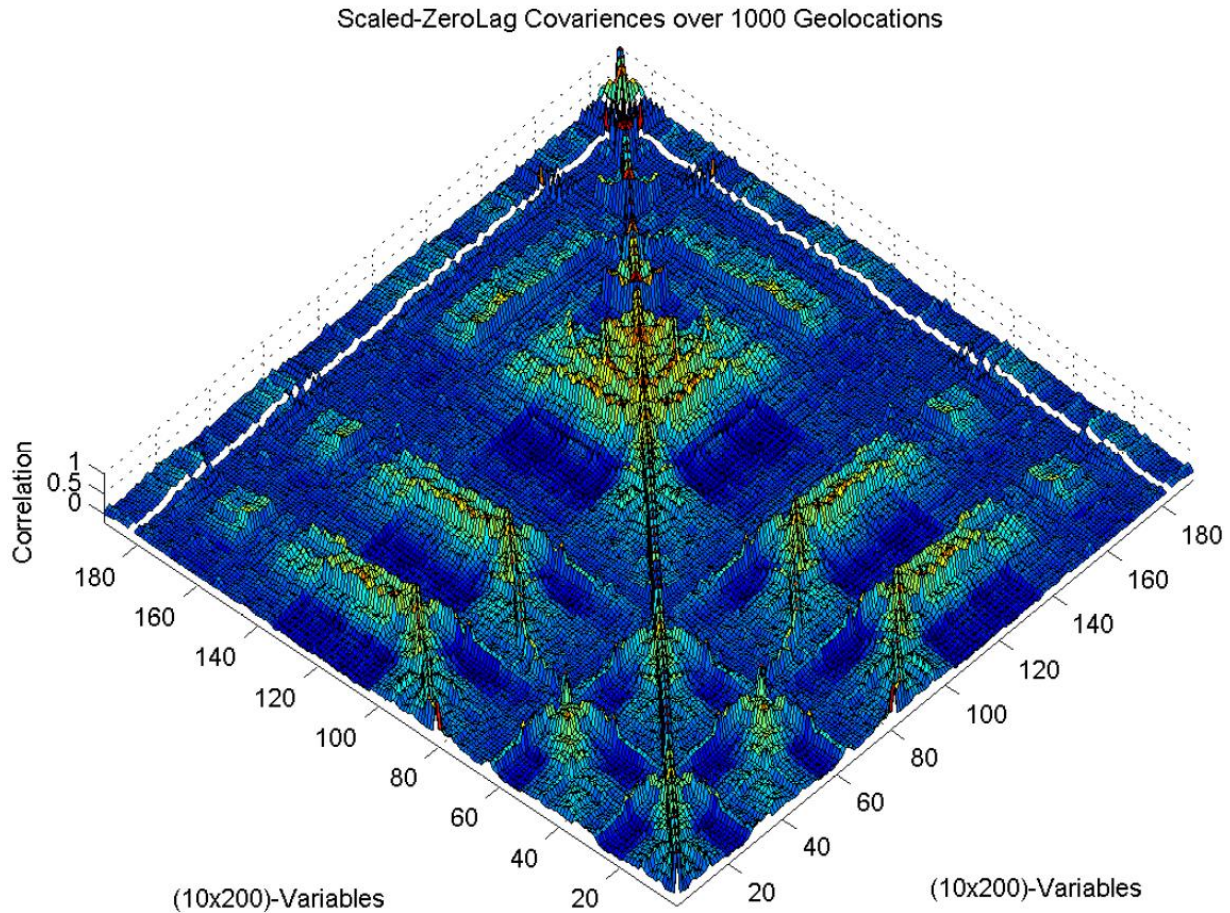


Figure 47- Scaled-ZeroLag Covariances over Census tract locations for 100 employment variables.

Certain patterns can be seen in the covariance matrices. There are large boxes that correspond to the variable groupings, for instance the group of questions asking about income. Each box of covariances basically asks the question: how similarly do areas in which people have the same income, interact with their environment? Or how are driving patterns similar or different between areas that have a certain degree difference in income?

This helped pass the test of making sure that data is reasonable. Some of the findings were:

- Locations with a high population of people between the ages of 18-25 tend to have a higher education level. Meaning the covariance matrix was able to find Universities, or more likely, the University of Maryland, which has a high population of younger people.
- Locations with houses with a high number of bedrooms correlates more closely to income than does home price. This means that it is more realistic to compare income with the number of bedrooms than the price of the home (in the state of Maryland)

These sorts of tests can help to find areas that have similar responses in their values to socio-economic factors. The final test is to determine if matching these locations amongst themselves to the particular individual's travel patterns will help to lead to a more narrowed ranged of socio-economic characteristics.

The variable list that was created to estimate income is shown below: These variables were chosen due to their relation with determining the income of a person in the area; dealing with poverty, raw income, houses in the area for rent, amount of people employed in the area, and average income.

Variable list:

- Income
- Poverty Status (number of households under the poverty line) at trip origin location
- Poverty Status (number of households under the poverty line) at trip destination location
- Dwellings for rent (number of houses up for rent) at trip origin location
- Dwellings for rent (number of houses up for rent) at trip destination location
- Number of employed people at origin

- Number of employed people at destination
- Average income in the area (origin)
- Average income in the area (destination)
- Population

Each of the variables that are not an average are also divided by the population of that area. This way, the houses for rent for instance are not higher simply because the census tract has more inhabitants. Population may not be directly related to income, but was included to find out the significance of each of these variables.

8.4 Linear Regression

The first model created to test whether socio-economic factors could be estimated through GPS points is a simple linear regression. By running one model with only the information known through the survey the participants filled out, and another using travel information, it is possible to see if this additional information can be helpful, and if so, how helpful is it?

8.4.1 Survey Data

SUMMARY OUTPUT

<i>Regression Statistics</i>	
Multiple R	0.567024366
R Square	0.321516632
Adjusted R Square	0.284508448
Standard Error	45574.75032
Observations	117

<i>ANOVA</i>					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	6	1.08E+11	1.8E+10	8.687717	9.66E-08
Residual	110	2.28E+11	2.08E+09		
Total	116	3.37E+11			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	76558.42334	24230.69	3.159565	0.002041	28538.89	124578	28538.89	124578
high school	43101.26909	19876.52	-2.16845	0.032278	-82491.9	-3710.68	-82491.9	-3710.68
bachelors	3436.696705	17808.61	0.19298	0.847331	-31855.8	38729.18	-31855.8	38729.18
masters	42098.98864	17632.06	2.387638	0.018662	7156.376	77041.6	7156.376	77041.6
male	5528.398068	15343.18	0.360316	0.719302	-24878.2	35934.99	-24878.2	35934.99
female	26725.69779	13259.81	-2.01554	0.046286	-53003.5	-447.86	-53003.5	-447.86
age	764.1237488	353.9715	2.158715	0.033046	62.63529	1465.612	62.63529	1465.612

8.4.2 Trip Data

<i>Regression Statistics</i>	
Multiple R	0.346391
R Square	0.119987
Adjusted R Square	-0.01805
Standard Error	54090.68
Observations	119

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	16	4.07E+10	2.54E+09	0.869209	0.605394
Residual	102	2.98E+11	2.93E+09		
Total	118	3.39E+11			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>
Intercept	108279.1	34227.35	3.163526	0.002054	40389.33	176168.9	40389.33	176168.9
poverty per person _O	105682.6	219349.4	0.4818	0.63098	-329396	540761	-329396	540761
vacant houses per person in labor force_o	258922.7	576845.7	0.44886	0.654485	-885248	1403093	-885248	1403093
Area Income Origin	-40888.4	68110.81	-0.60032	0.549623	-175986	94209.08	-175986	94209.08
BusinessUnit	0.386665	0.289771	1.334383	0.18505	-0.18809	0.961425	-0.18809	0.961425
GOV/PUB/Service	-215031	205157.3	-1.04813	0.297057	-621960	191897.5	-621960	191897.5
Restaurant	30669.62	55056.94	0.557053	0.578712	-78535.5	139874.8	-78535.5	139874.8
Undeveloped	85415.89	123899.6	0.689396	0.492139	-160338	331170.1	-160338	331170.1
Shops	146809.3	93936.93	1.56285	0.121185	-39514.2	333132.8	-39514.2	333132.8
Mixed_Use	419670.7	328101.3	1.279089	0.203768	-231117	1070458	-231117	1070458
RecreationSite	-39495	97126.79	-0.40663	0.68513	-232145	153155.6	-232145	153155.6
	-90383.7	102685.8	-0.8802	0.380822	-294060	113293.1	-294060	113293.1

Industrial	-77845.1	107077.3	-0.727	0.468891	-290232	134542.2	-290232	134542.2
Leisure	-1830.7	43475.85	-0.04211	0.966495	-88064.8	84403.44	-88064.8	84403.44
Institutional	3899.572	87682.02	0.044474	0.964614	-170017	177816.5	-170017	177816.5
Commercial	-76861	86866.58	-0.88482	0.378337	-249160	95438.5	-249160	95438.5
Residential	-40299.1	30856.09	-1.30603	0.194479	-101502	20903.82	-101502	20903.82

The survey data regression model is superior in every category: even with using fewer variables, it has significantly improved explanatory power. There are no statistically significant variables in the trip data regression model. Simply put, the regression model is not advanced enough to catch the nuance of the travel data. To run the data in the regression model, all travel data must be averaged and those average values used for the regression. For the survey data model, age, education, and gender were statistically significant. It will likely be difficult to beat the accuracy of this model with the machine learning techniques, but if this data is unknown to the researcher, and not possible to obtain, using the trip data may prove useful.

8.5 Machine Learning

Machine learning allows for a better taste of why one model is performing better than others, or if the rules derived by the algorithm tend to make sense. The usefulness of travel behavior data for determining socio-economics can be done by creating 2 different data sets: one with only the home location, and survey information included, and one with all the user's travel behavior recorded as individual variables. The variables were created to best get an idea of the surrounding environment of the user, and then averaged to find what the normal circumstances for the user was as they travel to different locations over the 70 day period. The model learning set were all trips for all users over the 70 day period. A decision tree structure was created with the J48 machine learning algorithm to maximize the correct predictions of the nominal variable Income. Once this model was created, it was tested on the test set, which was the average of all the user's trip variables (identical to the variables used in the learning set).

8.5.1 Trip data

=== Run information ===


```

Scheme:weka.classifiers.trees.J48
Relation:  nolanduse_create2
Instances:  20651
Attributes:  6
    income
    pop
    labor
    vacant
    poverty
    income_o
Time taken to build model: 1.84 seconds

```

```

=== Evaluation on test set ===
=== Summary ===

```

Correctly Classified Instances	33	27.7311 %
Incorrectly Classified Instances	86	72.2689 %
Kappa statistic	0.1329	
Mean absolute error	0.2173	
Root mean squared error	0.4201	
Relative absolute error	90.7797 %	
Root relative squared error	121.5587 %	
Total Number of Instances	119	

```

=== Confusion Matrix ===

```

```

a b c d e f g  <-- classified as
9 1 5 2 6 4 1 | a = 200K
4 8 1 2 3 1 0 | b = 150K
6 1 7 0 2 2 0 | c = 125K
6 3 2 3 2 2 1 | d = 100K
2 3 1 2 5 2 3 | e = 75K
4 2 2 3 1 1 0 | f = 50K
0 2 0 0 2 0 0 | g = 25K

```

8.5.2 Survey Data

```

=== Run information ===

```

```

Scheme:weka.classifiers.trees.J48
Relation:  survey_create
Instances:  20651
Attributes:  6
    income
    gender
    Masters

```

age
Bachelors
Highschool

Time taken to build model: 1.52 seconds

=== Summary ===

Correctly Classified Instances	25	21.3675 %
Incorrectly Classified Instances	92	78.6325 %
Kappa statistic	0.0276	
Mean absolute error	0.2323	
Root mean squared error	0.353	
Relative absolute error	97.1894 %	
Root relative squared error	102.1328 %	
Total Number of Instances	117	

=== Confusion Matrix ===

```
a b c d e f g <-- classified as
17 6 2 1 2 0 0 | a = 200K
12 4 0 2 1 0 0 | b = 150K
13 0 0 2 1 1 0 | c = 125K
6 8 1 0 1 3 0 | d = 100K
8 4 0 0 0 5 0 | e = 75K
1 4 1 0 3 4 0 | f = 50K
0 1 0 0 1 2 0 | g = 25K
```

The model accuracy is 28%, or the correct prediction of the income of the single individual on 28% of predictions made on the decision tree. The methodology is not a very straightforward approach for determining the income of an individual: first census data needs to be found, variables collected to match the socio-economic factor that is being searched, the GPS travel data is collected (assuming user agreement), data collected over 2 months, linking the two datasets, creating and running a learning set, then running the aggregate test set over the newly created learning set. While many steps of intensive data management are needed, it may become a future application in the field of advertisements with the rising agreement of users to be tracked, and lack of involvement in traditional surveys that give out sensitive information such as marital status, education level, age, income, etc. With something as simple as GPS location, applications like facebook, google or foursquare can create relatively accurate descriptors of individual users without their knowledge. Seeing as this GPS travel data is granted freely and legally, these databases with machine learned socio-economic data may be sold or used for the company's benefit.

8.6 Conclusions

The easiest and most accurate way to estimate an individual's income is by asking simple socio-economic variable questions. These include education, age, marital status, and home location. Many times, these pieces of information are either not available, or refused by the participant. With the advent of smartphones, and location based services, participant location can be easier to get a hold of than simple survey questions; many location based services come preloaded and agreed upon by default (Facebook, twitter, foursquare). Using only the location of the user, the individual's income can be estimated with 30% accuracy to within 12,500\$ (about twice as accurate as a random guess). While this accuracy is not very high, it is an interesting result from

the new location based services age. The field of location based socio-economics prediction could soon begin, with companies basing their advertisements on the last place that was travelled, or the average size of house in your area (all openly available Census data for all Census tracts in America). By combining this data with open source land use databases such as OpenStreetMap, it would not be difficult to determine how often an individual travels to their local McDonald's, where they live, how much money they make, and when they will be approaching a McDonald's outside their normal travel behavior. The future may see targeted advertisements based solely on travel patterns linked to open-source databases and advanced models.

CHAPTER 9: Conclusions

9.1 Realistic Implementation

The one major concern over this research is that of realism. Can the approaches defined in this paper truly be implemented in a real world environment? While much effort was put into making the experimental runs as pseudo-real-time as possible. A discussion of the time and computational requirements should be done. There are three main parts to the destination prediction suite of algorithms derived for this dissertation:

- 1) real time trip purpose derivation with historic travel patterns
- 2) Start of trip destination prediction
- 3) In-route trip route and destination prediction

On an i5-4590 Intel CPU with 8G rams computer, each of the three cases take approximately 8 hours (for 30 trip purpose learning set) , 10 hours (start of trip model), and 23 hours (90% trip completion, no restriction in-route trip route and destination prediction model). Run from start to finish the entire prediction process takes about 41 hours. While 41 hours sounds like a long time to run, this is for over 20,000 trips between 120 participants. Split between each of the participants the run time would be 21 minutes. Split between each trip that the individual is taking, the run time would be about 7 seconds. It is certainly realistic for the methodology explained in this dissertation to be implemented, as a destination prediction could be derived between the time the engine starts up (giving power to the onboard computer) and the vehicle begins moving.

9.2 Contributions to the Field

The dissertation has explored many concepts in the destination prediction field. Among the most important contributions that this research has provided is a method for the inclusion of trip purpose prediction in real time, with reasonable reliability to instantly raise the accuracy of both the start of trip model and the in-route prediction model. Through the use of machine learning and the combination of point of interest and land use data, the new approach can be a significant addition to any existing framework. Also, this data was collected in 2011 using relatively cheap GPS recorders. Technology continues to advance, giving more accurate GPS location, at lower cost with improved data storage. The ability to more accurately predict trip purpose and hence the location of destinations will continue to rise.

Additional to the derivation of trip purpose using no trip destination location, the advent of the tiered Markov model structure allows for a low learning period for good accuracy results. In the start of trip model, the accuracy increases to over 50% in less than 30 days, beating out older models that take over 300 days to learn. By including this modelling structure in the in-route procedure, similar learning times are needed, and an increase of accuracy between 15 and 30% can be found. Truly the inclusion of a Tiered Structure allows the model to use the variables that are available to it at the time of prediction, but fall back in lower accuracy models when they are not.

These two aspects of the overall research has made short term travel behavior prediction a more realistic and applicable field in transportation engineering.

9.3 Applications

This dissertation has covered the technical aspects of improving a destination prediction system. There is a new research field that has arisen from plug in hybrid electric vehicles: designing an efficient energy management system for the drive ahead.

Destination prediction will play a key role in this research. Basically the information that is needed is the upcoming slope, elevation, acceleration of the trip that is about to be taken. By taking into account these roadway characteristics, the energy management system decides when to use the internal combustion engine, and when to use electric power (Qi et al., 2015). The current research approach is to estimate the future roadway conditions based on current trip characteristics. But, by knowing the destination and route that is about to be taken, and much more knowledgeable energy management plan can be derived that actually knows the gradient of the route about to be taken. By saving gasoline on trips that are known to be less than the battery life of the car, significant energy can be saved. In the future, your car may know where the hybrid electric car is traveling based on destination prediction algorithms, and it can make an intelligent, informed decision on when to use gas and when to use electric power based upon the known route. Current research yields about a 12% fuel savings in this research field. With known destination locations, the fuel savings for hybrid electric cars could significantly improve and even pay for the inboard computer necessary for the computationally intensive destination prediction algorithms.

Below, further applications in personal, private, and government industries are listed.

9.3.1 Personal

A few applications instantly comes to mind including giving advanced information to the driver about changing travel patterns at the predicted destination before it occurs. This can be done via an on-board piece of equipment installed in the participant's vehicle, or something as simple as a smartphone kept on person at all times. It would predict your future location at the start of trip, give more accurate information as the participant is driving, and even alert the driver if roadway

conditions change near the predicted destination. With accuracies nearing 95% for home and work based trips, the technology could truly become a realistic and helpful application.

