Transportation Research Part B 123 (2019) 349–361

Contents lists available at ScienceDirect

Transportation Research Part B

journal homepage: www.elsevier.com/locate/trb

Short-term travel behavior prediction with GPS, land use, and point of interest data

Cory M. Krause[a,*], Lei Zhang[b]

[a] Noblis, 600 Maryland Avenue, S.W., Suite 700E, Washington, DC 20024, USA
[b] Department of Civil and Environmental Engineering, Maryland Transportation Institute, University of Maryland, 1173 Martin Hall, College Park, MD 20742, USA

ARTICLE INFO

Article history:
Received 22 March 2018
Revised 18 June 2018
Accepted 29 June 2018
Available online 9 July 2018

Keywords:
Destination prediction
Trip purpose
GPS data
Markov model
Land use
Point of interest

ABSTRACT

In everyday travel, U.S. commuters will each spend 38 h a year stuck in traffic and waste over $800 in fuel (TTI, 2015). Yet, despite this statistic, the regular commute of drivers is often predictable, leading many federal projects to aim at alleviating congestion through traveler information and intelligent transportation systems (e.g., INFLO, Queue WARN, CACC, EnableATIS, ATIS2.0). Short-term destination prediction is a developing field of research that can improve these approaches through real-traveler information, such as route, traffic incidence, and congestion levels. 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. To study these concepts, a database of GPS driving traces (260 participants for 70 days) was collected. To model the user's trip purpose in the prediction algorithm, 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 was calculated with machine learning methods. To take advantage of this data source, a new prediction model structure was developed that uses trip purpose when it is available and that falls back on traditional spatial temporal Markov models when it is not. For the first time, there is an understanding of "why" a trip is taken (not just "where" and "when"), allowing the use of "why" in the prediction model. This paper explores the baseline model followed by the inclusion of trip purpose. First, a baseline tiered time origin model was developed using the Markov Chain approach. This modelling structure allows for a short training period of current modeling techniques. The other major advantage to this structure is it allows for easy implementation of the trip purpose module. Then, a machine learning technique derived the trip purpose on 5-, 15- and 30-trip learning sets, followed by results organized by purpose, time, and origin. The machine learning technique does not require future land use data and is feasible for applicable use. This model is the first to use trip purpose to make a short-term destination prediction in pseudo real-time. Results show improved accuracy and speed over the current start-of-trip destination prediction models.

[*] Corresponding author.
E-mail addresses: cory.krause@noblis.org (C.M. Krause), lei@umd.edu (L. Zhang).

https://doi.org/10.1016/j.trb.2018.06.012

1. Introduction

Moving vehicle data are becoming more prevalent and accessible as traveling with GPS-enabled smart phones and in-vehicle systems become the norm. With the increase in vehicle tracking, many new applications are becoming available. Vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communication are now a primary focus of research at institutions such as the U.S. Federal Highway Administration (FHWA) 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. Private companies are also tracking speed profiles and collecting millions of GPS traces on public roads for the purpose of producing real-time traffic information (TTI, 2015). 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 the future location of a vehicle and a traveler can be accurately predicted, smarter suggestions of travel routing can be provided and travel demand information more accurately predicted in real time. Also, as in-vehicle computer systems become more advanced, travel information and history can be stored in the 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 the determined route. The TMC would obtain not just up-to-date information, but future vehicle location: the ability to see network conditions before they occur. With accurately-predicted destination in real time, upon entering the car, a passenger's on-board GPS might greet them with a suggestion to take an alternate route to work due to unexpected congestion. Even before the passenger enters their vehicle, their smart phone may suggest an alternative departure time knowing where you will go next. Mobile advertising could also benefit from real-time destination prediction algorithms. This research has applications for individual users, the network as a whole, transportation management centers, and mobility-as-a-service companies.

This paper explores the prediction of vehicle destination in real time. Models currently exist that use past GPS traces to predict future location through learning or probabilistic models. However, trip purpose and land use data have not been applied to these approaches to improve their accuracy. This research is the first to incorporate imputed trip purpose information into short-term destination prediction algorithms. The derivation of trip purpose in this paper is based on passively collected vehicle traces and land use information around the vehicle from open source data. Through machine learning, trip purpose is first imputed and then Markov models are derived using geospatial (how often does a passenger arrive at a location, and why) and temporal information (what time does the passenger tend to arrive there?). By combining imputed trip purpose with Markov model approaches, this research can produce significantly increased destination prediction accuracy after a short period of travel behavior learning. This approach also improves prediction accuracy where previous models have struggled; for trips such as school, social, shopping, and driving. This is accomplished with the integration of land use data to previous locations and with machine learning methods that associate locations with particular trip types. The methods developed in this paper increase accuracy for real-time destination prediction by 4 to 7 percentage points on average, with some types of trips seeing an accuracy increase by as much as 15 percentage points. The model is also shown to learn user behavior and improves its own predictive accuracy over time.

The remainder of this paper is structured as follows. First, a literature review is performed on existing real-time destination prediction models, focusing on the gap in trip purpose and early trip/pre-trip prediction procedures. The Data section will explain the GPS, demographic, and land use data employed in this research. The methods for deriving trip purpose is then discussed, followed by an introduction of the hierarchical Markov model for destination prediction. Results from the model and the included trip purpose module are predicted and discussed. Finally, conclusions and suggestions for future work are offered.

2. Background

2.1. Previous research on location prediction

There are several approaches for real-time destination prediction. Trajectory data and historical GPS points are used to create a basis for future trips. Ashbrook and Starner (2003) developed probabilistic models to predict future vehicle destinations with GPS data. Markov models are used to find the most likely, probabilistic next location based on a subset of previous locations. Liao et al. (2007) developed multimodal destination prediction methods. Their hierarchical modeling approach was able to increase accuracy with Bayesian inference while including different modes other than personal vehicles.

The previous research discussed thus far requires a map database for linking locations to roadways and inferring previous trips' routes to determine location. These types of models create a personalized prediction algorithm that bases the next trip prediction on previous trips made by a user. Some models have moved away from this personalized approach and instead use only the trajectory data presented from the current trip (Xue et al., 2013). This approach requires less data and less data post processing, but the drawback is lower accuracy and inability to increase accuracy through past 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 this application, the modelers have advanced knowledge of the intended vehicle destination. The algorithms showed promising results in route and time estimation, with improvements made possible with increased point frequency and accuracy.

The set of information that can be added to the vehicle trajectory is ever increasing. These include travel time, trip length, road conditions, driving habits, time of day, day of week, and velocity (Krumm and Horvitz, 2006, 2007; Horvitz and Krumm, 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 of similar type in the area. 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.” This paper increases the early stages of prediction to provide services on predicted destinations. Only the end of trips—i.e., when a user is close to their destination—are supported by their methodology.

Miyashita et al. (2008) built on 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 was not built on the previous paper. Alvarez-Garcia et al. (2010) proposed a destination prediction model that uses current location at the beginning of the trip. The support map of the algorithm is generated purely by the GPS points and is thus independent from a street map database. The modeling structure is most similar to the one used in this paper’s approach. The Alvarez-Garcia paper does not give a prediction at the start of trip; the prediction is provided 25% of the way through the trip (48.54% accuracy). As a comparison, the methodology in this paper will show an accuracy of 50.03% at the very start of the trip, providing a slight increase at an earlier starting point.

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 en-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.

The Markov model approach proposed for this research has been previously applied in other domains of transportation research. Yeon et al. (2008) developed a discrete time Markov chain to estimate travel time by discretizing links as congested or not congested. Li (2009) combines a Markov chain model and Bayesian analysis to produce a closed-form and computationally more efficient solution to the transit OD estimation problem. Lei et al. (2011) used a sub-trajectory synthesis that employs an offline Markov model to predict the probability of any given trajectory that is made online. Han and Sohn (2016) imputed activity sequences from travel smart-card data. The methodology employed land-use characteristics in a continuous hidden Markov model. Ma et al. (2017) estimated trip travel time distribution using a generalized Markov chain approach. Ulmer et al. (2017) proposed a new Markov Decision Process (MDP) for dynamic routing problems. They extended the traditional MDP for both stochastic and dynamic optimization problems. The new MDP is route-based, showing similar results to non-route-based models.

2.2. Trip purpose derivation

There is a wealth of research on trip purpose identification using GPS data. This work began with Wolf et al. (2001), when data loggers were used to replace or supplement electronic travel diaries. Since this initial work, several papers have aimed at increasing the accuracy of trip purpose derivation: Axhausen et al. (2003), Griffin and Huang (2005), McGowen (2006), Bohte and Matt (2009), Schüssler and Axhausen (2009), and Shen and Stopher (2013). Deng and Ji (2010) began to derive rules for trip purpose identification. Using a machine learning decision tree approach, the overall accuracy was 87.6% for 226 trips. Most recently, reports show that accuracy between 80 and 85% can be reached by employing a random-forest machine learning approach (Montini et al., 2014) on larger datasets consisting of trip records among 150 participants for an entire week.

The methodology from Lu et al. (2013) and Lu and Zhang (2015) is used as a baseline in the machine learning portion of this research due to the similarity in datasets, access to the model, and overall model accuracy. More information on trip purpose derivation is provided in Section 5. Using closest point of interest (POI) information, their models predicted the correct trip purpose for 80.6% of trips. Lu and Zhang (2014) further explored trip purpose estimation for urban travel by using NHTS add-on data. Their model showed an accuracy above 80% for home, work, school, and shopping trip types, but unsatisfactory results for social, other, and driving trip types.

The fields of trip purpose derivation and vehicle destination prediction have been gaining much interest and have seen several advances. Despite the gains that could be made in vehicle destination prediction algorithms with the use of trip purpose estimation, it has not yet been explored. This paper aims to start a new branch in vehicle destination prediction with the addition of trip purpose as a classifier to base predictions.

3. Data

3.1. GPS data

GPS data was collected as a 260-participant dataset, recording vehicle location every minute for 70 days. In total, the dataset consists of 36,000 trips. A representative sample was ensured by selecting participants from the full set of 800 applicants. Individuals provided their demographic information via an online survey and participants were selected based

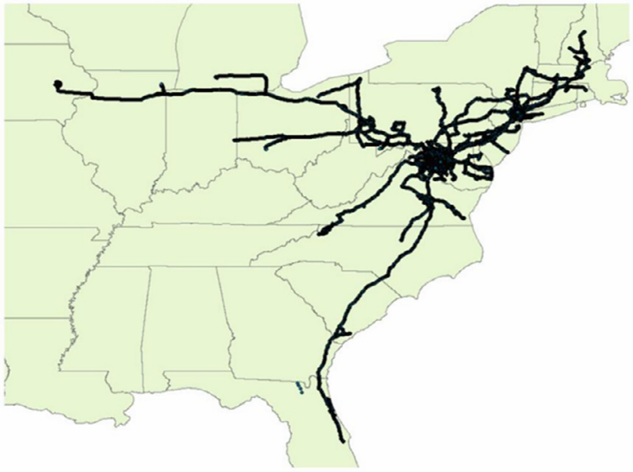

Fig. 1. GPS data representation.

on their response. A full report on data collection can be found in Krause (2012). The GPS data has gone through extensive cleaning and only error-free trip strings were used in this research. Fig. 1

The data spans a total of 22 states and long-distance trips are considered in the modeling. Due to long-distance travel destinations being difficult to predict, the accuracy of the model may be lower than those who use only short-distance trips.

3.2. Land use data

The point of interest data that was used to derive trip purpose for each trip was collected and aggregated from the open crowd-sourced data collection company, OpenStreetMap (http://www.openstreetmap.org/). The data has been converted from the original XML format into Geographic Information Systems (GIS) layer files for linking with the Global Positioning System (GPS) data. The image below shows the intricate POI data. In total, there are roughly 26,000,000 points of interest in Maryland, Washington, D.C., and Virginia (the locations where the majority of trips occurred in the dataset). Fig. 2

The POI data linked to the GPS data via a spatial join in the GIS environment. The distance threshold for POI to GPS linking is 300 m. Due to the computational time of linking POI and GPS points, only the DC/Maryland/Virginia region links 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 is likely minor.

3.3. Origin/ destination identification

It is important to explain how origins and destinations are derived and designated as the same location as another. First, all origins and destinations are defined as locations. A location is set when a vehicle is unmoving at the same latitude longitude point for more than three minutes. A 2 minute and 55 s stop to drop someone off, or to go through a fast food restaurant drive-thru would not be caught and considered a stop; however, if the threshold were lowered, then locations may be designated when idling at intersections or during traffic congestion. With each trip and origin/destination signified, the location is given an identifier so that similar locations can be grouped together. A location is clustered with others that are within 300 m of any other location of a user. The locations that match with others are given specific matching location identifier tags. A distance of 300 m was chosen due to previous papers (Ashbrook and Starner, 2003) using a similar distance threshold, but this value can be changed at the discretion of the stakeholder.

Of the 41,304 locations in the dataset, 4,832 locations were not within 300 m of any other location for the 70-day survey period. Using the methodology described in this paper, it is impossible to predict unique locations as the destination of trips.