9.3.2 Private Sector / Advertising

Private sector companies such as Google and Facebook would have much to gain from knowing the location that the user is about to travel to and the route that they will take to get there. Using the knowledge of nearby locations to the travel, selected advertisements can be made for locations that the user is about to be near. The advertisers would certainly be interested in knowing that their advertisements are going to drivers that are about to travel near one of their stores. This could be a new field in advertising based on predicted destination location.

Private GPS device companies may also be interested in the technology developed in this dissertation by building the next most advanced GPS device for installation in a personal vehicle. By making the predictions for the driver and tying in real road congestion information, it would greatly reduce the involvement from the driver, and make for a better piece of technology. Vehicle manufacturers may want to tie in an onboard computer that stores GPS data in the vehicle and no information needs to be transmitted to the infrastructure. For consumers who are concerned with protecting their privacy, this could be a great feature for vehicle manufacturers.

9.3.3 Government Services/ Large Scale Connectivity

By sharing much of the prediction information with Government services such as FHWA, there are many applications that could be deployed in the large scale. Real-time Origin Destination Prediction would be greatly improved by the advent of accurate short term travel behavior algorithms. With more accurate Origin/Destinations in real time, agencies would be more able to determine congested areas in real-time or possibly before they even occur.

In addition to Origin/Destination determination, advanced traveler information can be given to drivers on the roadway either through a Vehicle to Infrastructure based technology, or by roadside variable message signs at specific locations. The usefulness of a new technology that can predict travel behavior before they occur is vast, and the applications that can be put in place to take advantage of it are limitless.

9.4 Closing

This dissertation has aimed at predicting the future location of a traveling vehicle through GPS data, land use information, points of interest, and Census data. Through the use of a new methodology that employs information as it becomes available, the speed of prediction improvement is better than existing modeling frameworks. The greatest advancement comes from the inclusion of a machine learning model of trip purpose before the trip takes place. By correctly identifying the purpose of a trip before it starts, the accuracy improves over previously start of trip models by upwards of 10%. By applying this methodology to an in-route model that constantly updates the probabilities of all given locations, the trip purpose model is able to only consider those viable locations that will achieve the predicted trip purpose. Through these applications it has become feasible to create a working pseudo-real-time application that can be interfaced with local DOT's to use predicted travel to truly impact roadway conditions in real time. Finally, by using travel behavior information, it has been shown that socio-economics can be derived more accurately than with simple survey data itself. The initial application may be in locational advertisements based on socio-economics and intended travel destinations.

While this marks the end of the dissertation, future work is still to be considered. It can be seen in chapter 8 that the in-route model loses some accuracy on its competition at the end of the trip. This

may be due to an incorrect trip purpose prediction at the start of the trip. To solve this, a more accurate trip purpose could be attempted, or a system that discounts trip purpose the farther along a trip is. For example, after 30 minutes, with constantly updated information from the roadway, consider less that the trip has been described as “driving” type, and leave it open to recreation, or shopping. Also, a new method of in-route trip purpose derivation could be explored. Instead of simply deriving trip purpose at the start of the trip, have the machine learning model learn from past trip behavior in-route, that way as more information is gained, the trip purpose is more accurate. While this idea is conceivable, the amount of data that it would require would be immense. Due to the computational power and time required, it may be a difficult undertaking for a real world application.

This field of work is still in it’s infancy, and while the overall accuracy of the model does not exceed 75% when a trip is half finished, by studying further advances, it may be a prime factor in determining roadway conditions in 5-10 years. It should be noted that the study of connected vehicle technologies would be a driving force in the implementation of in-route vehicle prediction. Along with the increase in data availability, the aggregation of large datasets will allow for synergies that do not currently exist (platooning vehicles, vehicles that are destined for the same location, etc....). This is an exciting and advancing field that will likely expand in the coming years.

APPENDIX

Appendix A: Data processing code (VBA)

Raw data to trip format:

```
Sub check()
Dim original As Worksheet
Dim a As Long
Dim b As Integer
Dim OD As Worksheet
Set original = ActiveWorkbook.Sheets("Input")
Set OD = ActiveWorkbook.Sheets("Output")
a = 3
b = 2
While a < 50000
    ' If original.Cells(i, 2).Value - 0.001 > original.Cells(i + 1, 2).Value Or original.Cells(i, 2).Value + 0.001 <
original.Cells(i + 1, 2).Value Or original.Cells(i, 4).Value - 0.001 > original.Cells(i + 1, 4).Value Or original.Cells(i,
4).Value + 0.001 < original.Cells(i + 1, 4).Value Then
        If original.Cells(a, 8).Value > 1 Or original.Cells(a - 1, 8).Value > 1 Or original.Cells(a - 2, 8).Value > 1 Then
            OD.Cells(b, 2).Value = original.Cells(a, 2).Value
            OD.Cells(b, 3).Value = original.Cells(a, 4).Value
            OD.Cells(b, 4).Value = original.Cells(a, 6).Value
            OD.Cells(b, 5).Value = original.Cells(a, 7).Value
            OD.Cells(b, 6).Value = original.Cells(a, 8).Value
            OD.Cells(b, 7).Value = original.Cells(a, 9).Value
            OD.Cells(b, 1).Value = original.Cells(a, 13).Value
            b = b + 1
        End If
        a = a + 1
    Wend
'Dim original As Worksheet
Dim i As Integer
Dim j As Long
Dim k As Integer
Dim m As Double
Dim n As Double
Dim p As Integer
Dim q As Integer
Dim Z As Integer
'Dim OD As Worksheet
'Set original = ActiveWorkbook.Sheets("Input")
'Set OD = ActiveWorkbook.Sheets("Output")
Set Var = ActiveWorkbook.Sheets("Set")
lLoop = 1
i = 1
'iteration
```

```

j = 3
'row
k = 0
'#turns
m = 0
'distance
n = 0
'summed distance
p = 0
'number of gps points per trip
q = 0
'number of stops on trip
Z = 2
'sheet 3 counter
While i < 5000
    If Abs(OD.Cells(j, 7).Value - OD.Cells(j - 1, 7).Value) > 90 And Abs(OD.Cells(j, 7).Value - OD.Cells(j - 1, 7).Value) < 270 Then
        k = k + 1
        'check heading at each data point, if greater than 90 degrees, count turn integer'
        End If
    n = n + m
    m = Sqr((OD.Cells(j, 2).Value - OD.Cells(j - 1, 2).Value) ^ 2 + (OD.Cells(j, 3).Value - OD.Cells(j - 1, 3).Value) ^ 2)
    'calculates algebraic distance based upon lat long data'
    OD.Cells(j - 1, 9).Value = m
    'point to point distance'
    If OD.Cells(j - 1, 6).Value < 1 Then
        q = q + 1
        End If
    p = p + 1
    'iterates time counter'
    If Abs(OD.Cells(j, 4).Value - OD.Cells(j - 1, 4).Value) > 200 And OD.Cells(j - 1, 17).Value <> 59 Then
        'if the vehicle stops for more than 2 minutes then...'
        Var.Cells(Z, 2).Value = OD.Cells(j - p, 2)
        Var.Cells(Z, 3).Value = OD.Cells(j - p, 3)
        Var.Cells(Z, 4).Value = OD.Cells(j - 1, 2)
        Var.Cells(Z, 5).Value = OD.Cells(j - 1, 3)
        OD.Cells(j - 1, 8).Value = "Stop"
        'Destination of trip
        OD.Cells(j - 1, 14).Value = p - 1
        'inputs travel time for that trip
        OD.Cells(j, 8).Value = Abs(OD.Cells(j, 4).Value - OD.Cells(j - 1, 4).Value)
        'Soak time from the end of one trip to the start of the other'
        OD.Cells(j - 1, 12).Value = k
        'number of turns over 90 degrees taken on this trip
        OD.Cells(j - 1, 10).Value = n
        'prints summed distances by trip'
        OD.Cells(j - 1, 13).Value = q
        'prints summed stops by trip'
        Var.Cells(Z, 9).Value = OD.Cells(j - 1, 4).Value - OD.Cells(j - p, 4).Value
        'inputs travel time for that trip
        Var.Cells(Z, 6).Value = k
        'number of turns over 90 degrees taken on this trip
        Var.Cells(Z, 7).Value = n

```

```

        'prints summed distances by trip'
        'Var.Cells(Z, 8).Value = q
        'prints summed stops by trip'
        lLoop = lLoop + 1
        k = 0
        n = 0
        m = 0
        p = 0
        q = 0
        'resets counters
        Z = Z + 1
        End If
    j = j + 1
    i = i + 1
    'iterates counter and row'
Wend
End Sub

```

Create unique OD identifiers

```

Sub Trip()

    Dim i As Long
    Dim j As Long
    Dim r As Range
    Dim T As Worksheet
    Dim C As Worksheet
    Set C = ActiveWorkbook.Sheets("Cost")
    Set T = ActiveWorkbook.Sheets("Trip")

    ' Dim Olat As String
    ' Dim Olong As String
    ' Dim DLat As String
    ' Dim DLong As String

    i = 1
    j = 1

    Set r = Range("B2:E122")
    While i < 300
        If C.Cells(i, 27).Value < 1 Then
            While j < 300
                If Abs(r.Cells(i, 1) - r.Cells(i + j, 1)) < 0.003 And Abs(r.Cells(i, 2) - r.Cells(i + j, 2)) < 0.003 And
                Abs(r.Cells(i, 3) - r.Cells(i + j, 3)) < 0.003 And Abs(r.Cells(i, 4) - r.Cells(i + j, 4)) < 0.003 Then
                    C.Cells(i + j + 1, 27).Value = n
                    C.Cells(i + 1, 27).Value = n
                End If

                If Abs(r.Cells(i, 1) - r.Cells(i + j, 3)) < 0.003 And Abs(r.Cells(i, 2) - r.Cells(i + j, 4)) < 0.003 And
                Abs(r.Cells(i, 3) - r.Cells(i + j, 1)) < 0.003 And Abs(r.Cells(i, 4) - r.Cells(i + j, 2)) < 0.003 Then
                    C.Cells(i + 1, 27).Value = n
                    C.Cells(i + j + 1, 27).Value = n
                End If
            End While
        End If
    End While

```

```

        If Abs(r.Cells(i, 1) - r.Cells(i + 2, 3)) < 0.003 And Abs(r.Cells(i, 2) - r.Cells(i + 2, 4)) < 0.003 And
Abs(r.Cells(i, 3) - r.Cells(i + 1, 1)) < 0.003 And Abs(r.Cells(i, 4) - r.Cells(i + 1, 2)) < 0.003 Then
            C.Cells(i + 1, 27).Value = n
            C.Cells(i + 2, 27).Value = n
            C.Cells(i + 3, 27).Value = n
        End If
        j = j + 1
        n = n + 1
    Wend
    j = 1
End If
i = i + 1
n = n + 1
Wend

```

End Sub

Sub Trip()

```

    Dim q As Integer
    Dim i As Long
    Dim j As Long
    Dim r As Range
    Dim n As Long
    Dim k As Long
    Dim T As Worksheet
    Dim C As Worksheet
    Set T = ActiveWorkbook.Sheets("Cost")
    Set C = ActiveWorkbook.Sheets("Trip")

    ' Dim Olat As String
    ' Dim Olong As String
    ' Dim DLat As String
    ' Dim DLong As String
    q = 0
    i = 1
    j = 1
    k = 2

    While k < 30000
        While C.Cells(k, 1) - C.Cells(k + 1, 1) = 0
            Set r = Range("B2:E29080")
            ' While i < 300
                If C.Cells(i, 27).Value < 1 Then
                    While j < 300
                        If Abs(r.Cells(i, 1) - r.Cells(i + j, 1)) < 0.003 And Abs(r.Cells(i, 2) - r.Cells(i + j, 2)) < 0.003 And
Abs(r.Cells(i, 3) - r.Cells(i + j, 3)) < 0.003 And Abs(r.Cells(i, 4) - r.Cells(i + j, 4)) < 0.003 Then
                            If C.Cells(i + j + 1, 27).Value < 1 Then
                                C.Cells(i + j + 1, 27).Value = n
                                q = q + 1
                            End If
                            If C.Cells(i + 1, 27).Value < 1 Then
                                C.Cells(i + 1, 27).Value = n
                                q = q + 1
                            End If
                        End If
                    End If
                End If
            End If
        End If
    End If

```

```

        If Abs(r.Cells(i, 1) - r.Cells(i + j, 3)) < 0.003 And Abs(r.Cells(i, 2) - r.Cells(i + j, 4)) < 0.003 And
Abs(r.Cells(i, 3) - r.Cells(i + j, 1)) < 0.003 And Abs(r.Cells(i, 4) - r.Cells(i + j, 2)) < 0.003 Then
            If C.Cells(i + 1, 27).Value < 1 Then
                C.Cells(i + 1, 27).Value = n
                q = q + 1
            End If
            If C.Cells(i + j + 1, 27).Value < 1 Then
                C.Cells(i + j + 1, 27).Value = n
                q = q + 1
            End If
        End If

        If Abs(r.Cells(i, 1) - r.Cells(i + 2, 3)) < 0.003 And Abs(r.Cells(i, 2) - r.Cells(i + 2, 4)) < 0.003 And
Abs(r.Cells(i, 3) - r.Cells(i + 1, 1)) < 0.003 And Abs(r.Cells(i, 4) - r.Cells(i + 1, 2)) < 0.003 Then
            If C.Cells(i + 1, 27).Value < 1 Then
                C.Cells(i + 1, 27).Value = n
                q = q + 1
            End If
            If C.Cells(i + 2, 27).Value < 1 Then
                C.Cells(i + 2, 27).Value = n
                q = q + 1
            End If
            If C.Cells(i + 3, 27).Value < 1 Then
                C.Cells(i + 3, 27).Value = n
                q = q + 1
            End If
            ' C.Cells(i + 2, 27).Value = n
            ' C.Cells(i + 3, 27).Value = n
        End If
        j = j + 1
        ' n = n + 1
    Wend
    j = 1
End If
C.Cells(i + 1, 28).Value = q
i = i + 1
k = k + 1
n = n + 1
q = 0
'Wend

'i = 1
'j = 1
Wend
k = k + 1
Wend
End Sub

```

Converting Lat/Long to distance:

```

Sub NHTS()
Dim i As Long
Dim l1 As String
Dim l2 As String
Set A = ActiveWorkbook.Sheets("A")
i = 80000
l1 = 0.68059