4. Hierarchical Markov Model (HMM)

The model estimates the trip's destination then learns about the trip after it has begun. The next trip will have more information in four major categories, which 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). Fig. 3 shows a graphical breakdown of how the model works, followed by an explanation of each step.

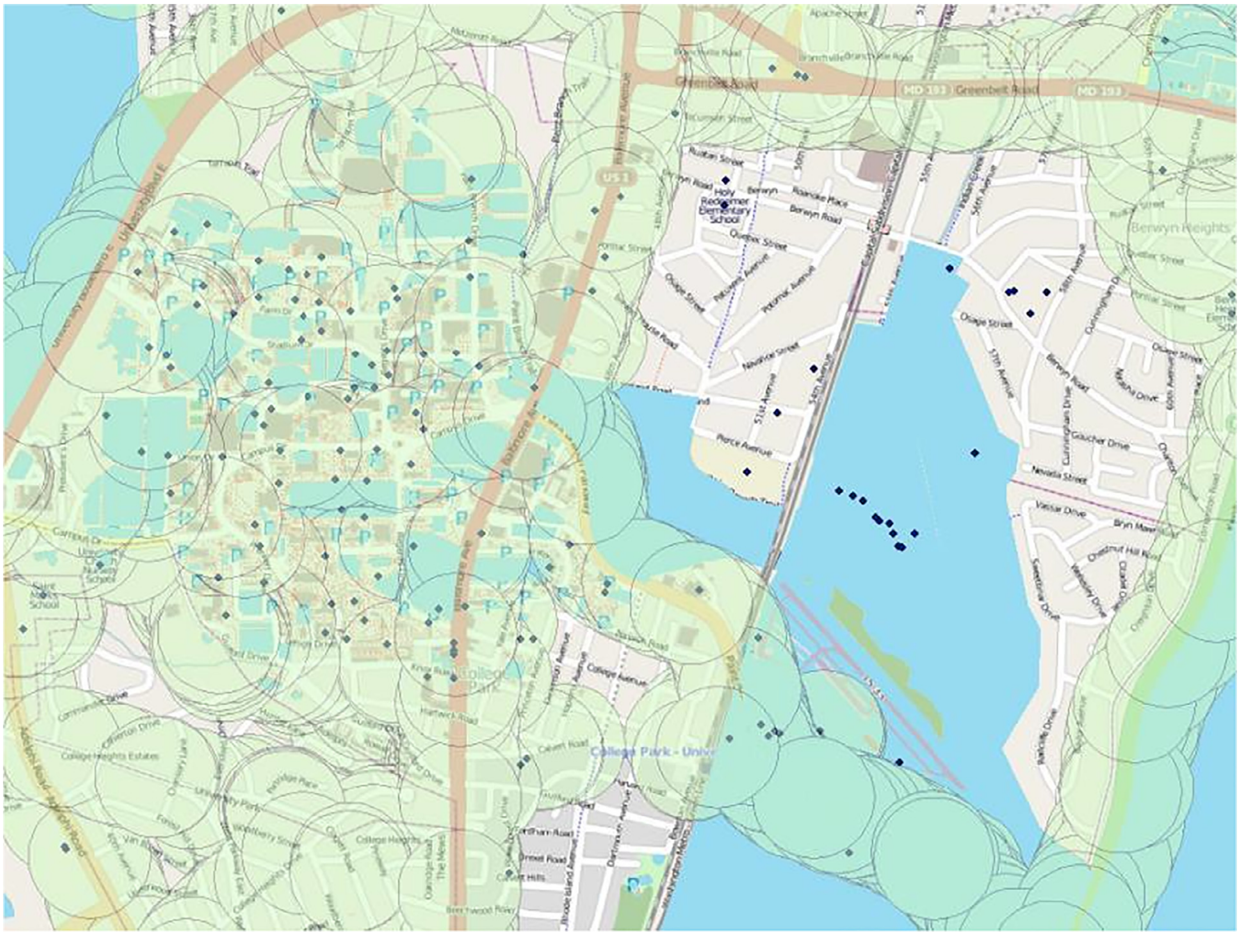

Fig. 2. GPS point buffers and POI spatial join around the University of Maryland.

The prediction model starts by first 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 began. 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 then 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 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 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 because no one destination has a higher likelihood of occurring than any other destination, the model moves on to the next tier.

At Tier 2, a search is done for all other previous trips that occurred at both the same hour of the day and the same origin as the current trip that it is estimating. If there is information in this array, then the most common destination of the same origin and time of day are made similar to the previous tier. There is a chance that the destination prediction that has just been made is the same as the previous correct destination. This means that the script is probably incorrect, and the algorithm goes on to Tier 3. The same steps are then taken to make an estimation on simply the time of day, and the check occurs again to see if it must move on to Tier 4. At Tier 4, the algorithm looks up the origin, and then selects the destination that is most common for a trip that originates from the selected origin ID. At Tier 5, the algorithm has run out of possible searches and simply selects the most abundant location that has not yet been used from the previous four tiers. Once the estimation is made, the script checks whether it is correct, loads the information into the appropriate arrays for future estimation, and then moves onto the next trip. 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.

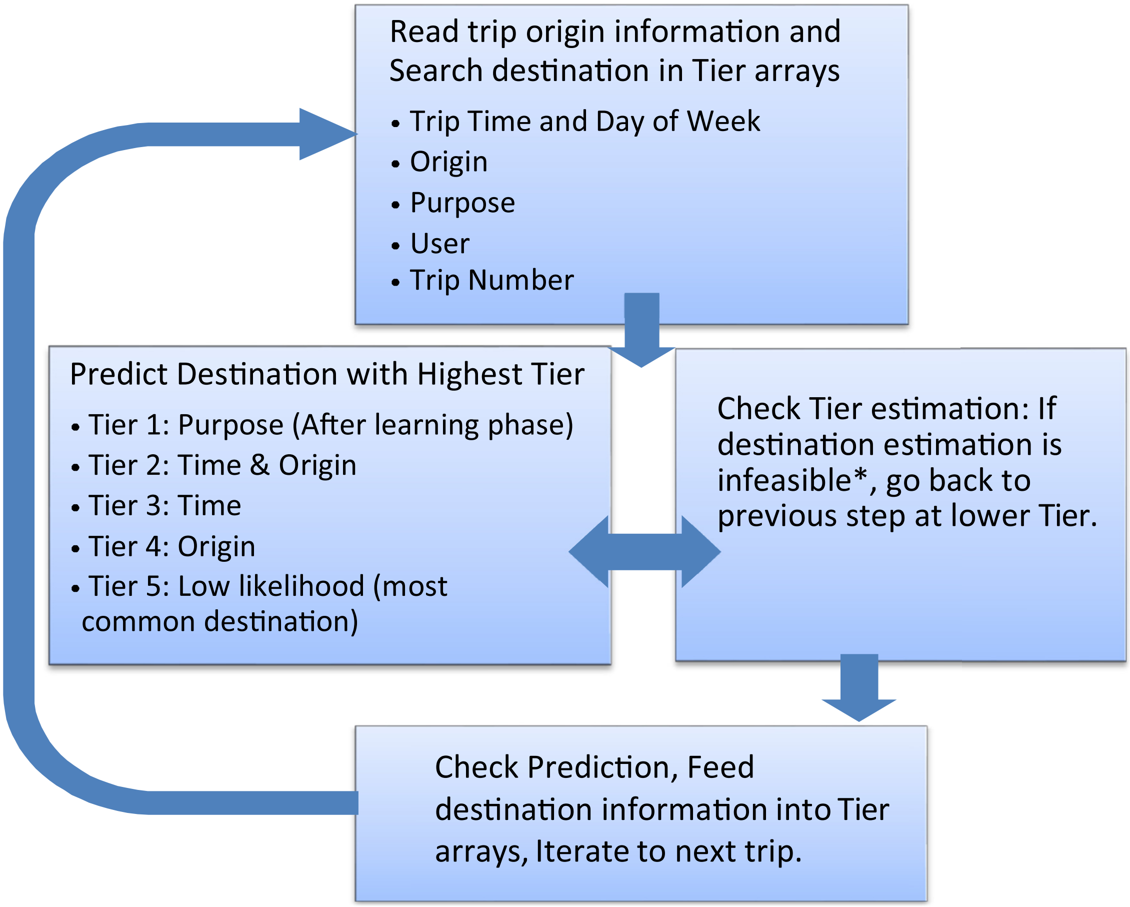
Read trip origin information and Search destination in Tier arrays
• Trip Time and Day of Week
• Origin
• Purpose
• User
• Trip Number
Predict Destination with Highest Tier
• Tier 1: Purpose (After learning phase)
• Tier 2: Time & Origin
• Tier 3: Time
• Tier 4: Origin
• Tier 5: Low likelihood (most common destination)
Check Tier estimation: If destination estimation is infeasible*, go back to previous step at lower Tier.
Check Prediction, Feed destination information into Tier arrays, Iterate to next trip.

Fig. 3. Graphical representation of prediction algorithm. *: infeasible selections occur when the existing trip database does not have a pre-existing trip of that type for that individual. Example: Tier 1 trip prediction would fail if that individual has not yet taken a "School" type trip when the current trip is estimated as "School". Tier 2 would fail if that individual has not previously taken a trip from "Home" at 11am, if the current trip is from "Home" and occurring at 11am.

4.1. Model formulation

Each equation number links directly to the accompanying tier number in Fig. 3.
Trip Purpose:

$$P^{p}\ (v_i = l|u_i = u,\, p_i = p_k) = \frac{\sum\{v_r|v_r \in V, u_i \in U, p_i = p_k\}}{\sum\{v_r, u_i|\ v_r \in V,\ \ u_i \in U, p_i = P\}} \tag{1}$$

Time and Origin:

$$P^{T\&O}\ (v_i = l|u_i = u,\ t_i = T, v_{i-1} = l_k) = \frac{\sum\{v_r|v_r \in V,\ \ u_i \in U,\ \ t_i = t_k, v_{r-1} =\ \ l_k\}}{\sum\{v_r, u_i|\ v_r \in V,\ \ u_i \in U,\ \ t_i = t_k,\ \ v_{r-1} =\ \ l_k\ \}} \tag{2}$$

Time:

$$P^{T}\ (v_i = l|u_i = u,\ t_i = T) = \frac{\sum\{v_r|v_r \in V,\ \ u_i \in U,\ \ t_i = t_k\}}{\sum\{v_r, u_i|\ v_r \in V,\ \ u_i \in U,\ \ t_i \in T\}} \tag{3}$$

Origin:

$$P^{O}\ (v_i = l|u_i = u,\ v_{i-1} = l_k) = \frac{\sum\{v_r|v_r \in V,\ \ u_i \in U,\ \ v_r = l, v_{r-1} =\ \ l_k\ \}}{\sum\{v_r, u_i|\ v_r \in V,\ \ u_i \in U,\ \ v_{r-1} =\ \ l_k\ \}} \tag{4}$$

Most Frequent Visited:

$$P^{MFV}\ (v_i = l|u_i = u,\ v_{i-1} = l_k) = \frac{\sum\{v_r|v_r \in V,\ \ u_i \in U,\ \ v_r = l\}}{\sum\{v_r, u_i|\ v_r \in V,\ \ u_i \in U\}} \tag{5}$$

Where:

$V = \{v_1, v_2, \ldots, v_n\}$ is the set of all visited locations
$U = \{u_1, u_2, \ldots, u_n\}$ is the set of all users
$T = \{t_1, t_2, \ldots, t_{24}\}$ is the set of all time intervals
$P = \{p_1, p_2, \ldots, p_7\}$ is the set of all purposes (home, work, other, driving, school, social, shopping)
l_k is the previous location
v_r is the next visited location
u_i is the current user
t_k is the previous time period
t_i is the current time period

5. 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 with 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. However, 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 probable location. By using demographic information of the user in the purpose model, the algorithm may rule out locations that the trajectory model would otherwise consider (e.g., a driver with no children would unlikely be headed to an elementary school despite their vehicle approaching it).

For each participant's first 15 trips, the trip characteristics were recorded into a separate database for machine learning. Once the first 15 trips are taken, a pre-established rule system was run to estimate the purpose for each of those trips. This rule-based system has shown to be 81% accurate in previous research (Lu et al., 2013). However, this trip purpose was based on end of trip location, which must be removed. Using the first 15 trips and a J48 machine-learning algorithm, a rule set was created that determined trip purpose using only start-of-trip information. This rule set was applied to all future trips. The process and data needed are shown in Fig. 4.

Due to the size of the rule set, the full list is not included in this paper. For access, please contact the authors. The approach derived a trip purpose estimation based on previous trips, which requires no land use or destination information from the current trip after the start of the car engine.

As an example, one 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 < 0.4 miles and
if time since last trip < 14 h 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 searches for the most common previously-visited location that has 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 HMM. This trip purpose module is added to the hierarchical Markov model as the 'top tier'. When this prediction estimation is available (after the first 15 trips), the purpose module is used to estimate destination.

6. Results

The results are given in two categories; with and without the trip purpose module. When the trip purpose module of the algorithm made a prediction, then that prediction was used. If, however, it did not have a prediction based on the user's purpose, then the baseline model was used. Fig. 5 compares the accuracy of the two models.