```

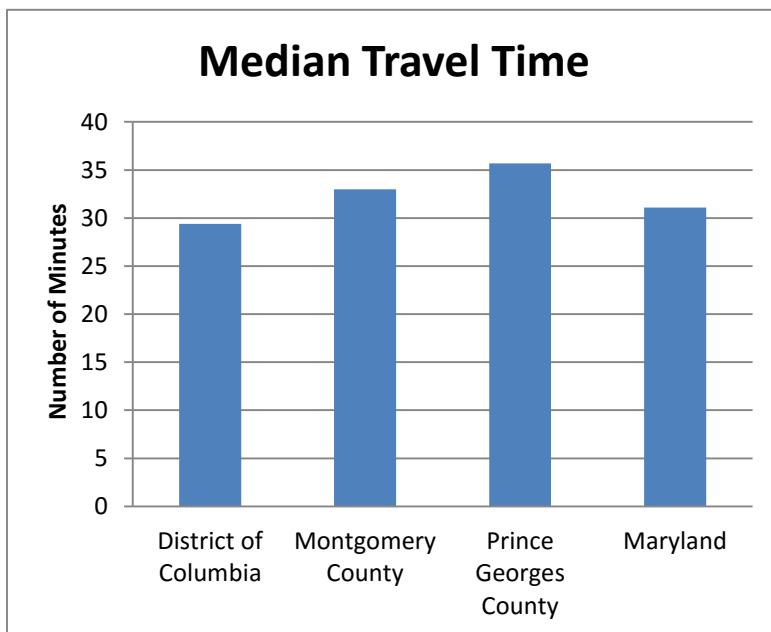
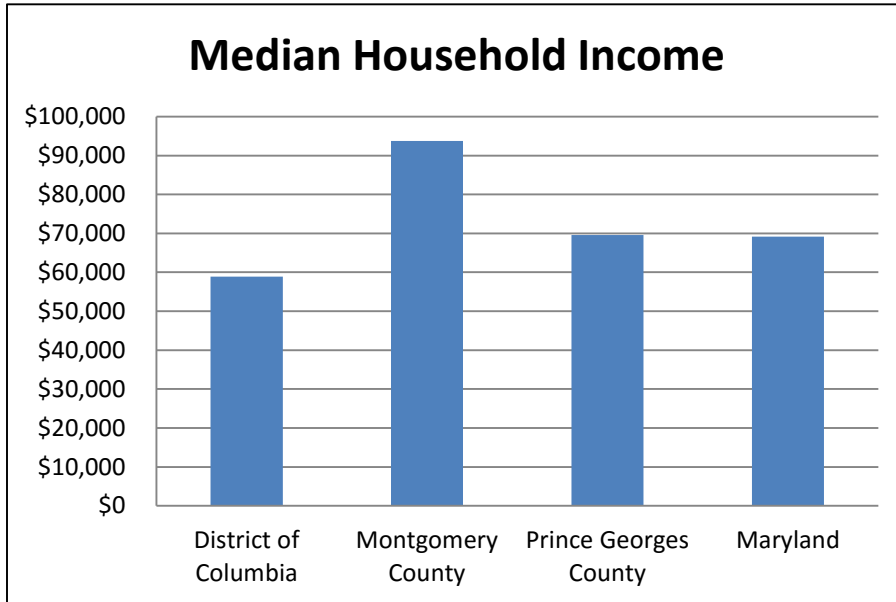
```

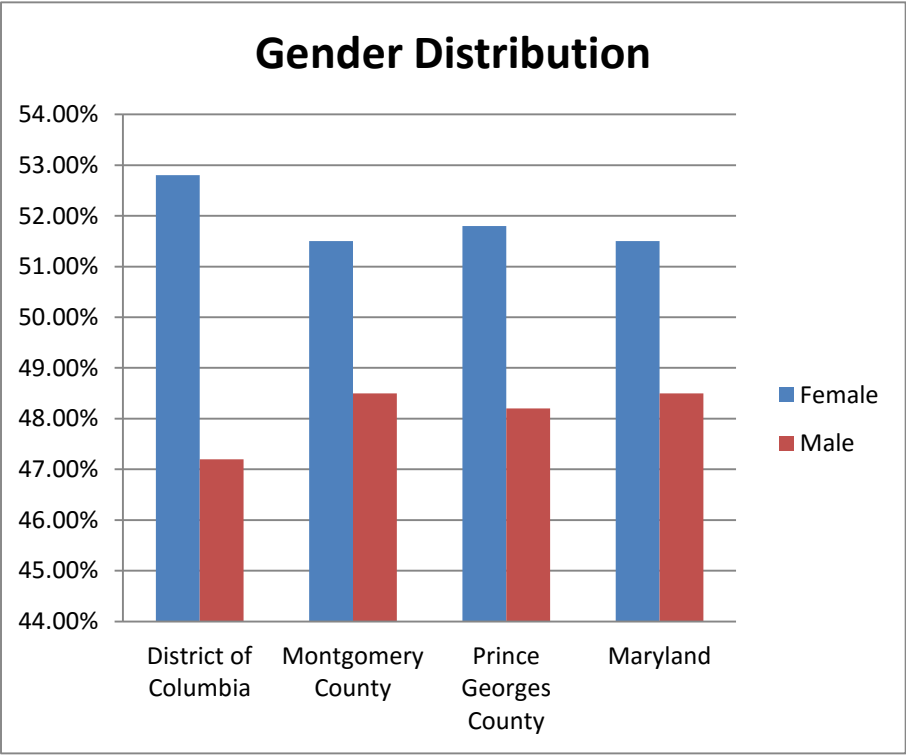
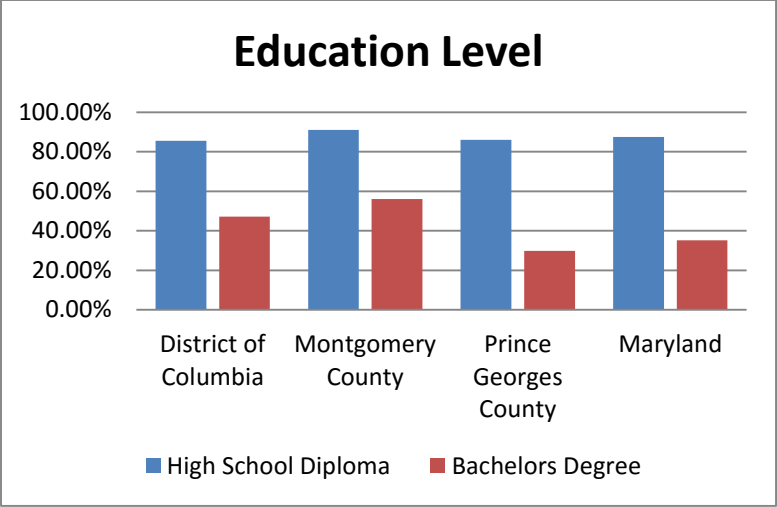
l2 = 1.342265
While i < 200000
If A.Cells(i - 1, 1).Value <> A.Cells(i, 1).Value Then
' l1 = A.Cells(i, 16).Value
' l2 = A.Cells(i, 17).Value
End If
A.Cells(i, 26).Value = Application.Acos(Sin(l1) * Sin(A.Cells(i, 16).Value) + Cos(l1) * Cos(A.Cells(i, 16).Value) *
Cos(A.Cells(i, 17).Value - l2)) * 3963.1676
i = i + 1
If A.Cells(i + 1, 1).Value <> A.Cells(i, 1).Value Then
l1 = A.Cells(i + 1, 16).Value
l2 = A.Cells(i + 1, 17).Value
End If
i = i + 1
Wend
End Sub

```

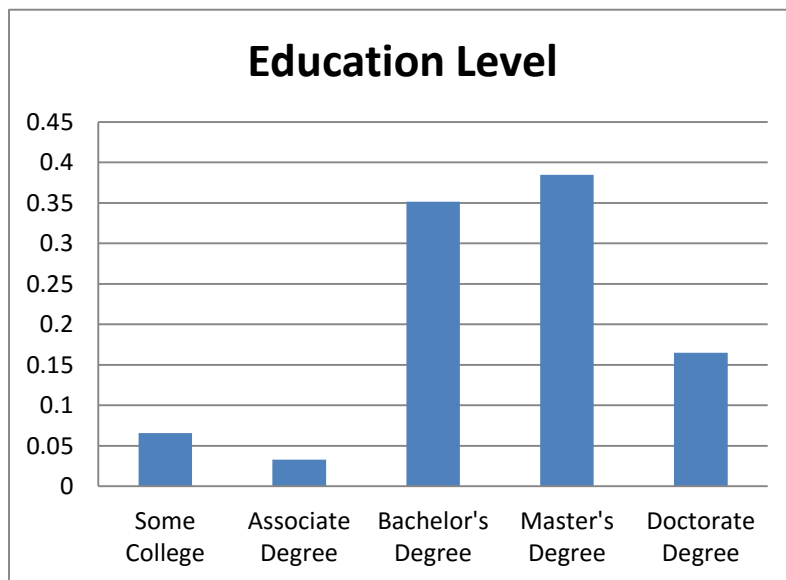
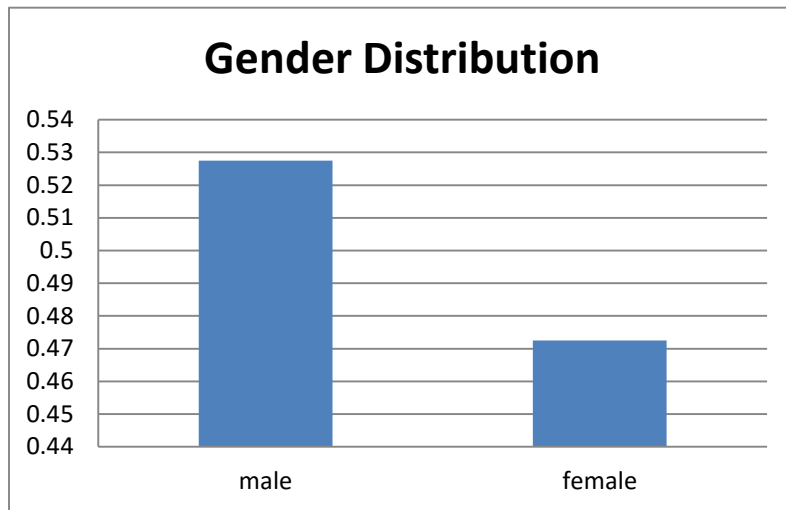

Appendix B, GPS Survey demographic information:

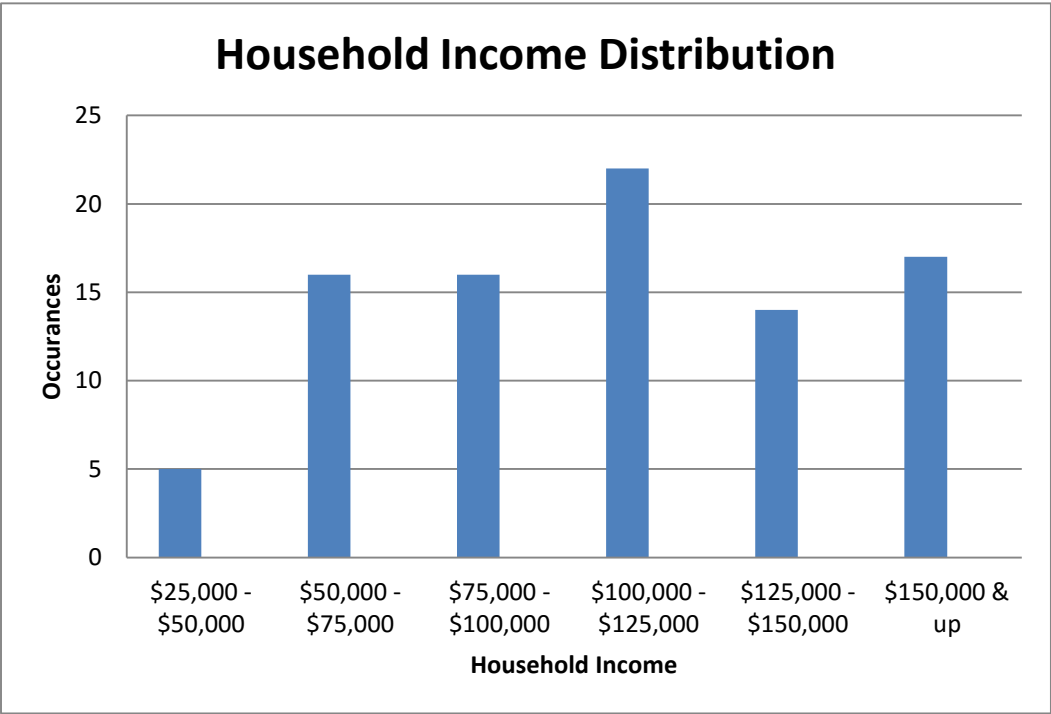
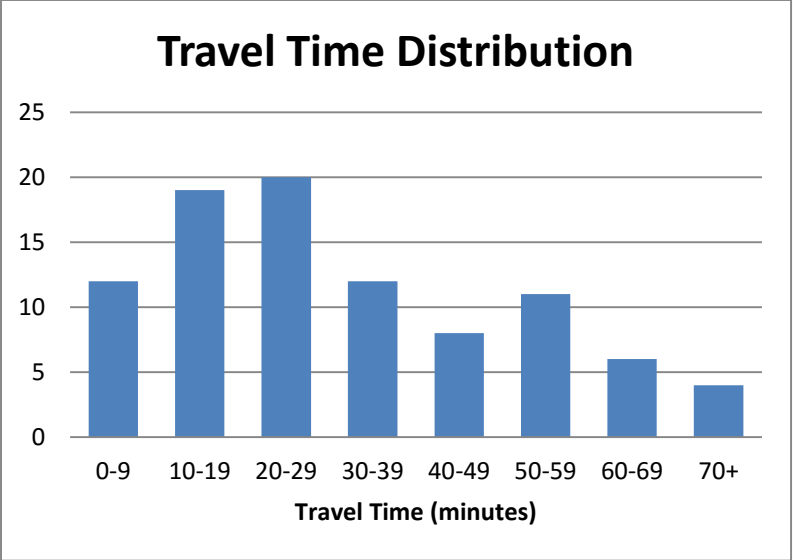
Surrounding Area





Pilot Study Participants



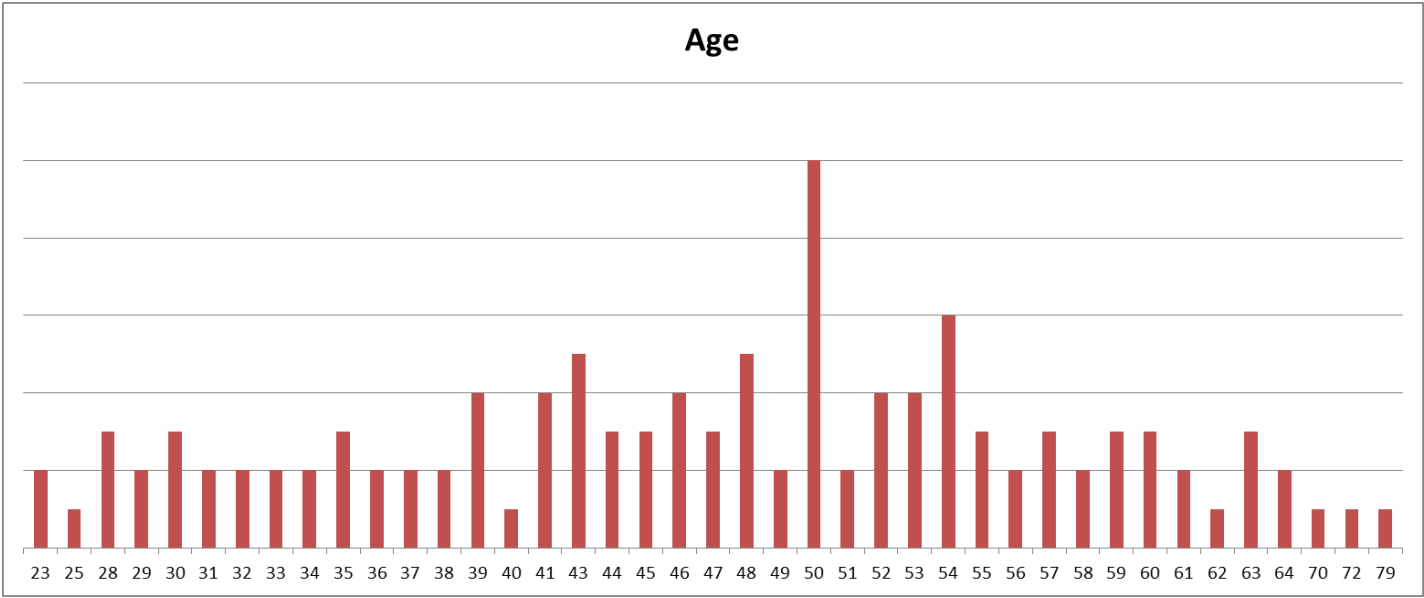


\$0-	\$24,999-	\$49,999-	\$74,999-	\$99,999-	\$124,999-	\$150,000-
\$24,999	\$49,999	\$74,999	\$99,999	\$124,999	\$149,999	

#	of	4	13	18	19	18	19	28
Participants								
Percentage		3%	11%	15%	16%	15%	16%	24%

Income distribution of participants selected.

Age:







Appendix C: GPS Forms and information mailed to survey participants:

Travel Diary

Travel Diary Example:

This form is a simple example of how to fill out the travel survey. If you are having trouble understanding the layout or how to fill out the diary, please contact Cory Krause at ckrause@umd.edu with your questions.

First I input my name and for which day I am filling out the survey for.



Home > GPS Participation

GPS Survey

Thank you for taking part in the University of Maryland's GPS travel survey. Please fill out this form as fully as possible.
If you have any questions or have trouble filling out the survey, please email me at ckrause@umd.edu or call me at 301-852-3392.
Thanks for your help
Cory Krause

First Name (*)
Last Name (*)

John
Smith

Enter date for which you are filling out your travel diary: (*)

10.08.2011

October 2011

Su	Mo	Tu	We	Th	Fr	Sa
25	26	27	28	29	30	1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5

NEXT

Copyright 2010 - 2013 The University of Maryland. All rights reserved.

All information given to the research team will be strictly used for research purposes.
Data will not be sold to an outside agency for any reason.

159

GPS Survey

Each page will be used for an individual trip. For example, if you drove from work to the grocery store, then to the day care then home, that is 3 separate trips (and would use 3 pages of this survey):

Where were you at 3am?

Home

Did you have at least one trip this day (either by car, walking, train, any mode of transportation) (*)

☐ Did not have any trips ☒ I had at least one trip

What time did you leave this location? (hh:mm am/pm)

08:00 am

What mode of transportation did you take on this trip?

Drive Alone

What was the purpose of the trip?

Work

What time did you arrive at the destination of this trip?
(hh:mm am/pm)

08:30 am

Nearest intersection:

Route 1 & Campus

Nearest landmark? (Building, park, school, church, etc...)

Kim Engineering

If you have no more trips to report, please click submit to finish the form.

Submit

If you have more trips for the day, please click on next.

PREV

NEXT

On the following page I fill out my first trip of the day: my drive to work. I leave at 8 am and arrive at 8:30 am. I fill out the nearest intersection as well as the nearest landmark. Since I have another trip for the day, I click 'NEXT'.

GPS Survey

What time did you leave this location? (hh:mm am/pm)

12:00 pm

What mode of transportation did you take on this trip?

Bike

What was the purpose of the trip?

Social/Recreational

What time did you arrive at the destination of this trip? (hh:mm am/pm)

12:15 pm

Where is that destination?

Nearest intersection:

Campus dr. & Knox rd

Nearest landmark? (Building, park, school, church, etc...)

Stamp Student Union

If you have no more trips to report, please click submit to finish the form.

Submit

If you have more trips for the day, please click on next.

PREV

NEXT

Here I fill out my trip to get lunch. I leave at 12 pm (noon) and arrive at my location at 12:15pm.

I rode my bike there and it is classified as social/recreational.

GPS Survey

What time did you leave this location? (hh:mm am/pm)

1:00 pm

What mode of transportation did you take on this trip?

Bike

What was the purpose of the trip?

Work

What time did you arrive at the destination of this trip? (hh:mm am/pm)

1:10 pm

Nearest intersection:

Route 1 & Campus dr.

Nearest landmark? (Building, park, school, church, etc...)

Kim Engineering

If you have no more trips to report, please click submit to finish the form.

Submit

If you have more trips for the day, please click on next.

PREV

NEXT

Shown above is my return trip from lunch.

GPS Survey

What time did you leave this location? (hh:mm am/pm)

4:45 pm

What mode of transportation did you take on this trip?

Drive Alone

What was the purpose of the trip?

Home

What time did you arrive at the destination of this trip? (hh:mm am/pm)

5:10 pm

Nearest intersection:

48th Place & Route 1

Nearest landmark? (Building, park, school, church, etc...)

CP Fire Station

If you have no more trips to report, please click submit to finish the form.

Submit

If you have more trips for the day, please click on next.