Improvements were 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 was seen. This

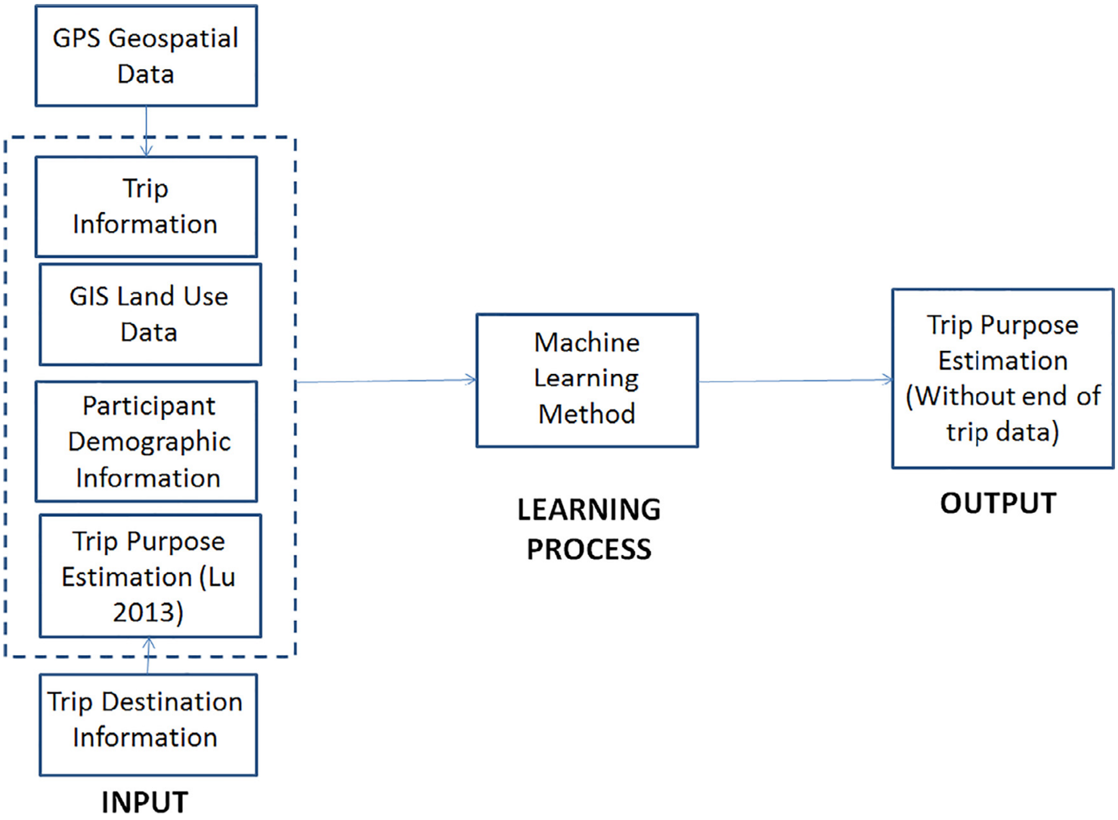
GPS Geospatial Data
Trip Information
GIS Land Use Data
Participant Demographic Information
Trip Purpose Estimation (Lu 2013)
Trip Destination Information
INPUT
Machine Learning Method
LEARNING PROCESS
Trip Purpose Estimation (Without end of trip data)
OUTPUT

Fig. 4. Machine learning for the trip purpose model using only start-of-trip information.

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 learning.

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, which is no longer helping overall accuracy. Zhao et al. (2018) estimated the probability that a driving pattern change occurs under multiple behavior dimensions. The Bayesian method described therein can successfully locate these behavior inflection points in travel patterns. Future research may consider these behavior pattern changes and alter the learning model to discount historical travel data before pattern changes. This may marginally increase overall model accuracy.

6.1. Prediction accuracy by trip purpose

The increased accuracy of the purpose model was made by classifying five trip purposes: work, social, school, driving, and shopping. “Home” and “other” showed either a slight decrease in accuracy or very marginal improvement. This can be explained by the land use characteristics that help to identify these zones, providing increased information about the area. The land use characteristics that identify these locations are: business unit, commercial, government/public or service, residential, restaurant, mixed use, institutional, industrial, leisure, recreation site, shops, and undeveloped. School-type trips doubled in accuracy due to matching similar trips that used the land use type government/public/service. Driving made a major improvement as well, however the trip type made up a very small percentage of overall trips and therefore did not make 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, and school locations. Even with this difficulty, there is about a 5% improvement over the HMM without trip purpose. Fig. 6

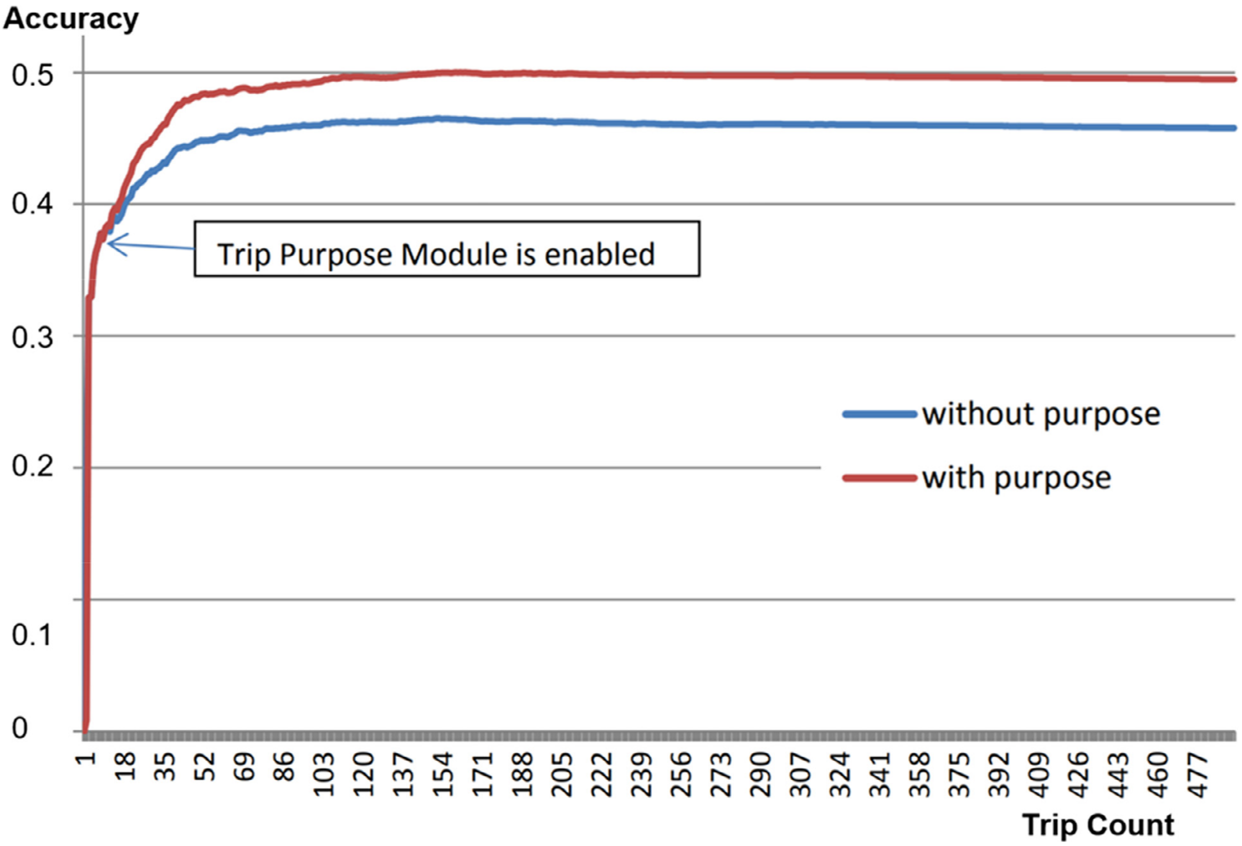

Fig. 5. Prediction algorithm accuracy: with and without trip purpose module.

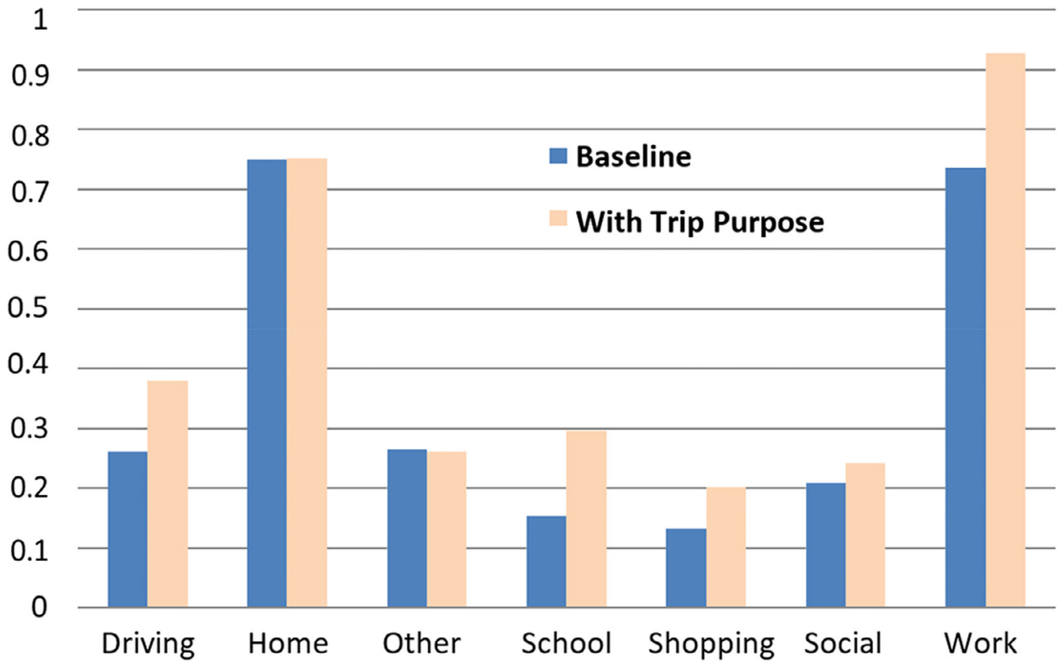

Fig. 6. Prediction accuracy by trip purpose.

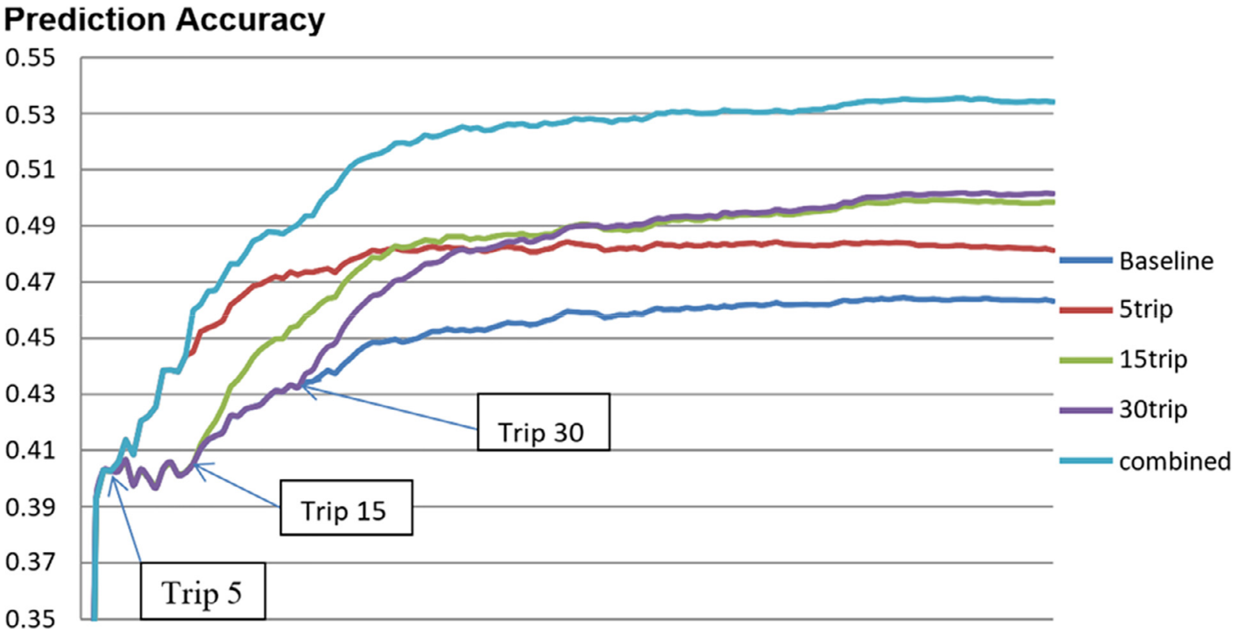

Fig. 7. Mode prediction accuracy by trip purpose learning period. *Note:* Starting at trip 3 increasing to trip 130. Trips 131–500 are removed for ease of viewing.

6.2. 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. Similar to 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 trips' accuracy in order to provide more accuracy of trips 31–500, or does a small learning set of 5 trips suffice in increasing the model accuracy after the learning set is enacted (without sacrificing accuracy for trips 6–30)? Fig. 7 shows the overall accuracy with similar graphs.