PREV

NEXT

And finally, my trip home for the day. If this is your last trip of the day, please click Submit to save the form.

Again, if you have any questions, feel free to call or email me. Thank you for your help.

Cory Krause

Transportation Systems Research Lab

University of Maryland

ckrause@umd.edu

301-852-3392

GPS Installation

GPS Installation Instructions

Thank you for taking our online travel survey and taking part in the GPS portion of our project.

This document will show you how to install the GPS device in your vehicle. If at any time you

have a problem with the directions, don't hesitate to email me at ckrause@umd.edu

Step 1: Plug the car charger into the vehicle that you take on a regular basis (most used vehicle) via the cigarette lighter receptacle.



Step 2: Insert the mini USB end of the car charger in the GPS port on the side of the device.



Step3: Move the switch on the side of the GPS all the way to the left so that it is on the “LOG” section.



Step 4: Drive normally and do not touch any of the buttons on the device during your travel survey period. When not driving, leave the device in the car, it will go into sleep mode after not moving for a few minutes. After the trial period is over, simply turn off the device, unplug it, and return using the shipping label we will have given you.

Step 5: Twice during the trial period, I will send you an email that asks you to validate your data. Please see the included document for an example of how to fill out the travel survey.

Thanks again for your help,

Cory Krause

Transportation Systems Research

University of Maryland

ckrause@umd.edu

301-852-3392

Return Shipping

Return Shipment Instructions:

- Fill out W-9 form
- Put GPS device and charger in the box. (make sure the GPS device is turned off)
- Sign Honorarium stating you participated in a Survey. Fill out all information including mailing address.
- Sign and date the consent form
- Take the return shipment label from inside the box and adhere it to the outside of the box. You only need to fill out section 1 of the form (your address information). All other sections can be left blank or are already filled out.
- Make sure that previous shipping labels are removed or covered by the new shipping label.
- Ship the package from any FedEx location. (All shipping costs have been prepaid).

Appendix D: 30 Day trip purpose training model

30 day training model tree:

Filename: 30day.model

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation:30daytrain

Attributes: 40

- S_Time
- Weekday
- Weekend
- Associate
- Bachelor
- HighSchool
- Master
- finalage
- FirstTrip
- MedTrip
- BusinessUnit
- Commercial
- GOV/PUB/Service
- Residential
- Restaurant
- Mixed_Use
- Institutional
- Industrial
- Leisure
- RecreationSite
- Shops
- Undeveloped
- Person ID
- gender
- Lat
- Long
- Home
- Work
- Home Lat
- Home Long
- Work Lat
- Work Long
- D_Home
- D_home2
- D_work
- income
- Purpose
- Soak
- Origin ID
- trip number

=== Classifier model ===

J48 pruned tree

```
D_Home <= 0.41
| Soak <= 14100
| | income <= 49999
| | | Home Lat <= 39.140853
| | | | Lat <= 38.985637: 3:Other (9.0)
| | | | Lat > 38.985637
| | | | | MedTrip <= 0: 4:School/D (13.0)
| | | | | MedTrip > 0
| | | | | | trip number <= 14: 3:Other (5.0/2.0)
| | | | | | trip number > 14: 4:School/D (3.0)
| | | | Home Lat > 39.140853: 2:Home (5.0)
| | | income > 49999
| | | Master <= 0
| | | | Home Lat <= 38.994672
| | | | | Weekday <= 0: 2:Home (2.0)
| | | | | Weekday > 0
| | | | | | finalage <= 52: 3:Other (2.0)
| | | | | | finalage > 52: 7:Work (5.0)
| | | | Home Lat > 38.994672: 3:Other (69.0/2.0)
| | | Master > 0
| | | | Work Lat <= 38.991534
| | | | | Person ID <= 255: 7:Work (14.0)
| | | | | Person ID > 255: 2:Home (3.0)
| | | | Work Lat > 38.991534
| | | | | Soak <= 6900
| | | | | | Mixed_Use <= 0
| | | | | | | Home Lat <= 39.041926: 3:Other (20.0)
| | | | | | | Home Lat > 39.041926
| | | | | | | Residential <= 0
| | | | | | | | Home Long <= -77.141932: 3:Other (5.0/1.0)
| | | | | | | | Home Long > -77.141932: 7:Work (5.0)
| | | | | | | Residential > 0: 7:Work (3.0)
| | | | | | Mixed_Use > 0
| | | | | | | MedTrip <= 0
| | | | | | | | D_Home <= 0.11: 7:Work (6.0/1.0)
| | | | | | | | D_Home > 0.11: 3:Other (2.0)
| | | | | | | MedTrip > 0: 7:Work (6.0/1.0)
| | | | | Soak > 6900: 7:Work (13.0)
| | Soak > 14100
| | | S_Time <= 0.668071: 7:Work (331.35/4.0)
| | | S_Time > 0.668071
| | | | Bachelor <= 0
| | | | | income <= 49999: 2:Home (5.0)
| | | | | income > 49999: 7:Work (44.71/1.0)
| | | | Bachelor > 0
| | | | | Work Long <= -76.847621: 3:Other (41.0/2.0)
```

```

| | | | Work Long > -76.847621: 7:Work (2.47)
D_Home > 0.41
| Soak <= 19200
| | Institutional <= 0
| | | Shops <= 0
| | | Residential <= 0
| | | Soak <= 9600
| | | Leisure <= 0
| | | | S_Time <= 0.309629
| | | | income <= 49999
| | | | | D_Home <= 2.85: 3:Other (2.0)
| | | | | D_Home > 2.85: 4:School/D (13.0)
| | | | | income > 49999
| | | | | Mixed_Use <= 0
| | | | | D_Home <= 8.61
| | | | | Commercial <= 0
| | | | | finalage <= 38
| | | | | | FirstTrip <= 0: 5:Shopping (2.0/1.0)
| | | | | | FirstTrip > 0: 3:Other (4.0)
| | | | | finalage > 38
| | | | | GOV/PUB/Service <= 0
| | | | | Restaurant <= 0
| | | | | Person ID <= 39: 3:Other (3.0)
| | | | | Person ID > 39
| | | | | Weekday <= 0: 2:Home (2.0)
| | | | | Weekday > 0
| | | | | Work Long <= -77.052049
| | | | | trip number <= 5: 3:Other (2.0)
| | | | | trip number > 5
| | | | | D_Home <= 4.21
| | | | | | S_Time <= 0.106771: 2:Home (2.0)
| | | | | | S_Time > 0.106771: 3:Other (2.0)
| | | | | D_Home > 4.21: 2:Home (10.0)
| | | | | Work Long > -77.052049: 3:Other (2.0)
| | | | | Restaurant > 0: 3:Other (2.0)
| | | | | GOV/PUB/Service > 0: 3:Other (2.0)
| | | | | Commercial > 0: 3:Other (2.0/1.0)
| | | | | D_Home > 8.61
| | | | | trip number <= 22
| | | | | Master <= 0: 5:Shopping (26.0/4.0)
| | | | | Master > 0
| | | | | income <= 124999: 5:Shopping (8.0/2.0)
| | | | | income > 124999: 3:Other (3.0)
| | | | | trip number > 22
| | | | | S_Time <= 0.010663: 4:School/D (5.0)
| | | | | S_Time > 0.010663
| | | | | FirstTrip <= 0: 2:Home (2.0)
| | | | | FirstTrip > 0: 5:Shopping (4.0/1.0)
| | | | | Mixed_Use > 0
| | | | | S_Time <= 0.024313: 6:Social/R (2.0)
| | | | | S_Time > 0.024313: 3:Other (2.0)

```

					S_Time > 0.309629
				D_home2 <= 0.34	
				D_home2 <= 0	
				Associate <= 0	
				S_Time <= 0.519813: 3:Other (37.0/3.0)	
				S_Time > 0.519813	
				income <= 49999	
				GOV/PUB/Service <= 0	
				Master <= 0	
				Weekday <= 0: 3:Other (9.0)	
				Weekday > 0	
				FirstTrip <= 0	
				Person ID <= 148	
				D_Home <= 3.97: 3:Other (3.0)	
				D_Home > 3.97: 5:Shopping (5.0)	
				Person ID > 148: 3:Other (3.0/1.0)	
				FirstTrip > 0: 3:Other (8.0/3.0)	
				Master > 0: 5:Shopping (3.0/1.0)	
				GOV/PUB/Service > 0: 5:Shopping (5.0/1.0)	
				income > 49999	
				Person ID <= 258	
				trip number <= 6	
				Person ID <= 33: 3:Other (9.0)	
				Person ID > 33	
				finalage <= 48: 5:Shopping (16.0/5.0)	
				finalage > 48	
				finalage <= 55: 3:Other (7.0)	
				finalage > 55: 5:Shopping (4.0)	
				trip number > 6	
				finalage <= 60	
				Industrial <= 0	
				GOV/PUB/Service <= 0	
				Home Lat <= 39.039408: 3:Other (27.0/4.0)	
				Home Lat > 39.039408	
				Restaurant <= 0	
				FirstTrip <= 0	
				Work Lat <= 38.942441: 5:Shopping (3.0)	
				Work Lat > 38.942441	
				S_Time <= 0.933429: 3:Other (27.0/7.0)	
				S_Time > 0.933429: 5:Shopping (5.0)	
				FirstTrip > 0	
				S_Time <= 0.590033: 3:Other (5.0)	
				S_Time > 0.590033: 5:Shopping (14.0/3.0)	
				Restaurant > 0	
				Soak <= 1400: 5:Shopping (4.0)	
				Soak > 1400: 3:Other (9.0/1.0)	
				GOV/PUB/Service > 0	
				income <= 124999: 3:Other (5.0)	
				income > 124999	
				FirstTrip <= 0: 5:Shopping (4.0)	
				FirstTrip > 0	

```

D_work <= 0.11: 5:Shopping (4.0/1.0)
D_work > 0.11: 3:Other (2.0)
Industrial > 0: 5:Shopping (3.0/1.0)
finalage > 60
D_Home <= 33.83: 5:Shopping (14.0/1.0)
D_Home > 33.83: 3:Other (2.0)
Person ID > 258: 3:Other (12.0)
Associate > 0
Origin ID <= 53758
income <= 99999: 5:Shopping (8.0/2.0)
income > 99999: 2:Home (2.0/1.0)
Origin ID > 53758: 3:Other (7.0/1.0)
D_home2 > 0
Undeveloped <= 0
Restaurant <= 0
Associate <= 0
Soak <= 697
S_Time <= 0.967725
Commercial <= 0
Person ID <= 217
Industrial <= 0
Lat <= 39.03387
GOV/PUB/Service <= 0: 3:Other (32.0/4.0)
GOV/PUB/Service > 0
D_home2 <= 0.03: 3:Other (4.0/1.0)
D_home2 > 0.03: 5:Shopping (3.0)
Lat > 39.03387
RecreationSite <= 0
Work Long <= -76.946833
Long <= -77.200864: 3:Other (7.0)
Long > -77.200864
HighSchool <= 0
Home Lat <= 39.121198: 2:Home (14.0/2.0)
Home Lat > 39.121198: 3:Other (5.0/1.0)
HighSchool > 0: 3:Other (3.0/1.0)
Work Long > -76.946833
finalage <= 55: 3:Other (6.0)
finalage > 55: 5:Shopping (2.0)
RecreationSite > 0: 3:Other (2.0/1.0)
Industrial > 0: 2:Home (2.0/1.0)
Person ID > 217: 3:Other (36.0/8.0)
Commercial > 0: 5:Shopping (4.0/1.0)
S_Time > 0.967725: 4:School/D (3.0/1.0)
Soak > 697
RecreationSite <= 0
D_Home <= 1.32
Origin ID <= 12181
Soak <= 2300: 3:Other (2.0/1.0)
Soak > 2300: 5:Shopping (4.0)
Origin ID > 12181: 3:Other (15.0/1.0)
D_Home > 1.32

```

[illegible]

[illegible]

		D_Home <= 25.37: 3:Other (4.0)
		D_Home > 25.37: 5:Shopping (2.0)
		Person ID > 258: 5:Shopping (3.0/1.0)
		Industrial > 0
		HighSchool <= 0: 5:Shopping (5.0)
		HighSchool > 0: 3:Other (2.0/1.0)
		D_work > 0.33
		Master <= 0: 5:Shopping (7.0/1.0)
		Master > 0
		Person ID <= 65: 5:Shopping (2.0)
		Person ID > 65: 3:Other (4.0/1.0)
		GOV/PUB/Service > 0
		FirstTrip <= 0
		Master <= 0
		MedTrip <= 0: 2:Home (6.0/1.0)
		MedTrip > 0: 3:Other (3.0/1.0)
		Master > 0
		Origin ID <= 59582: 3:Other (7.0)
		Origin ID > 59582: 2:Home (2.0/1.0)
		FirstTrip > 0: 5:Shopping (7.0/3.0)
		RecreationSite > 0
		D_Home <= 1.12: 2:Home (2.0/1.0)
		D_Home > 1.12
		HighSchool <= 0
		Home Long <= -76.896507
		Work Lat <= 39.001428
		D_home2 <= 0.06: 5:Shopping (4.0)
		D_home2 > 0.06: 3:Other (2.0)
		Work Lat > 39.001428: 3:Other (11.0)
		Home Long > -76.896507: 5:Shopping (4.0/1.0)
		HighSchool > 0: 5:Shopping (6.0)
		Associate > 0
		Work Long <= -76.891298
		Soak <= 4519
		finalage <= 58
		Soak <= 700: 2:Home (2.0/1.0)
		Soak > 700
		S_Time <= 0.858158: 3:Other (7.0/1.0)
		S_Time > 0.858158: 5:Shopping (4.0/1.0)
		finalage > 58: 5:Shopping (3.0)
		Soak > 4519: 2:Home (2.0)
		Work Long > -76.891298
		Work Lat <= 38.987017
		S_Time <= 0.608063: 2:Home (2.0)
		S_Time > 0.608063: 3:Other (3.0)
		Work Lat > 38.987017: 2:Home (14.0)
		Restaurant > 0
		income <= 74999: 5:Shopping (14.0/1.0)
		income > 74999
		income <= 124999
		Commercial <= 0

[illegible]


```

| | | | | | | | | | MedTrip > 0
| | | | | | | | | |   Origin ID <= 78166
| | | | | | | | | |     S_Time <= 0.899775: 3:Other (14.0)
| | | | | | | | | |     S_Time > 0.899775: 6:Social/R (3.0/1.0)
| | | | | | | | | |     Origin ID > 78166: 6:Social/R (2.0)
| | | | | | | | | |     D_work > 0.24: 6:Social/R (4.0/1.0)
| | | | | | | | | |     FirstTrip > 0: 3:Other (7.0/1.0)
| | | | | | | | | |     Origin ID > 88294: 5:Shopping (2.0)
| | | | | Soak > 9600
| | | | |   FirstTrip <= 0
| | | | |     Commercial <= 0
| | | | |       MedTrip <= 0
| | | | |         Leisure <= 0
| | | | |           D_home2 <= 0: 3:Other (2.0/1.0)
| | | | |           D_home2 > 0
| | | | |             GOV/PUB/Service <= 0: 2:Home (18.0/1.0)
| | | | |             GOV/PUB/Service > 0: 3:Other (3.0/1.0)
| | | | |             Leisure > 0: 3:Other (2.0)
| | | | |   MedTrip > 0
| | | | |     Leisure <= 0
| | | | |       Weekday <= 0
| | | | |         income <= 74999: 6:Social/R (6.0/2.0)
| | | | |         income > 74999
| | | | |           GOV/PUB/Service <= 0
| | | | |             D_work <= 0.42: 3:Other (15.0/2.0)
| | | | |             D_work > 0.42: 6:Social/R (3.0)
| | | | |             GOV/PUB/Service > 0: 6:Social/R (2.0)
| | | | |       Weekday > 0
| | | | |         Work Long <= -77.248586: 2:Home (3.0)
| | | | |         Work Long > -77.248586
| | | | |           Lat <= 38.897921
| | | | |             Lat <= 38.650237: 3:Other (3.0)
| | | | |             Lat > 38.650237: 5:Shopping (2.0)
| | | | |             Lat > 38.897921
| | | | |               Home Long <= -76.935978
| | | | |                 D_Home <= 1.02: 2:Home (2.0/1.0)
| | | | |                 D_Home > 1.02: 3:Other (41.0/3.0)
| | | | |                 Home Long > -76.935978
| | | | |                   D_Home <= 11.5
| | | | |                     Home Long <= -76.896507: 2:Home (6.0)
| | | | |                     Home Long > -76.896507: 3:Other (2.0)
| | | | |                     D_Home > 11.5: 3:Other (9.0)
| | | | |           Leisure > 0
| | | | |             Home Long <= -76.935541: 6:Social/R (10.0/1.0)
| | | | |             Home Long > -76.935541: 2:Home (3.0)
| | | | |       Commercial > 0: 5:Shopping (2.0)
| | | | |   FirstTrip > 0
| | | | |     Commercial <= 0
| | | | |       finalage <= 52
| | | | |         Bachelor <= 0
| | | | |         Leisure <= 0