The 5-trip purpose module clearly has an accuracy advantage early 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 in order 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 three minutes, running a new learning set between trips may require significant computing power.

6.3. Prediction variation

The trip purpose model increases the prediction accuracy of the HMM by 7%, but it should be noted the nature of the predictions. Since this is a start-of-trip prediction model, the prediction may be altered as time passes and en-route algorithms can take over. It is important to get the prediction correct, but also, if it is not correct, give a feasible subset for possible locations. This was studied by looking at the prediction types in the HMM and trip purpose model.

The hierarchical Markov model estimates either home or work location over 90% of 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 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 that 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. Clearly, additional time to learn from

Table 1
Model Accuracy Compared to Trip Prediction Type.

Model	Percent of Destinations Estimated as Home or Work	Overall Model Accuracy	Accuracy After Learning Period
Tiered Time Origin (Baseline)	90.5%	45.7%	N/A
Trip Purpose Learning (30 Trip)	77.3%	50.0%	51.2%
Trip Purpose Learning (15 Trip)	82.0%	49.4%	50.1%
Trip Purpose Learning (5 Trip)	89.6%	47.1%	47.2%
Combined 5–15–30 Trip Purpose	62.4%	52.3%	52.7%

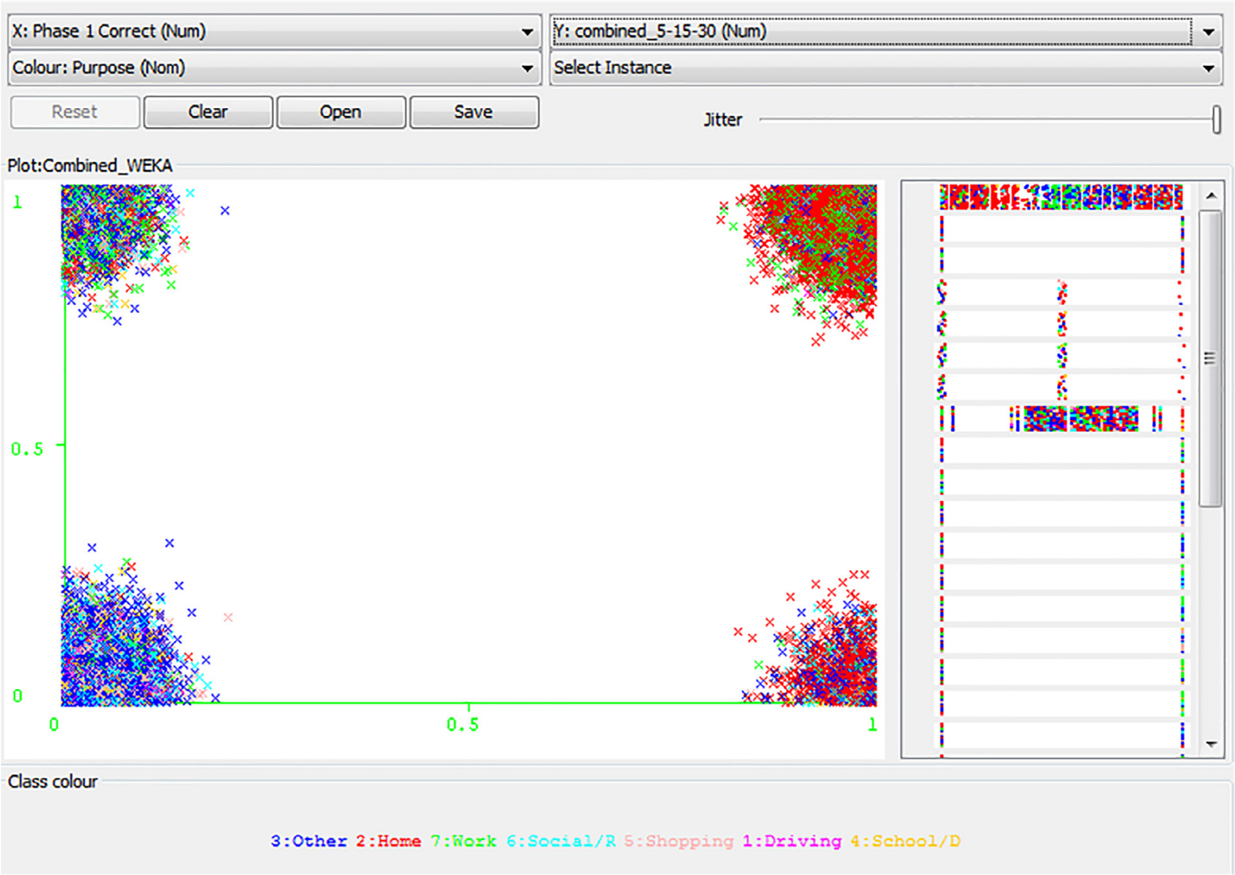
X: Phase 1 Correct (Num)
Y: combined_5-15-30 (Num)
Colour: Purpose (Nom)
Select Instance
Reset
Clear
Open
Save
Jitter
Plot:Combined_WEKA
1
0.5
0
0
0.5
1
Class colour
3:Other 2:Home 7:Work 6:Social/R 5:Shopping 1:Driving 4:School/D

Fig. 8. Accuracy comparison of baseline model (bottom right), trip purpose model (top left), both (top right), and neither (bottom left) by trip purpose.

user's trip purpose activity does improve the model despite only marginal increase in accuracy. When this model is applied to en-route destination prediction, a major increase in accuracy may be seen.

Table 1 shows that the 5-15-30 combined trip training model resorts to choosing non-home and work locations 37.6% of the time. Fig. 8 (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 difficulty with “other” trips.

The Trip purpose model both predicts a wider range of locations and tends to get those locations (shopping, social, driving, school, other) correct more often.

6.4. Comparison with state-of-the-art models

The literature gives the best start-of-trip model in Gao’s (2012) most-frequent hour-day model (MFHD). Their dataset included 3,373 locations and a 50-day training set. The HMM (baseline model) in this paper uses most frequent hour-day categorization and has an accuracy of 45.7% with the GPS data. Our dataset included 41,304 locations and no training period. The purpose model, with a combined 5-15-30 trip training period and 41,304 locations has an after-training accuracy of 52.7%, improving the best approach in the literature by 7 percentage point.

6.5. Future work and en-route prediction

Future work will focus on en-route prediction of destination. There have been several en-route destination prediction algorithms, particularly in computer sciences. The en-route models began in 2008 (Krumm 2008), where the GPS locations were set into a grid network and future destination was 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 highest 90% route completion model in the literature was done by Alvarez-Garcia et al. (2010) at roughly 79% accuracy. The model maximized 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 of going to a destination can be determined with fair accuracy. The model increases in accuracy considerably as the trip progresses. By including the work in this paper, the authors believe that an increase in accuracy could be found over the existing computer science-based research.