```

					S_Time <= 0.566375: 3:Other (5.0)
					S_Time > 0.566375
					income <= 149999: 6:Social/R (9.0)
					income > 149999
					trip number <= 11: 6:Social/R (5.0/1.0)
					trip number > 11: 3:Other (4.0)
					Leisure > 0: 6:Social/R (6.0)
					Bachelor > 0
					Work Long <= -77.10198: 2:Home (3.0/1.0)
					Work Long > -77.10198: 6:Social/R (6.0)
					finalage > 52
					D_home2 <= 0
					Person ID <= 253: 6:Social/R (11.0/2.0)
					Person ID > 253: 5:Shopping (3.0/1.0)
					D_home2 > 0
					Home Lat <= 39.084969: 2:Home (7.0)
					Home Lat > 39.084969: 6:Social/R (3.0)
					Commercial > 0: 3:Other (4.0/1.0)
					Residential > 0
					Work Long <= -76.876958
					finalage <= 55
					Mixed_Use <= 0
					Home <= 85177
					Bachelor <= 0
					Associate <= 0
					D_Home <= 6.96
					Work Long <= -77.218229: 6:Social/R (3.0)
					Work Long > -77.218229
					Home Lat <= 39.140853
					D_Home <= 0.73: 5:Shopping (3.0)
					D_Home > 0.73
					D_Home <= 5.12: 2:Home (39.0/2.0)
					D_Home > 5.12
					income <= 74999: 2:Home (4.0/1.0)
					income > 74999: 3:Other (3.0)
					Home Lat > 39.140853: 3:Other (2.0/1.0)
					D_Home > 6.96
					S_Time <= 0.722204
					Soak <= 11400
					Home Long <= -77.074487
					finalage <= 41: 5:Shopping (3.0)
					finalage > 41
					Long <= -77.15427
					D_home2 <= 0.06: 6:Social/R (5.0/1.0)
					D_home2 > 0.06: 4:School/D (2.0)
					Long > -77.15427: 5:Shopping (2.0)
					Home Long > -77.074487
					Long <= -76.989072: 1:Driving (4.0/1.0)
					Long > -76.989072: 4:School/D (4.0)
					Soak > 11400: 3:Other (2.0)
					S Time > 0.722204

[illegible]

```

| | | | | | | | | | Master <= 0
| | | | | | | | | | Home Lat <= 39.074142
| | | | | | | | | | trip number <= 23: 4:School/D (6.0/1.0)
| | | | | | | | | | trip number > 23: 3:Other (3.0/1.0)
| | | | | | | | | | Home Lat > 39.074142: 3:Other (4.0)
| | | | | | | | | | Master > 0
| | | | | | | | | | income <= 124999: 5:Shopping (4.0/1.0)
| | | | | | | | | | income > 124999: 3:Other (2.0)
| | | | | | | | | | FirstTrip > 0
| | | | | | | | | | GOV/PUB/Service <= 0: 6:Social/R (4.0)
| | | | | | | | | | GOV/PUB/Service > 0: 5:Shopping (5.0/1.0)
| | | | | | | | | | finalage > 55
| | | | | | | | | | Weekday <= 0
| | | | | | | | | | Soak <= 2100: 1:Driving (7.0)
| | | | | | | | | | Soak > 2100: 2:Home (4.0/1.0)
| | | | | | | | | | Weekday > 0
| | | | | | | | | | Bachelor <= 0
| | | | | | | | | | Home Lat <= 39.117919
| | | | | | | | | | Mixed_Use <= 0
| | | | | | | | | | Long <= -77.117984: 3:Other (6.0/1.0)
| | | | | | | | | | Long > -77.117984: 2:Home (8.0/2.0)
| | | | | | | | | | Mixed_Use > 0
| | | | | | | | | | Soak <= 2900: 2:Home (4.0)
| | | | | | | | | | Soak > 2900: 1:Driving (2.0)
| | | | | | | | | | Home Lat > 39.117919: 1:Driving (3.0)
| | | | | | | | | | Bachelor > 0
| | | | | | | | | | Origin ID <= 86192: 3:Other (14.0/2.0)
| | | | | | | | | | Origin ID > 86192: 2:Home (2.0)
| | | | | | | | | | Work Long > -76.876958
| | | | | | | | | | finalage <= 41: 4:School/D (4.0/1.0)
| | | | | | | | | | finalage > 41
| | | | | | | | | | Associate <= 0
| | | | | | | | | | Master <= 0: 1:Driving (2.0/1.0)
| | | | | | | | | | Master > 0
| | | | | | | | | | S_Time <= 0.750571: 1:Driving (7.0)
| | | | | | | | | | S_Time > 0.750571: 2:Home (2.0)
| | | | | | | | | | Associate > 0
| | | | | | | | | | S_Time <= 0.691371: 1:Driving (2.0)
| | | | | | | | | | S_Time > 0.691371: 5:Shopping (2.0)
| | | | | | | | | | Shops > 0
| | | | | | | | | | Person ID <= 12: 3:Other (4.0/1.0)
| | | | | | | | | | Person ID > 12
| | | | | | | | | | D_work <= 0.27
| | | | | | | | | | Long <= -76.873999: 5:Shopping (23.0/1.0)
| | | | | | | | | | Long > -76.873999: 3:Other (4.0/1.0)
| | | | | | | | | | D_work > 0.27: 3:Other (2.0)
| | | | | | | | | | Institutional > 0
| | | | | | | | | | finalage <= 44
| | | | | | | | | | S_Time <= 0.720154
| | | | | | | | | | D_Home <= 2.08: 1:Driving (3.0/1.0)
| | | | | | | | | | D_Home > 2.08: 3:Other (14.0/1.0)

```

```

| | | | S_Time > 0.720154
| | | | | Weekday <= 0
| | | | | | Soak <= 2900
| | | | | | | Person ID <= 114: 6:Social/R (2.0)
| | | | | | | Person ID > 114: 3:Other (2.0)
| | | | | | Soak > 2900: 2:Home (2.0)
| | | | | Weekday > 0
| | | | | | income <= 49999: 1:Driving (2.0)
| | | | | | income > 49999
| | | | | | | Mixed_Use <= 0
| | | | | | | | Bachelor <= 0: 6:Social/R (6.0/1.0)
| | | | | | | | Bachelor > 0: 2:Home (3.0)
| | | | | | | Mixed_Use > 0: 6:Social/R (10.0/1.0)
| | | | finalage > 44
| | | | | Soak <= 9800
| | | | | | Bachelor <= 0: 4:School/D (39.0/2.0)
| | | | | | Bachelor > 0
| | | | | | | finalage <= 54
| | | | | | | | Residential <= 0: 4:School/D (10.0/1.0)
| | | | | | | | Residential > 0: 2:Home (3.0/1.0)
| | | | | | | finalage > 54: 6:Social/R (4.0)
| | | | | Soak > 9800: 6:Social/R (10.0/1.0)
| | Soak > 19200
| | | D_home2 <= 0
| | | | finalage <= 32
| | | | | Residential <= 0
| | | | | | Leisure <= 0
| | | | | | | S_Time <= 0.295838: 2:Home (3.0)
| | | | | | | S_Time > 0.295838
| | | | | | | | trip number <= 6: 3:Other (6.0/1.0)
| | | | | | | | trip number > 6
| | | | | | | | | Home Long <= -77.144835: 6:Social/R (5.0)
| | | | | | | | | Home Long > -77.144835
| | | | | | | | | Work Long <= -76.979234: 3:Other (5.0/1.0)
| | | | | | | | | Work Long > -76.979234
| | | | | | | | | MedTrip <= 0: 6:Social/R (2.0)
| | | | | | | | | MedTrip > 0: 3:Other (2.0)
| | | | | | | | Leisure > 0: 3:Other (2.0)
| | | | | | Residential > 0: 2:Home (4.0)
| | | | finalage > 32
| | | | | Leisure <= 0
| | | | | | Residential <= 0
| | | | | | | D_Home <= 31.84
| | | | | | | | Soak <= 24500: 3:Other (4.0)
| | | | | | | | Soak > 24500: 6:Social/R (64.68/10.68)
| | | | | | | D_Home > 31.84
| | | | | | | | Home Lat <= 39.115942: 3:Other (11.0/1.0)
| | | | | | | | Home Lat > 39.115942: 6:Social/R (3.14/0.14)
| | | | | | Residential > 0: 6:Social/R (18.0/4.0)
| | | | | Leisure > 0: 6:Social/R (15.0)
| | | D_home2 > 0

```

```

| | | D_home2 <= 1.98
| | | Soak <= 194300
| | | | D_Home <= 51.66
| | | | Mixed_Use <= 0
| | | | MedTrip <= 0
| | | | Industrial <= 0
| | | | D_home2 <= 0.01
| | | | | Person ID <= 268
| | | | | FirstTrip <= 0: 2:Home (35.0/1.0)
| | | | | FirstTrip > 0
| | | | | Soak <= 89206
| | | | | | finalage <= 45
| | | | | | Soak <= 33130: 3:Other (4.0)
| | | | | | Soak > 33130: 6:Social/R (3.8/0.8)
| | | | | | finalage > 45
| | | | | | Soak <= 79700: 2:Home (6.0/1.0)
| | | | | | Soak > 79700: 3:Other (2.0)
| | | | | | Soak > 89206: 2:Home (15.0)
| | | | | | Person ID > 268: 6:Social/R (3.0)
| | | | | D_home2 > 0.01
| | | | | Soak <= 179145: 2:Home (496.0/15.0)
| | | | | Soak > 179145
| | | | | | D_home2 <= 0.26
| | | | | | income <= 149999: 2:Home (20.0/1.0)
| | | | | | income > 149999: 6:Social/R (3.0/1.0)
| | | | | | D_home2 > 0.26
| | | | | | Long <= -77.15153: 6:Social/R (2.0)
| | | | | | Long > -77.15153: 3:Other (2.0)
| | | | | Industrial > 0
| | | | | | HighSchool <= 0: 6:Social/R (2.0)
| | | | | | HighSchool > 0
| | | | | | finalage <= 32: 3:Other (2.0)
| | | | | | finalage > 32: 2:Home (9.0)
| | | | | MedTrip > 0
| | | | | | HighSchool <= 0
| | | | | | Leisure <= 0
| | | | | | Soak <= 90300
| | | | | | Person ID <= 268
| | | | | | Residential <= 0
| | | | | | Institutional <= 0
| | | | | | D_home2 <= 0.07
| | | | | | | S_Time <= 0.801496: 2:Home (38.8/4.8)
| | | | | | | S_Time > 0.801496
| | | | | | | Soak <= 20700: 2:Home (2.0)
| | | | | | | Soak > 20700: 6:Social/R (2.0)
| | | | | | D_home2 > 0.07
| | | | | | | S_Time <= 0.608063
| | | | | | | D_home2 <= 0.14: 6:Social/R (4.0/1.0)
| | | | | | | D_home2 > 0.14: 2:Home (6.0)
| | | | | | | S_Time > 0.608063: 3:Other (9.0/3.0)
| | | | | | Institutional > 0: 2:Home (2.0/1.0)

```

```

| | | | | | | | | | Residential > 0: 2:Home (27.0/1.0)
| | | | | | | | | | Person ID > 268
| | | | | | | | | | Origin ID <= 74904: 6:Social/R (2.0)
| | | | | | | | | | Origin ID > 74904: 3:Other (6.0)
| | | | | | | | | | Soak > 90300: 2:Home (50.0)
| | | | | | | | | | Leisure > 0
| | | | | | | | | | Weekday <= 0: 2:Home (5.0/1.0)
| | | | | | | | | | Weekday > 0: 6:Social/R (3.0)
| | | | | | | | | | HighSchool > 0
| | | | | | | | | | Person ID <= 98: 6:Social/R (7.0/2.0)
| | | | | | | | | | Person ID > 98
| | | | | | | | | | D_work <= 0.15
| | | | | | | | | | Residential <= 0: 6:Social/R (5.0/1.0)
| | | | | | | | | | Residential > 0: 2:Home (2.0)
| | | | | | | | | | D_work > 0.15: 2:Home (7.0)
| | | | | | | | | | Mixed_Use > 0
| | | | | | | | | | finalage <= 53
| | | | | | | | | | Weekday <= 0: 6:Social/R (4.0)
| | | | | | | | | | Weekday > 0: 3:Other (4.0/1.0)
| | | | | | | | | | finalage > 53: 2:Home (12.0/1.0)
| | | | | | | | | | D_Home > 51.66
| | | | | | | | | | income <= 74999: 2:Home (13.0/2.0)
| | | | | | | | | | income > 74999
| | | | | | | | | | Bachelor <= 0
| | | | | | | | | | Long <= -76.904439: 6:Social/R (4.05/0.05)
| | | | | | | | | | Long > -76.904439: 3:Other (3.0)
| | | | | | | | | | Bachelor > 0: 3:Other (7.0/1.0)
| | | | | | | | | | Soak > 194300
| | | | | | | | | | Restaurant <= 0
| | | | | | | | | | S_Time <= 0.967725
| | | | | | | | | | Soak <= 233400
| | | | | | | | | | Institutional <= 0
| | | | | | | | | | S_Time <= 0.480246: 2:Home (13.0)
| | | | | | | | | | S_Time > 0.480246
| | | | | | | | | | D_home2 <= 0.31
| | | | | | | | | | Mixed_Use <= 0
| | | | | | | | | | Soak <= 202500
| | | | | | | | | | HighSchool <= 0: 3:Other (2.0)
| | | | | | | | | | HighSchool > 0: 6:Social/R (3.0)
| | | | | | | | | | Soak > 202500
| | | | | | | | | | FirstTrip <= 0
| | | | | | | | | | D_work <= 0.06: 2:Home (22.0/1.0)
| | | | | | | | | | D_work > 0.06
| | | | | | | | | | D_home2 <= 0.03: 2:Home (5.0/1.0)
| | | | | | | | | | D_home2 > 0.03
| | | | | | | | | | Origin ID <= 6954: 2:Home (4.0/1.0)
| | | | | | | | | | Origin ID > 6954: 6:Social/R (8.0/1.0)
| | | | | | | | | | FirstTrip > 0
| | | | | | | | | | S_Time <= 0.92635: 2:Home (3.0/1.0)
| | | | | | | | | | S_Time > 0.92635: 6:Social/R (2.0)
| | | | | | | | | | Mixed_Use > 0

```

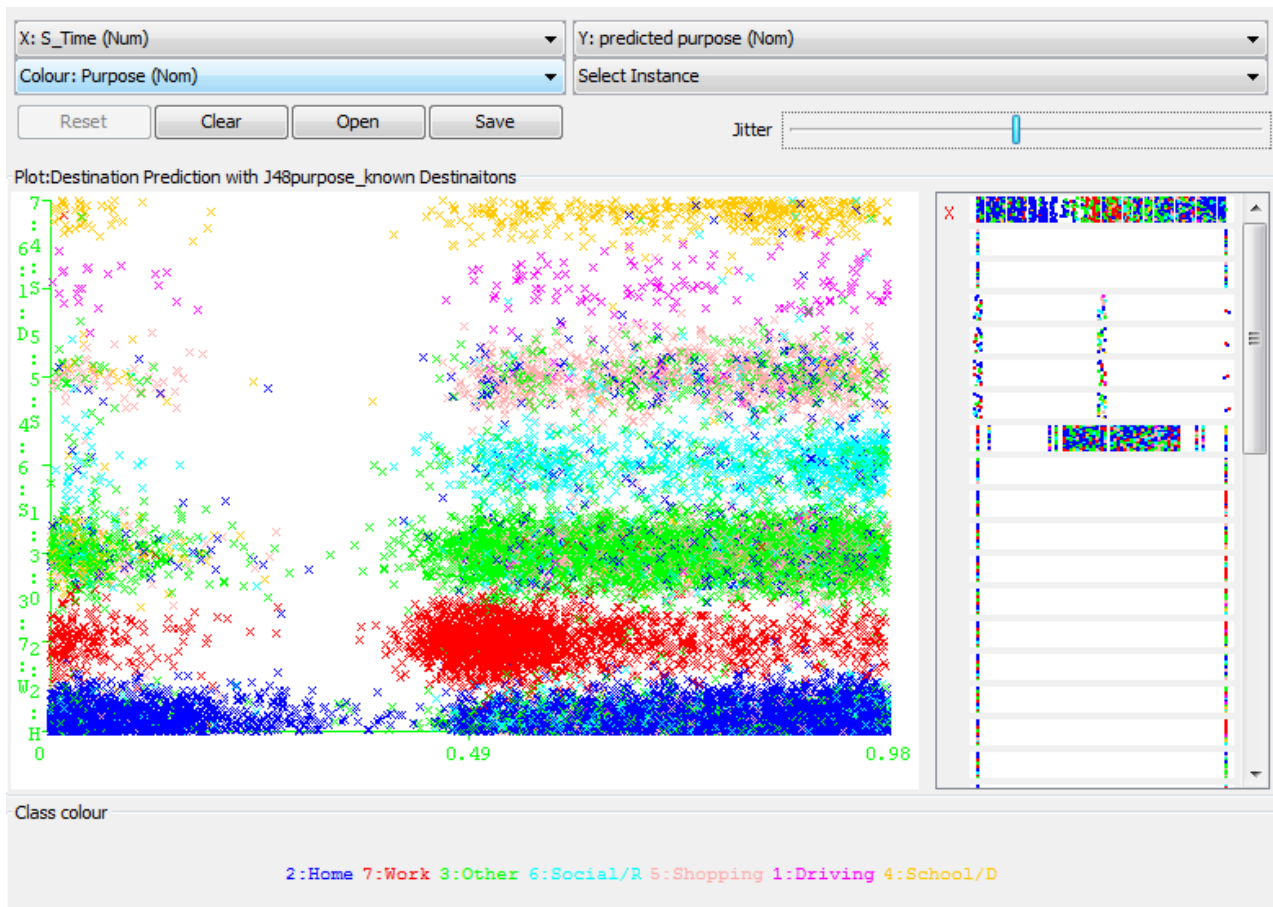
```

| | | | | | | | | | | Home Lat <= 39.072464: 2:Home (5.0/1.0)
| | | | | | | | | | | Home Lat > 39.072464: 3:Other (3.0)
| | | | | | | | | | | D_home2 > 0.31
| | | | | | | | | | | Bachelor <= 0: 3:Other (5.0/1.0)
| | | | | | | | | | | Bachelor > 0: 6:Social/R (4.0)
| | | | | | | | | | | Institutional > 0: 2:Home (4.0/1.0)
| | | | | | | | | | | Soak > 233400: 3:Other (6.0/1.0)
| | | | | | | | | | | S_Time > 0.967725
| | | | | | | | | | | Residential <= 0: 6:Social/R (15.0/3.0)
| | | | | | | | | | | Residential > 0
| | | | | | | | | | | Bachelor <= 0: 2:Home (3.0)
| | | | | | | | | | | Bachelor > 0
| | | | | | | | | | | Person ID <= 47: 2:Home (3.0/1.0)
| | | | | | | | | | | Person ID > 47: 6:Social/R (6.0)
| | | | | | | | | | | Restaurant > 0: 6:Social/R (2.0)
| | | | | | | | | | | D_home2 > 1.98
| | | | | | | | | | | S_Time <= 0.858158
| | | | | | | | | | | Home Long <= -77.057113: 6:Social/R (3.0/1.0)
| | | | | | | | | | | Home Long > -77.057113: 3:Other (6.0)
| | | | | | | | | | | S_Time > 0.858158: 6:Social/R (5.0)

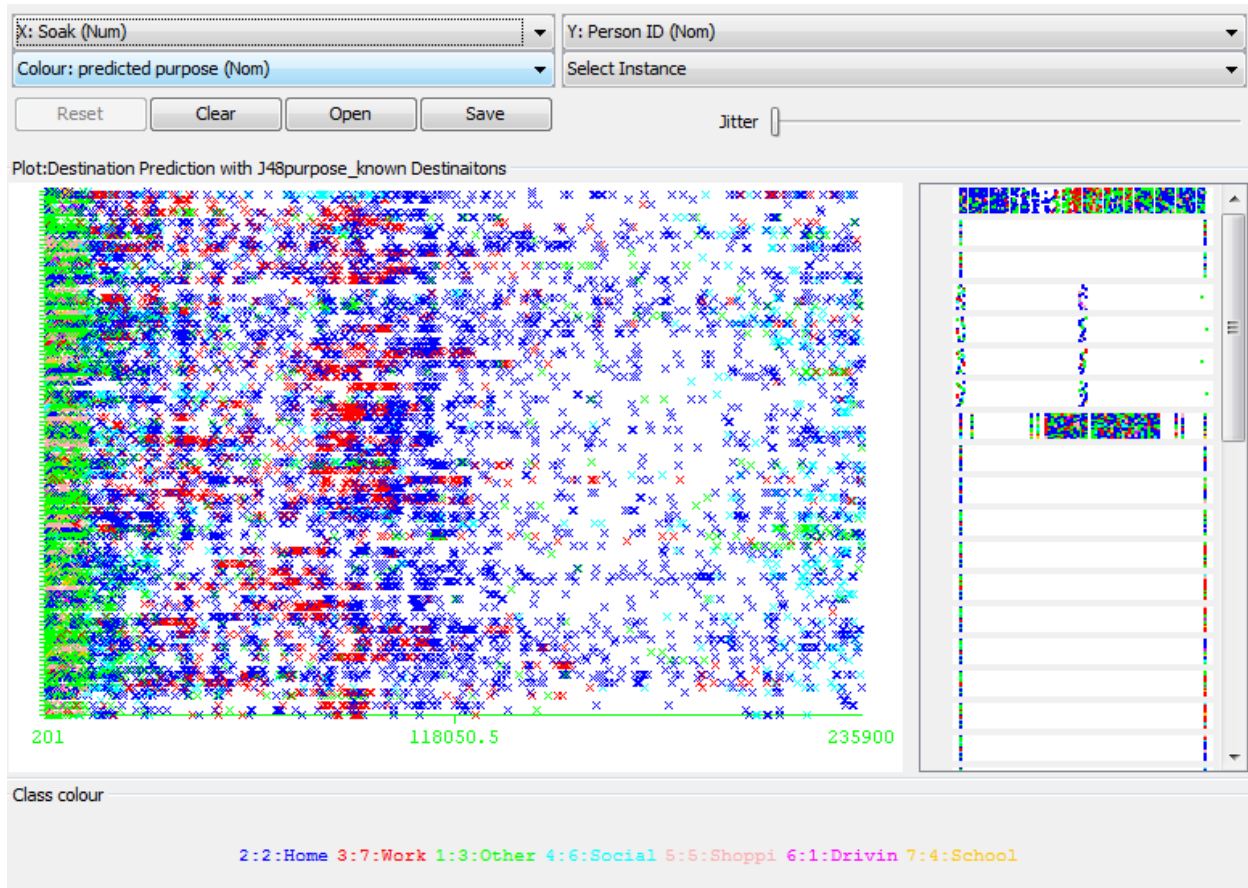
```

Number of Leaves : 387

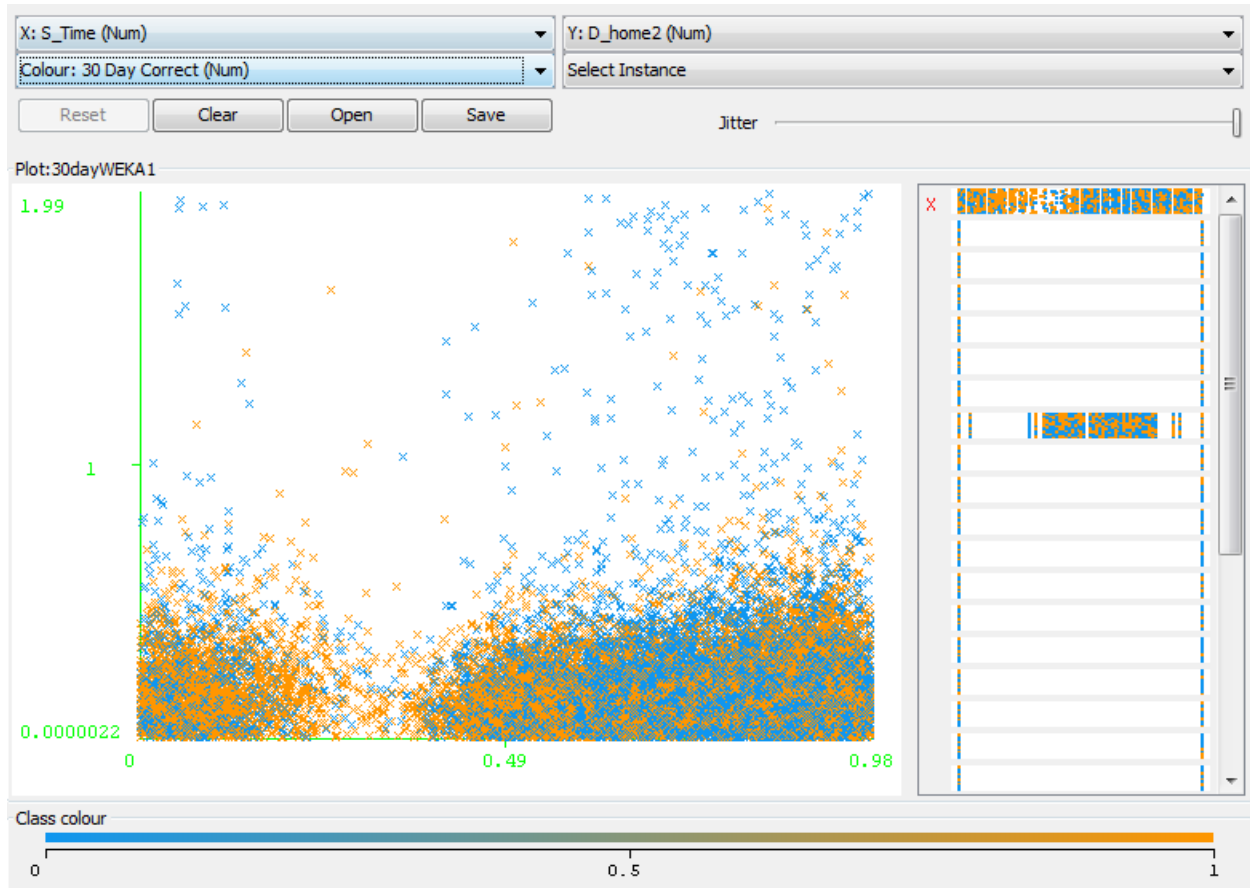
Appendix E: Purpose Graphs



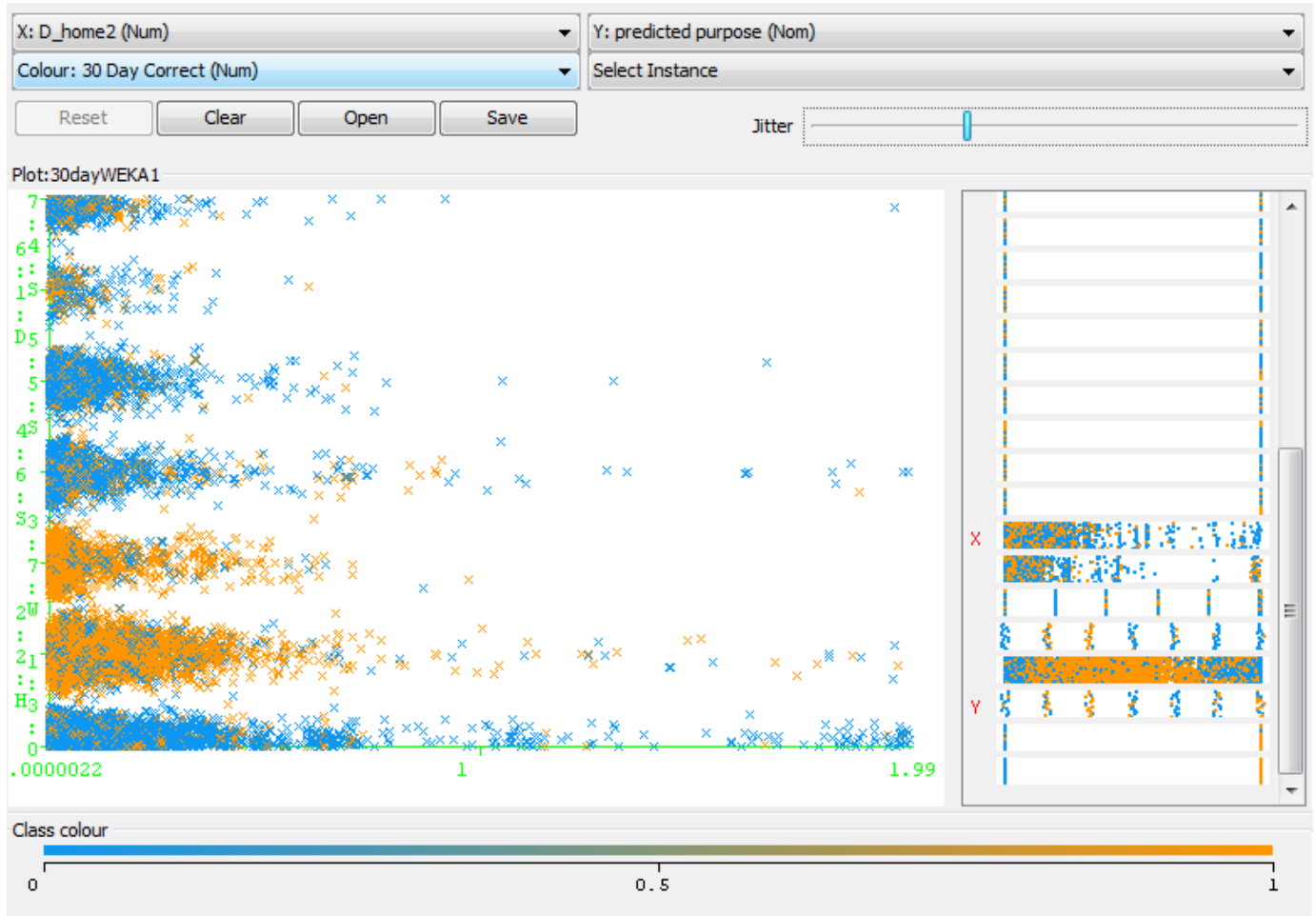
Predicted Purpose versus actual, by time of day