7. Conclusion

This is the first research to use trip purpose for predicting destination location. A literature review showed 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 was used to estimate an accurate start-of-trip destination. By incorporating demographic and land use data, trip purpose was derived, which improves the accuracy of this start-of-trip model. Overall accuracy of a hierarchical Markov model was improved by approximately 7% by the end of the survey period, but improvements can be seen as early as trip number 6, which showed a fast improvement with little drawback. The model was shown to be over 90% accurate at predicting work trip destinations. 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 en-route destination prediction algorithms.

References

Alvarez-Garcia, J.A., Ortega, J.A., Gonzalez-Abril, L., Velasco, F., 2010. Trip destination prediction based on past GPS log using a Hidden Markov Model. Expert Syst. Appl. 37 (12), 8166–8171.

Ashbrook, D., Starner, T., 2003. Using GPS to learn significant locations and predict movement across multiple users. Pers. Ubiquit. Comput. 7 (5), 275–286.

Axhausen, K.W., Schoenfelder, S., Wolf, J., Oliveira, M., Samaga, U., 2003. 80 weeks of GPS-traces: approaches to enriching the trip information. 83rd Transportation Research Board Meeting. ETH Eidgenössische Technische Hochschule Zürich, Institut für Verkehrsplanung und Transportsysteme.

Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: a large-scale application in the Netherlands. Transport. Res. Part C: Emerg. Technol. 17 (3), 285–297.

Deng, Z., Ji, M., 2010. 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.

Gao, H., 2012. In: *Mobile Location Prediction in Spatio-Temporal Contex*t, 2, pp. 1–4.

Griffin, T., Huang, Y., 2005. A decision tree classification model to automate trip purpose derivation. In: Proceedings of the 18th International Conference on Computer Applications in Industry and Engineering 2005. Honolulu. CAINE.

Han, G., Sohn, K., 2016. Activity imputation for trip-chains elicited from smart-card data using a continuous hidden Markov model. Transport. Res. Part B: Methodol. 83, 121–135.

Horvitz, E., Krumm, J., 2012. Some help on the way: opportunistic routing under uncertainty. In: Proceedings of the 2012 ACM Conference on Ubiquitous Computing. ACM, pp. 371–380.

Karbassi, A., Barth, M., 2003. Vehicle route prediction and time of arrival estimation techniques for improved transportation system management. In: Proceedings of the Intelligent Vehicles Symposium. IEEE.

Krause, C., 2012. A Positive Model of Route Choice Behavior and Value of Time Calculation Using Longitudinal GPS Survey Data. University of Maryland, Digital Repository at the University of Maryland, College Park.

Krumm, J., 2008. A markov model for driver turn prediction. SAE SP 2193 (1).

Krumm, J., Horvitz, E., 2006. Predestination: inferring destinations from partial trajectories. In: Proceedings of UbiComp. ACM, pp. 243–260.

Krumm, J., Horvitz, E., 2007. Predestination: where do you want to go today. IEEE Comput. 105–107.

Lei, P.R., Shen, T.J., Peng, W.C., Su., I.J., 2011. Exploring spatial-temporal trajectory model for location prediction. In: IEEE 12th International Conference on Mobile Data Management, pp. 58–67. doi:10.1109/MDM.2011.61.

Li, B., 2009. Markov models for Bayesian analysis about transit route origin–destination matrices. Transport. Res. Part B: Methodol. 43, 301–310.

Liao, L., Fox, D., Kautz, H., 2007. Learning and inferring transportation routines. Artif. Intell. 171 (5-6), 311–331.

Lu, Y., Zhu, S., Zhang, L., 2013. Imputing trip purpose based on GPS travel survey data and machine learning methods. Transportation Research Board 92nd Annual Meeting. Transportation Research Board.

Lu, Y., Zhang, L., 2014. 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. Transportation Research Board.

Lu, Y., Zhang, L., 2015. Imputing trip purposes for long-distance travel. Transport. 42 (4), 581–595.

Ma, Z., Koutsopoulos, H.N., Ferreira, L., Mesbah, M., 2017. Estimation of trip travel time distribution using a generalized Markov chain approach. Transport. Res. Part C: Emerg. Technol. 74, 1–21.

McGowen, P.T., 2006. Ph.D. dissertation. University of California, Irvine.

Miyashita, K., Terada, T., Nishio, S., 2008. A map matching algorithm for car navigation systems that predict user destination. 22nd International Conference on Advanced Information Networking and Applications Workshops/Symposia.

Montini, L., Rieser-Schüssler, N., Horni, A., Axhausen, K.W., 2014. Trip purpose identification from GPS Tracks. Transportation Research Board 93rd Annual Meeting. Transportation Research Board.

Schüssler, N., Axhausen, K.W., 2009. Processing GPS raw data without additional information. Transport. Res. Rec. J. Transport. Res. Board 2105 (1), 28–36.

Shen, L., Stopher, P.R., 2013. A process for trip purpose imputation from Global Positioning System data. Transport. Res. Part C: Emerg. Technol. 36, 261–267.

Terada, T., Miyamae, M., Kishino, Y., Tanaka, K., Nishio, S., Nakagawa, T., Yamaguchi, Y., 2006. Design of a car navigation system that predicts user destination. 7th International Conference on Mobile Data Management. IEEE.

TTI, 2015. Urban Mobility Report. Texas A&M Transportation Institute Report.

Ulmer, M.W., Goodson, J.C., Mattfeld, D., Thomas, B., 2017. Route-Based Markov Decision Processes for Dynamic Vehicle Routing Problems. Technical University Braunschweig, Braunschweig Working Paper.

Wolf, J., Guensler, R., Bachman, W., 2001. Elimination of the travel diary: experiment to derive trip purpose from global positioning system travel data. Transport. Res. Rec. 1768, 125–134.

Xue, A.Y., Zhang, R., Zheng, Y., Xie, X., Huang, J., Xu, Z., 2013. Destination prediction by sub-trajectory synthesis and privacy protection against such prediction. IEEE 29th International Conference on Data Engineering (ICDE). IEEE.

Yeon, J., Elefteriadou, L., Lawphongpanich, S., 2008. Travel time estimation on a freeway using Discrete Time Markov Chains. Transport. Research Part B: Methodol. 42, 325–338.

Zhao, Z., Koutsopoulos, H., Zhao, J., 2018. Detecting pattern changes in individual travel behavior: a Bayesian approach. Transport. Res. Part B: Methodol. 112, 73–88.