Purpose by Soak time



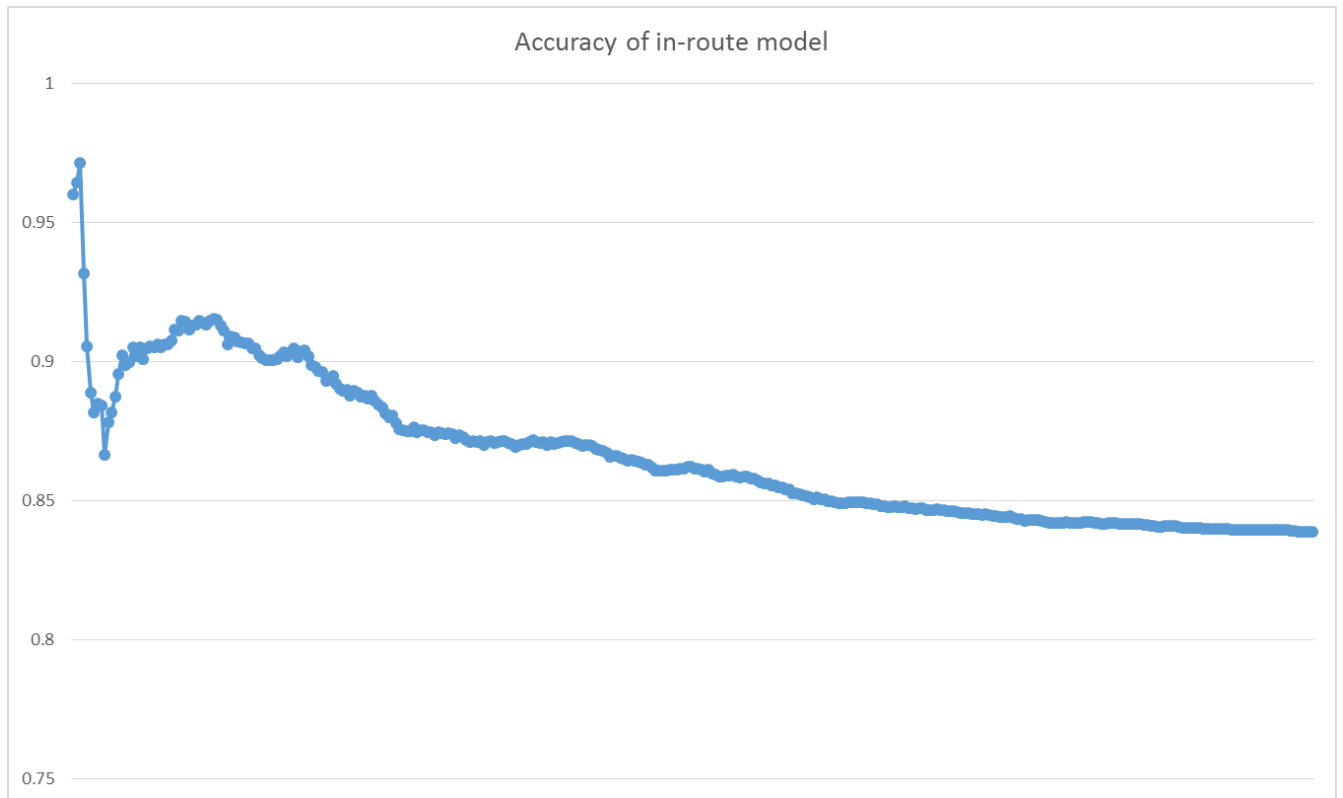
Accuracy by time of day and purpose

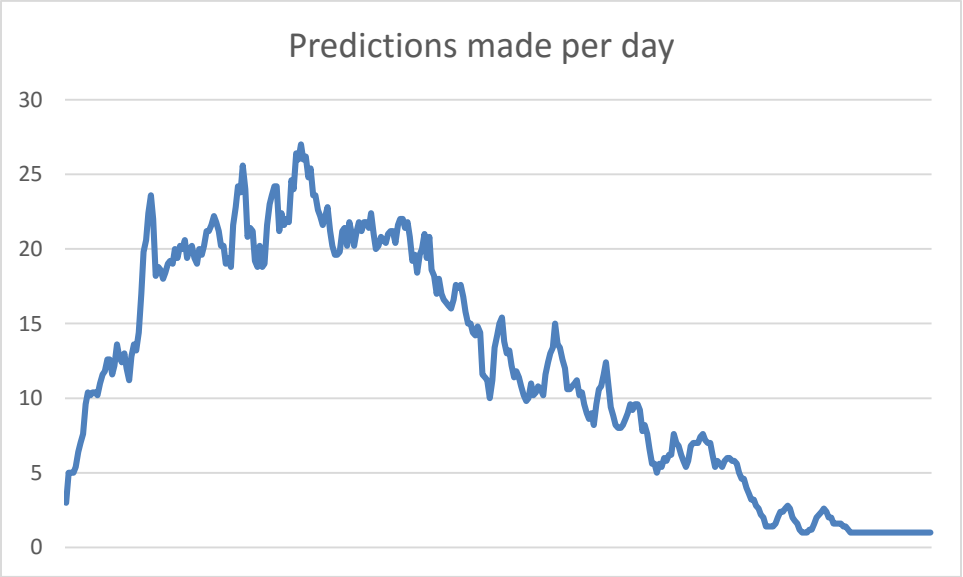
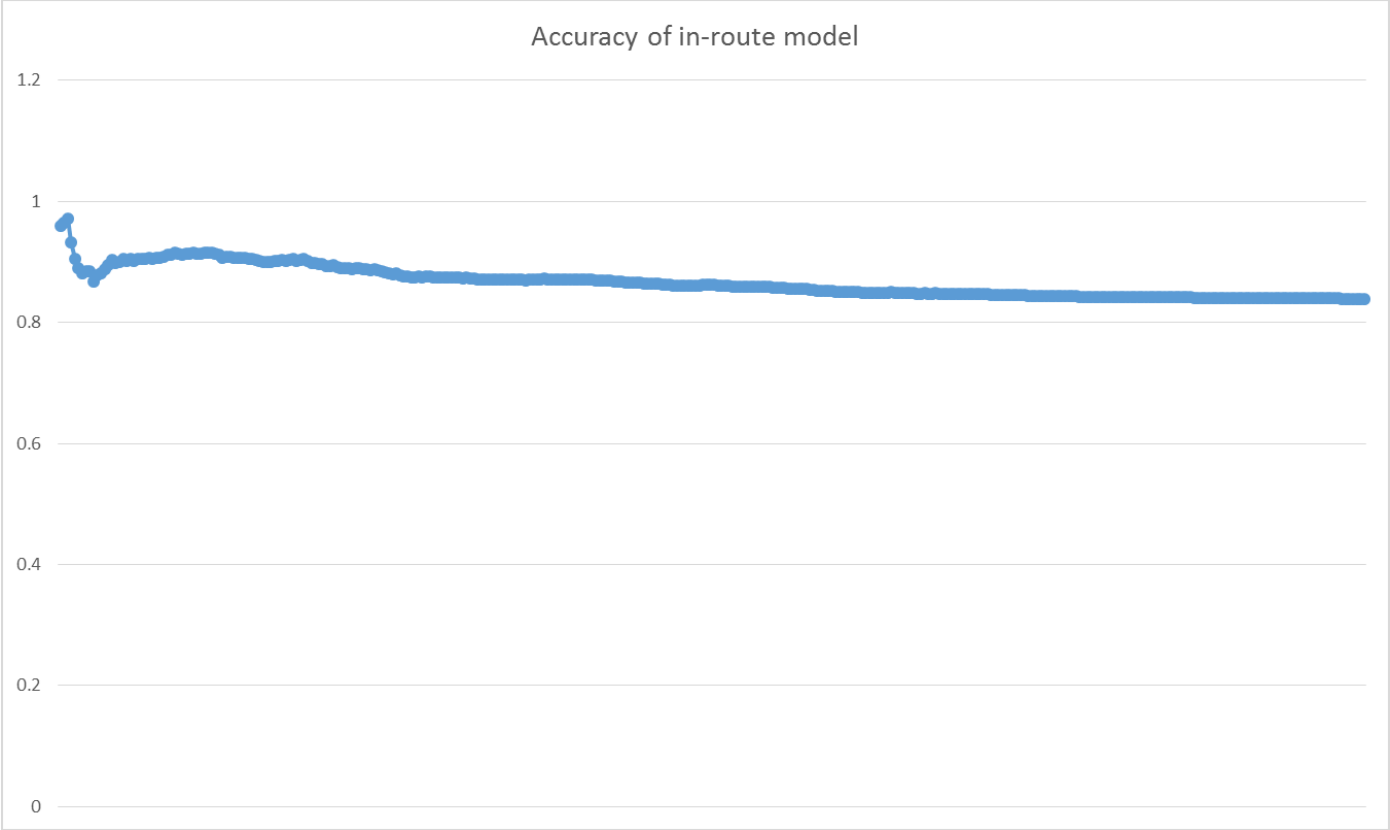


Purpose, distance to home, and accuracy

Appendix F: In-route prediction accuracy by day

Purpose_9_.85. Basically this is the highest accuracy model, with the highest possible threshold for route selection.





Appendix G: Raw In-route Model accuracies

Select by:	Origin	Origin	Origin	Origin	Origin	Origin	Origin	Purpose	Purpose	Purpose	Purpose	Purpose	Purpose	Purpose	None	None	None	None	None	None	None
Trip Amount	0.25	0.25	0.25	0.5	0.5	0.5	0.5	0.25	0.25	0.25	0.5	0.5	0.5	0.5	0.25	0.25	0.25	0.5	0.5	0.5	0.5
Match percentage	All	0.1	0.2	All	0.1	0.2	0.45	All	0.1	0.2	All	0.1	0.2	0.45	All	0.1	0.2	All	0.1	0.2	0.45
Total estimations	10583	9246	7238	11028	10097	9306	5930	10888	9025	6359	12356	11037	9793	5287	13319	11409	9723	14518	13483	12300	8953
Correct total	4193	3869	3087	4850	4683	4486	2958	6091	5435	4014	7116	6761	6289	3643	5041	4555	3895	6081	5914	5588	4140
Percentage accurate	0.3962	0.4185	0.4265	0.4398	0.4638	0.4821	0.4988	0.5594	0.6022	0.6312	0.5759	0.6126	0.6422	0.6890	0.3785	0.3992	0.4006	0.4189	0.4386	0.4543	0.4624
Percent Estimated	0.5865	0.5124	0.4011	0.6112	0.5596	0.5157	0.3286	0.6034	0.5002	0.3524	0.6848	0.6117	0.5427	0.2930	0.7381	0.6323	0.5388	0.8046	0.7472	0.6817	0.4962

Select by:	Origin	Origin	Origin	Origin	Origin	Origin	Origin	Purpose	Purpose	Purpose	Purpose	Purpose	Purpose	Purpose	None	None	None	None	None	None	None
Trip Amount	0.75	0.75	0.75	0.9	0.9	0.9	0.9	0.75	0.75	0.75	0.9	0.9	0.9	0.9	0.75	0.75	0.75	0.9	0.9	0.9	0.9
Match percentage	All	0.4	0.6	All	0.6	0.75	0.85	All	0.4	0.6	All	0.6	0.75	0.85	All	0.4	0.6	All	0.6	0.75	0.85
Total estimations	11389	8746	6950	11609	7924	6445	4616	13440	9648	7087	14078	8907	6948	4748	15241	12148	9311	15536	11269	9060	6486
Correct total	5459	4841	4043	5919	4930	4174	3008	7929	6655	5183	8463	6601	5355	3704	6923	6162	4941	7607	6454	5374	3808
Percentage accurate	0.479	0.554	0.582	0.510	0.622	0.648	0.652	0.590	0.690	0.731	0.601	0.741	0.771	0.780	0.454	0.507	0.531	0.490	0.573	0.593	0.587
Precent Estimated	0.631	0.485	0.385	0.643	0.439	0.357	0.256	0.745	0.535	0.393	0.780	0.494	0.385	0.263	0.845	0.673	0.516	0.861	0.625	0.502	0.359

Appendix H: Income Prediction Model (J48 in WEKA)

=== Run information ===

Scheme:weka.classifiers.trees.J48 -C 0.25 -M 2

Relation: nolanduse_create2

Instances: 20651

Attributes: 6

income

pop

labor

vacant

poverty

income_o

Time taken to build model: 1.84 seconds

=== Evaluation on test set ===

=== Summary ===

Correctly Classified Instances	33	27.7311 %
Incorrectly Classified Instances	86	72.2689 %
Kappa statistic	0.1329	
Mean absolute error	0.2173	
Root mean squared error	0.4201	
Relative absolute error	90.7797 %	
Root relative squared error	121.5587 %	
Total Number of Instances	119	

=== Detailed Accuracy By Class ===

TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
0.321	0.242	0.29	0.321	0.305	0.562	200K
0.389	0.069	0.5	0.389	0.438	0.639	125K
0.421	0.16	0.333	0.421	0.372	0.651	150K
0.158	0.09	0.25	0.158	0.194	0.547	100K
0.278	0.158	0.238	0.278	0.256	0.559	75K
0.077	0.104	0.083	0.077	0.08	0.417	50K
0	0.043	0	0	0	0.516	25K
Weighted Avg.	0.277	0.144	0.282	0.277	0.276	0.568

=== Confusion Matrix ===

a b c d e f g <-- classified as

9 1 5 2 6 4 1 | a = 200K

6 7 1 0 2 2 0 | b = 125K

4 1 8 2 3 1 0 | c = 150K
6 2 3 3 2 2 1 | d = 100K
2 1 3 2 5 2 3 | e = 75K
4 2 2 3 1 1 0 | f = 50K
0 0 2 0 2 0 0 | g = 25K

Appendix I: Census Metadata File

To save space, the URL is below for the full Census Metadata file used for Chapter 8 of this dissertation:

<http://www2.census.gov/geo/docs/maps-data/data/tiger/prejoined/ACSMetadata2010.txt>

Bibliography

1. Alvarez-Garcia, J. A., J. A. Ortega, L. Gonzalez-Abril, and F. Velasco. "Trip destination prediction based on past GPS log using a Hidden Markov Model." *Expert Systems with Applications* 37, no. 12 (2010): 8166-71.
2. Anderson, R., and Costinett, P. (2007). "Ohio Statewide Passenger Transit Calibration", presented at the 11th TRB National Transportation Planning Applications Conference, Daytona Beach, Florida.
3. Arentze, T. A. and Timmermans, H.J.P. (2000). ALBATROSS: A Learning-Based Transportation Oriented Simulation System, Eindhoven, EIRSS.
4. Arentze, T. A. and Timmermans, H.J.P. (2003). Modeling learning and adaptation processes in activity-travel choice: A framework and numerical experiment, *Transportation* 30: 37–62.
5. Arentze, T. A. and Timmermans, H.J.P. (2004). A learning-based transportation oriented simulation system, *Transportation Research Part B* 38: 613–633.
6. Arentze, T. A. and Timmermans, H.J.P. (2005). Information gain, novelty seeking and travel: a model of dynamic activity-travel behavior under conditions of uncertainty, *Transportation Research Part A* 39: 125–145.
7. Arentze, T. A. and Timmermans, H.J.P. (2005). Modeling learning and adaptation in transportation contexts, *Transportmetrica* 1(1): 13–22.
8. Arentze, T., Pelizaro, C. and Timmermans, H. (2005). Implementation of a model of dynamic activity-travel rescheduling decisions: an agent-based micro-simulation framework, *Proceedings of CUPUM 05, Computers in Urban Planning and Urban Management*, 30-Jun-2005, London, CASA Centre for Advanced Spatial Analysis - University College London, London, pp. paper– 48.
9. Ashiabor, S., Baik, H., & Trani, A. (2007). Logit models to forecast nationwide intercity travel demand in the United States. *Journal of the Transportation Research Record*, 2007, 1–12.
10. Ashok, K., and M. E. Ben-Akiva. "Alternative approaches for real-time estimation and prediction of time-dependent Origin-Destination flows." *Transportation Science* 34, no. 1 (2000): 21-36.
11. Auld, J., Mohammadian K., Doherty, S. (2008). Analysis of activity conflict resolution strategies, 1-24.
12. Auld, J., Williams, C., Mohammadian, K., Nelson, P. (2008). AN AUTOMATED GPS-BASED PROMPTED RECALL SURVEY WITH LEARNING ALGORITHMS, 500.
13. Axhausen, Kay W.; Schoenfelder, S.; Wolf, J.; Oliveira, M.; Samaga, U. "80 weeks of GPS-traces: approaches to enriching the trip information." Submitted to the 83rd Transportation Research Board Meeting. ETH, Eidgenössische Technische Hochschule Zürich, Institut für Verkehrsplanung und Transportsysteme, 2003.

14. Bachmann, Anja , Christian Borgelt, and Győző Gidófalvi. "Incremental Frequent Route Based Trajectory Prediction." In Proceedings of the Sixth ACM SIGSPATIAL International Workshop on Computational Transportation Science, 49-54. Orlando, FL, USA: ACM, 2013.
15. Becker, G. S. (1965) A Theory of the Allocation of Time.pdf.
16. Ben-Akiva (1999). Activity Based Disaggregate travel demand model system with activity schedules, Massachusetts Institute of Technology, transportation Research Part A 35 (2000) 1-28.
17. Ben-Akiva, M. (2008). Travel Demand Modeling, Transportation Systems Analysis: Demand & Economics, Massachusetts Institute of Technology 1.201 / 11.545 / ESD.210
18. Ben-Akiva, M. and de Palma, A. (1986). Analysis of a dynamic residential location choice model with transaction costs. *Journal of Regional Science*, 26: 321–341.
19. Ben-Akiva, M., de Palma, A. and Kaysi, I. (1991). Dynamic network models and driver information systems, *Transportation Research Part A* 25(5): 251–266.
20. Bohlooli, Ali, and Kamal Jamshidi. "A GPS-free method for vehicle future movement directions prediction using SOM for VANET." *Applied Intelligence* 36, no. 3 (2012): 685-97.
21. Bohte, Wendy, and Kees Maat. "Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands." *Transportation Research Part C: Emerging Technologies* 17, no. 3 (2009): 285-97.
22. Boulmakoul, A. (2012). Moving Object Trajectories Meta-Model and Spatio-Temporal Queries. *International Journal of Database Management Systems*, 4(2), 35–54. doi:10.5121/ijdms.2012.4203
23. Boulmakoul, A. (2012). Moving Object Trajectories Meta-Model and Spatio-Temporal Queries. *International Journal of Database Management Systems*, 4(2), 35–54.
24. Bricka, S. (n.d.). Variations in Long-Distance Travel, 197-206. TRB Transportation Research Circular E-C026—Personal Travel: The Long and Short of It
25. Bureau of Transportation Statistics. (1997). 1995 American travel survey. Washington, DC: US Department of Transportation. Retrieved February 11, 2009, from http://www.bts.gov/publications/1995_american_travel_survey/
26. Burris, M. W., & Pendyala, R. M. (2002). Discrete choice models of traveler participation in differential time of day pricing programs. *Transport Policy*, 9(3), 241-251. doi:10.1016/S0967-070X(02)00002-1
27. Burris, M. W., & Stockton, B. R. (n.d.). HOT Lanes in Houston — Six Years of Experience, 1-21.
28. Burris, M. W., Konduru, K. K., & Swenson, C. R. (2004). Long-Run Changes in Driver Behavior Due to Variable Tolls, (1864), 78-85.

29. Cambridge Systematics. (2008). National travel demand forecasting model phase I final scope. NCHRP Project 08-36, Task 70, National Cooperative Highway Research Program. Washington, DC: Transportation Research Board.
30. Cao, L., & Krumm, J. (2009). From GPS traces to a routable road map. Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '09, 3. New York, New York, USA: ACM Press.
doi:10.1145/1653771.1653776
31. Chen, B. (2011). Study of Interregional Long-Distance Commuting Using NHTS data.
32. Cirillo, C., & Axhausen, K. W. (2006). Evidence on the distribution of values of travel time savings from a six-week diary. *Transportation Research Part A: Policy and Practice*, 40(5), 444–457. doi:10.1016/j.tra.2005.06.007
33. Costinett, P. and Stryker, A. (2007). "Calibrating the Ohio Statewide Travel Model", presented at the 11th TRB National Transportation Planning Applications Conference, Daytona Beach, Florida.
34. Deng, Zhongwei, and Minhe Ji. "Deriving rules for trip purpose identification from GPS travel survey data and land use data: A machine learning approach." 7th International Conference on Traffic and Transportation Studies, (ICTTS 2010), August 3, 2010 - August 5, 2010, 2010.
35. E. Horvitz and J. Krumm, "Some help on the way: opportunistic routing under uncertainty," in Proceedings of the 2012 ACM Conference on Ubiquitous Computing, ser. UbiComp '12, 2012, pp. 371–380.
36. Emmerink, R. H. M., Axhausen, K. W., Njikamp, P. and Rietveld, P. (1995). The potential of information provision in a simulated road transport network with non-recurrent congestion, *Transportation Research Part C* 3(5): 293–309.
37. Erhardt, G. D., Freedman, J., Stryker, A., Fujioka, H., & Anderson, R. (2007). Ohio Long-Distance Travel Model, (2003), 130-138. doi:10.3141/2003-16
38. Erhardt, G., & Freedman, J. (n.d.). CALIBRATION AND APPLICATION OF THE OHIO LONG DISTANCE TRAVEL MODEL.
39. Erhardt, G., Freedman, J., Stryker, A., Fujioka, H. and Anderson, R. (2007). "The Ohio Long Distance Travel Model", accepted for publication in *Transportation Research Record*, Washington, D.C.
40. Fadaei Oshyani, Masoud, Marcus Sundberg, and Anders Karlstrom. "Estimating flexible route choice models using sparse data." 2012 15th International IEEE Conference on Intelligent Transportation Systems, ITSC 2012, September 16, 2012 - September 19, 2012, 2012.
41. Filev, D., F. Tseng, J. Kristinsson, and R. McGee. "Contextual on-board learning and prediction of vehicle destinations." *Computational Intelligence in Vehicles and Transportation Systems (CIVTS)*, 2011 IEEE Symposium on, 2011.

42. Froehlich, Jon, and John Krumm. "Route prediction from trip observations." SAE SP 2193 (2008): 53.
43. Gambs, Sébastien, Marc-Olivier Killijian, and Miguel Núñez del Prado Cortez. "Next place prediction using mobility markov chains." *Proceedings of the First Workshop on Measurement, Privacy, and Mobility*. ACM, 2012.
44. Gao, H. (2012). Mobile Location Prediction in Spatio-Temporal Context, (2), 1–4.
45. Giaimo, G., R. Anderson, L. Wargelin, and P. Stopher. Will It Work? Pilot Results from First Large-Scale Global Positioning System–Based Household Travel Survey in the United States. In *Transportation Research Record: Journal of the Transportation Research Board*, No. 2176, Transportation Research Board of the National Academies, Washington, D.C.,
46. Giaimo, G.T., & Schiffer, R. (Eds.). (2005, August). Statewide travel demand modeling: A peer exchange. Transportation Research Circular, #E-C075. Washington, DC: Transportation Research Board.
47. Gong, Xiaowen, and Sathiamoorthy Manoharan. "On predicting vehicle tracks." 13th IEEE Pacific Rim Conference on Communications, Computers and Signal Processing, PACRIM 2011, August 23, 2011 - August 26, 2011, 2011.
48. Griffin, Terry, and Yan Huang. "A decision tree classification model to automate trip purpose derivation." 18th International Conference on Computer Applications in Industry and Engineering 2005, CAINE 2005, November 9, 2005 - November 11, 2005, 2005.
49. Guangtao, Xue, Li Zhongwei, Zhu Hongzi, and Liu Yunhuai. "Traffic-Known Urban Vehicular Route Prediction Based on Partial Mobility Patterns." *Parallel and Distributed Systems (ICPADS)*, 2009 15th International Conference on, 2009.
50. Hariharan, R., Krumm, J., & Horvitz, E. (2005). Web-Enhanced GPS, T. Strang and C. Linnhoff-Popien (Eds.): *LoCA 2005*, LNCS 3479, pp. 95 – 104, 2005.
51. Hess, S., Bierlaire, M., & Polak, J. W. (2005). Estimation of value of travel-time savings using mixed logit models. *Transportation Research Part A: Policy and Practice*, 39(2-3), 221-236. doi:10.1016/j.tra.2004.09.007
52. Horowitz, J. L. (1984). The stability of stochastic equilibrium in a two-link transportation network, *Transportation Research Part B* 18(1): 13–28.
53. Ibe, Oliver (2013) *Markov Processes for Stochastic Modeling*. Elsevier Inc. ISBN 978-0-12-407795-9
54. Jeung, Hoyoung, Man Lung Yiu, Xiaofang Zhou, and Christian S. Jensen. "Path prediction and predictive range querying in road network databases." *VLDB Journal* 19, no. 4 (2010): 585-602.

55. Karbassi, A., and M. Barth. "Vehicle route prediction and time of arrival estimation techniques for improved transportation system management." Intelligent Vehicles Symposium, 2003. Proceedings. IEEE, 2003.
56. Kaul, Sanjit, Marco Gruteser, Vinuth Rai, and John Kenney. "On predicting and compressing vehicular GPS traces." 2010 IEEE International Conference on Communications Workshops, ICC 2010, May 23, 2010 - May 27, 2010, 2010.
57. Krause, C. "A Positive Model of Route Choice Behavior and Value of Time Calculation Using Longitudinal GPS Survey Data". Digital Repository at the University of Maryland. 2012
58. Krumm, J. and E. Horvitz, "Predestination: Inferring destinations from partial trajectories," in Proc. UbiComp, 2006, pp. 243–260.
59. Krumm, J. and E. Horvitz, "Predestination: Where do you want to go today?" IEEE Computer, pp. 105–107, 2007.
60. Krumm, John, Robert Gruen, and Daniel Delling. "From destination prediction to route prediction." Journal of Location Based Services 7, no. 2 (2013): 98-120.
61. Krumm, John. "A markov model for driver turn prediction." SAE 2008 World Congress, April 14-17, 2008 Detroit MI USA. Paper number 2008-01-0195 (2008).
62. Krumm, John. "Trajectory Analysis for Driving." Chap. 7 In Computing with Spatial Trajectories, edited by Yu Zheng and Xiaofang Zhou, 213-41: Springer New York, 2011.
63. Lapparent, M. D. (2002). Nonlinearities in the Valuations of Travel Times at the Individual Level .
64. Lei Zhang, Frank Southworth, Chenfeng Xiong & Anthon Sonnenberg (2012): Methodological Options and Data Sources for the Development of Long-Distance Passenger Travel Demand Models: A Comprehensive Review, Transport Reviews: A Transnational Transdisciplinary Journal, 32:4, 399-433
65. Lei, P.-R., Shen, T.-J., Peng, W.-C., & Su, I.-J. (2011). Exploring Spatial-Temporal Trajectory Model for Location Prediction. 2011 IEEE 12th International Conference on Mobile Data Management, 58–67. doi:10.1109/MDM.2011.61
66. Li, H., Guensler, R., & Ogle, J. (2005). An Analysis of Morning Commute Route Choice Patterns Using GPS Based Vehicle Activity Data. TRB, 30332(404).
67. Li, Shi-Bao, and Li Hong. "Algorithm of route discovery based on distance prediction in MANET." Tongxin Xuebao/Journal on Communications 31, no. 11 (2010): 180-87.
68. Liao, L., Fox, D., & Kautz, H. (2003). Learning and Inferring Transportation Routines.
69. Lomax, T., Schrank, D., Eisele, B. (2012). Annual Urban Mobility Report. Texas A&M Transportation Institute.
<http://d2dtl5nnlpfr0r.cloudfront.net/tti.tamu.edu/documents/mobility-report-2012.pdf>

70. Lu, Yijing, and Lei Zhang. "Trip Purpose Estimation for Urban Travel in the U.S.: Model Development, NHTS Add-on Data Analysis, and Model Transferability Across Different States." Transportation Research Board 93rd Annual Meeting, 2014.
71. Lu, Yijing, Shanjiang Zhu, and Lei Zhang. "Imputing Trip Purpose Based on GPS Travel Survey Data and Machine Learning Methods." Transportation Research Board 92nd Annual Meeting, 2013.
72. Mallett, W. (n.d.). Long-Distance Travel by Low-Income Households. TRB Transportation Research Circular E-C026—Personal Travel: The Long and Short of It (169-177)
73. McGowen, Patric Tracy. "Predicting activity types from GPS and GIS data." University of California, Irvine, 2006.
74. Meghanathan, Natarajan. "A location prediction based routing protocol and its extensions for multicast and multi-path routing in mobile ad hoc networks." Ad Hoc Networks 9, no. 7 (2011): 1104-26.
75. Merah, AmarFarouk, Samer Samarah, Azzedine Boukerche, and Abdelhamid Mammeri. "A sequential Patterns Data Mining Approach Towards Vehicular Route Prediction in VANETs." Mobile Networks and Applications 18, no. 6 (2013): 788-802.
76. Miklucak, Toma, Michal Gregor, and Ale Janota. "Using neural networks for route and destination prediction in intelligent transport systems." 12th International Conference on Transport Systems Telematics, TST 2012, October 10, 2012 - October 13, 2012, 2012.
77. Miller, E.J. (2004). The trouble with intercity travel demand models. Transportation Research Record, 1895, 94–101.
78. Miyashita, Koichi, Tsutomu Terada, and Shojiro Nishio. "A map matching algorithm for car navigation systems that predict user destination." 22nd International Conference on Advanced Information Networking and Applications Workshops/Symposia, AINA 2008, March 25, 2008 - March 28, 2008, 2008.
79. Moiseeva, Anastasia, Joran Jessurun, and Harry Timmermans. "Semiautomatic Imputation of Activity Travel Diaries: Use of Global Positioning System Traces, Prompted Recall, and Context-Sensitive Learning Algorithms." Transportation Research Record: Journal of the Transportation Research Board, no. 2183 (2010): pp 60-68.
80. Montini, Lara, Nadine Rieser-Schüssler, Andreas Horni, and Kay W. Axhausen. "Trip Purpose Identification from GPS Tracks." Transportation Research Board 93rd Annual Meeting, 2014.
81. Morzy, M. (2007). Mining Frequent Trajectories of Moving Objects, 667–680.
82. Murakami, E. "Can using global positioning system (GPS) improve trip reporting?." Transportation research. Part E, Logistics and transportation review 7.2-3 (1999):149.
83. Murakami, E., & Young, J. (1997). Daily Travel by Persons with Low Income Daily Travel by Persons with Low Income.

84. Nakayama, S., Kitamura, R. and Fujii, S. (2001). Drivers' route choice rules and network behavior: Do drivers become rational and homogeneous through learning?, *Transportation Research Record: Journal of the Transportation Research Board* 1752: 62–68.
85. Oliveira, M. G. S., Vovsha, P., Wolf, J., Birotker, Y., Givon, D., & Paasche, J. (2011). Global Positioning System-Assisted Prompted Recall Household Travel Survey to Support Development of Advanced Travel Model in Jerusalem, Israel. *Transportation Research Record: Journal of the Transportation Research Board*, 2246(-1), 16-23.
doi:10.3141/2246-03\
86. Outwater, M., Sall, E., Tierney, K., Bradley, M., Kuppam, A., & Modugula, V. (n.d.). *California Statewide Model for High-Speed Rail*, 3(1), 58-83.
87. Patterson, D., et al., Inferring High-Level Behavior from Low-Level Sensors, in *UbiComp 2003: Ubiquitous Computing*. 2003, Springer: Seattle, Washington USA. P. 73-89
88. Persad-Maharaj, Narin, Sean J. Barbeau, Miguel A. Labrador, Philip L. Winters, Rafael Perez, and Nevine Labib Georggi. "Real-time travel path prediction using gps-enabled mobile phones." 15th World Congress on Intelligent Transport Systems and ITS America Annual Meeting 2008, November 16, 2008 - November 20, 2008, 2008.
89. Pfoser, D. (2004). Modeling , Storing and Mining Moving Object Databases. Sotiris Brakatsoulas Research Academic Computer Technology Institute. Organizing the Moving Object Database.
90. Pfoser, D. (2004). Modeling , Storing and Mining Moving Object Databases. Sotiris Brakatsoulas Research Academic Computer Technology Institute. Organizing the Moving Object Database.
91. Picado, R., Freedman, J., Stryker, A. and Erhardt, G. (2007). "Design, Estimation and Calibration of the Ohio Statewide Short Distance Travel Models", presented at the 11th TRB National Transportation Planning Applications Conference, Daytona Beach, Florida.
92. Prato, C. G. (2009). Route choice modeling : past , present and future research directions, 2(1), 65-100.
93. Qiu, Disheng, Paolo Papotti, and Lorenzo Blanco. "Future Locations Prediction with Uncertain Data." Chap. 27 In *Machine Learning and Knowledge Discovery in Databases*, edited by Hendrik Blockeel, Kristian Kersting, Siegfried Nijssen and Filip Železný. Lecture Notes in Computer Science, 417-32: Springer Berlin Heidelberg, 2013.
94. Richardson, A., SEETHALER, R. (n.d.). Estimating Long-Distance Travel Behavior from the Most Recent Trip. TRB Transportation Research Circular E-C026—Personal Travel: The Long and Short of It
95. Roth, Michael, Fredrik Gustafsson, and Umut Orguner. "On-road trajectory generation from GPS data: A particle filtering/smoothing application." 15th International

Conference on Information Fusion, FUSION 2012, September 7, 2012 - September 12, 2012, 2012.

96. Sanchez-Cambronero, S., A. Rivas, I. Gallego, and J. M. Menendez. "Predicting traffic flow in road networks using Bayesian networks with data from an optimal plate scanning device location." 2nd International Conference on Agents and Artificial Intelligence, ICAART 2010, January 22, 2010 - January 24, 2010.
97. Satellite, T., & Programme, A. (2011). International Recommendations for Tourism Statistics 2008 Draft Compilation Guide, (March).
98. Schönfelder, S., H. Li, R. Guensler, J. Ogle and K.W. Axhausen (2005) Analysis of Commute
99. Schonfelder, S., Li, H., Guensler, R., Ogle, J., Axhausen, K (2005). Analysis of Commute Atlanta Instrumented Vehicle GPS Data: destination Choice Behavior and Activity Spaces. Arbeitsberichte Verkehrs- und Raumplanung, 303, IVT, ETH, Zürich.
100. Schüssler, Nadine; Axhausen, Kay W. "Processing GPS raw data without additional information." Eidgenössische Technische Hochschule, Institut für Verkehrsplanung und Transportsysteme, 2008.
101. Seshadri, Anand, Hojong Baik, and Antonio Trani. "A model to estimate the origin-transfer-destination route flows and airport O-D segment flows across the continental United States." 53rd Air Traffic Control Association Annual Conference 2008, November 2, 2008 - November 5, 2008, 2008.
102. Shen, Li, and Peter R. Stopher. "A process for trip purpose imputation from Global Positioning System data." Transportation Research Part C: Emerging Technologies 36 (2013): 261-67.
103. Simmons, Reid, Brett Browning, Yilu Zhang, and Varsha Sadekar. "Learning to predict driver route and destination intent." ITSC 2006: 2006 IEEE Intelligent Transportation Systems Conference, September 17, 2006 - September 20, 2006, 2006.
104. Small, K., Winston, C., & Yan, J. (2005). Uncovering the distribution of motorists' preferences for travel time and reliability b, 73(4), 1367-1382.
105. Southworth, F., & Hu, P.S. (2010). The American long distance personal travel data and modeling program: A roadmap. Revised draft. Prepared for the Federal Highway Administration, Washington, DC.
106. Stopher, P., E. Clifford, J. Zhang, and C. Fitzgerald. "Deducing mode and purpose from GPS data". Institute of Transport and Logistics Studies Working Paper. University of Sydney, 2008.
107. Stopher, P., Q. Jiang, and Camden FitzGerald. "Processing GPS data from travel surveys." AUSTRALASIAN TRANSPORT RESEARCH FORUM (ATRF), 28TH, 2005, SYDNEY, NEW SOUTH WALES, AUSTRALIA 28 (2005): 17P.

108. Terada, Tsutomu, Masakazu Miyamae, Yasue Kishino, Kohei Tanaka, Shojiro Nishio, Takashi Nakagawa, and Yoshihisa Yamaguchi. "Design of a car navigation system that predicts user destination." 7th International Conference on Mobile Data Management, 2006. MDM 2006, May 10, 2006 - May 12, 2006, 2006.
109. Tiwari, V. S., A. Arya, and S. Chaturvedi. "Route prediction using trip observations and map matching." Advance Computing Conference (IACC), 2013 IEEE 3rd International, 2013.
110. Torkkola, K., et al., Traffic Advisories Based on Route Prediction, in Workshop on Mobile Interaction with the Real World (MIRW 2007). 2007: Singapore.
111. Wolf, J. (2000). Using GPS Data Loggers To Replace Travel Diaries In the Collection of Travel Data by, (July).
112. Wolf, J., Guensler, R., Bachman, W. Elimination of the Travel Diary : An Experiment to Derive Trip Purpose From GPS Travel Data, (01).
113. Wolf, J., Loechl, M., Myers, J., Arce, C. TRIP RATE ANALYSIS IN GPS-ENHANCED PERSONAL TRAVEL By, (August 2001).
114. Wolf, J., M. Loechl, M. Thompson, and C. Arce. Trip Rate Analysis in GPS-Enhanced Personal Travel Surveys. In Transport Survey Quality and Innovation (P. Stopher and P. M. Jones, eds.), Pergamon, Oxford, United Kingdom, 2003, pp. 483–498.
115. Wolf, Jean, Randall Guensler, and William Bachman. "Elimination of the travel diary: Experiment to derive trip purpose from Global Positioning System travel data." Transportation Research Record 1768 (2001): 125-34.
116. World Tourism Organization. (2011). International Recommendations for Tourism Statistics 2008 Draft Compilation Guide, (March).
117. Wu, Hsin-Te, and Wen-Shyong Hsieh. "Location-based vehicular moving predictions for wireless communication." International Journal of Ad Hoc and Ubiquitous Computing 10, no. 4 (2012): 197-206.
118. Xu, Y., Zuyeva, L., Kall, D., Elango, V., & Guensler, R. (2009). Mileage Based Value Pricing: Phase II Case Study Implications of Commute Atlanta Project. TRB, (18).
119. Xue, A. Y., Zhang Rui, Zheng Yu, Xie Xing, Huang Jin, and Xu Zhenghua. "Destination prediction by sub-trajectory synthesis and privacy protection against such prediction." Data Engineering (ICDE), 2013 IEEE 29th International Conference on, 2013.
120. Yang, J., & Hu, M. (2006). TrajPattern : Mining Sequential Patterns, 664–681.
121. Zhang, L. (2006a). Search, information, learning and knowledge in travel-decision making. Ph.D. Dissertation, Department of Civil Engineering, University of Minnesota.
122. Zhang, L. (2006b). An agent-based behavioral model of spatial learning and route choice. In the compendium of papers CD of the 85th Transportation Research Board Annual Meetings, Washington, DC.

123. Zhang, L. (2006c). Traffic diversion effect of ramp metering at individual and system levels. *Transportation Research Record: Journal of the Transportation Research Board*, 2012: 20-29.
124. Zhang, L. (2007). Developing a positive approach to travel demand analysis: Silk theory and behavioral user equilibrium, in R. E. Allsop, B. M. G. H. and B. G. Heydecker (eds), *Transportation and Traffic Theory*, Elsevier, chapter 34, pp. 791–812.
125. Zhang, L. and Levinson, D. (2004). Agent-Based Approach to Travel Demand Modeling. In *Transportation Research Record: Journal of the Transportation Research Board*, Vol. 1898, 2004, pp. 28–36.
126. Zhang, L., Levinson, D., and Zhu, S. (2008). Agent-Based Model of Price Competition, Capacity Choice, and Product Differentiation on Congested Networks (August 1, 2008). *Journal of Transport Economics and Policy*, 42(3): 435-461
127. Zhang, L., Xiong, C., & Berger, K. (2010). Multimodal inter-regional origin-destination demand estimation: A review of methodologies and their applicability to national-level travel analysis in the US WCTR. Lisbon, Portugal, July, 2010.
128. Zhang, M., Chen, B. (2011). Understanding Emerging Commuting Trends in a Weekly Travel Decision Frame--Implications for Mega Region Transportation Planning September 2011 Ming Zhang and Binbin Chen Report 161127 Center for Transportation Research University of Texas at Austin Southw (Vol. 7).
129. Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W. (2008). Understanding Mobility Based on GPS Data, (49).
130. Zheng, Y., Li, Q., Chen, Y., Xie, X., & Ma, W. (2008). Understanding Mobility Based on GPS Data, (49).
131. Zheng, Y., Liu, L., Wang, L., & Xie, X. (2008). Learning Transportation Mode from Raw GPS Data for Geographic Applications on the Web, 247-256.
132. Zhout, X. (2008). A Hybrid Prediction Model for Moving Objects. *IEEE 24th International Conference on Data Engineering. ICDE 2008*, 70–79. 978-1-4244-1837-4
133. Qi, Wu, Boriboonsomsin, Barth. A Novel Blended Real-time Energy Management Strategy for Plug-in Hybrid Electric Vehicle Commute Trips. 2015 IEEE 18th International Conference on Intelligent Transportation Systems.
134. Alvarez-Garcia, Juan Antonio, et al. "Trip destination prediction based on past GPS log using a hidden markov model." *Expert Systems with Applications* 37.12 (2010): 8166-8171.