Automated Transportation Mode Detection Using Smart Phone Applications via Machine Learning: Case Study Mega City of Tehran

Zahra Ansari Lari
Graduate Student
Department of Civil and Environment Engineering
Amirkabir University of Technology
Hafez Avenue, Tehran, Iran
Tel: +98 (21) 88336480
Email: zansarilari@aut.ac.ir

Amir Golroo
Assistant Professor
Department of Civil and Environment Engineering
Amirkabir University of Technology
Hafez Avenue, Tehran, Iran
Tel: +98 (21) 64543010
Fax: +98 (21) 66414213
Email: agolroo@aut.ac.ir

Paper submitted for:
Presentation at 94th Annual Transportation Research Board Meeting, January 2015 and
Publication in the Transportation Research Record: Journal of Transportation Research Board

Word Count: 6147 + 1,750 = 7,897 words
(Text) (4 Tables + 3 Figures)
ABSTRACT

Through the past few decades, travel behaviors have become more complicated especially in mega cities such as Tehran, the capital of Iran. Decision makers require more accurate and comprehensive information to plan for city transportation. As opposed to traditional paper-based and telephone-based surveys, a new efficient and effective data collection method has been recently applied using information technology such as the GPS-based data collection method which can track passengers’ trips. Having utilized this new method, the main aim of this study is to analyze the collected data in order to distinguish transportation modes used by passengers using a novel machine learning method called random forest. This model not only classifies transportation modes i.e., car, bus, and walking at a high accuracy of almost 96%, but also determines the most influential attributes in the process of classification based on two importance indices: mean decrease accuracy and Gini index. Results show that instant speed and accuracy of GPS track are the most influential attributes in the transportation modes classification. Transportation planners benefit a lot from this accurate and comprehensive travel behaviors data (used modes) for policy making.

Keywords: GPS-Based Data collection, Mode Recognition, Random Forest, Feature Selection.
INTRODUCTION

Over the past few decades, there is a growing attention toward transportation modes share, the sequential of used modes in a trip, origin-destination travelers’ information, and travelers’ characteristics, amongst different groups of policymakers. Urban planners and city officials, on one hand, have found this issue as a useful tool to conduct their programs in a way of providing the appropriate and adequate transport services for future travel demands. Health officials, on the other hand, use this information to take feasible and effective actions according to the global problem of obesity among people.

One of the essential information required for transportation planning is to understand passengers’ travel behavior in cities. In other words, understanding the modes used, origins, and destinations of daily trips could lead to comprehensive transportation planning. This planning needs a database with a large number of observations at a high quality. One of the major methods of acquiring data is to conduct a survey. Travel surveys have been undertaken through different methods by various researchers to collect required data (1-3). The first method of collecting travel data was paper-based questionnaires followed up computer-assisted telephone interview (CATI) which reached the survey participants who submitted the paper survey. This method was dependent on respondents’ memory and contained indispensable errors. Also, in the best condition a respondent was able to only describe his/her trips for one or two days without even an acceptable accuracy.

The recent progress in information technology gives researchers an excellent opportunity to gather the travel information more efficiently and effectively i.e., not only time and budget are saved, but significantly more data is collected. One of the useful tools which receives attention by transportation planners is GPS-based technology data collection methods. GPS traces are used widely for collecting activity-based travel diary data during recent years (4-7).

This study focuses on analyzing data through presenting a novel machine learning modeling formulation to use raw GPS data for mode recognition. Machine learning methods have performed well in both complex regression and classification problems. The formulation used in this study is a random forest which was proposed for the first time in 2001 by Breiman (8). Machine learning methods have been well performed in classification problems. Random forest is an ensemble classifier which is combination of N tree predictors (8). It is generally found that ensemble methods are more accurate than any of the individual classifiers creating the ensemble (9). An ensemble could be aggregate weighted or unweighted votes from each individual classifier to classify an unknown observation. In addition, feature selection aspect is examined based on a random forest model in this study. Three modes are examined in this study: car, bus, and walking. Subway and commuted rails trips are not being included in the model, because almost all the lanes of this mode in Tehran are underground so GPS records just showed the origin and destination stations which cannot result in mode recognition modeling process. One way to classify these trips is to utilize GIS maps including origin and destination stations to determine the associated mode and path.

LITERATURE REVIEW AND BACKGROUND

Extensive studies have conducted on travel behavior using GPS-based data owing to a high level precision, reliability, efficiency, effectiveness, time-saving, and capability of providing a large potential sample size (4-7 and 10-12). This is mainly because of significant advancements in information technology.

Vehicle-based GPS data collection was the first devices used to record GPS traces (13 and 14). This method had many limitations e.g., not covering other frequently used modes such as walking or public transport (15). To overcome this problem, passengers were equipped with a GPS device to collect data (4 and 16-19). However, the investment for this method was too
high (15). It also was a heavy device to be carried by a passenger. Besides, it was a chance that
a passenger forgets to take the device with him/ her during daily trips (15).

By the advent of smart phones, the GPS-based data collection methods enormously
shifted to a more convenient and comprehensive way. Although, it overcomes some problems
of previous methods (20 and 21), it has its own limitations at the beginning. For instance, the
first smart phone applications were very battery consuming and they were just capable to record
data about 5 to 6 hours a day (20). Battery life is still a big issue impacting the more widespread
use of smartphones in GPS data collection. Also, these first generation applications did not
allow passengers to use the smart phones while it was capturing data (20). This issue might
lead to some changes in him/ her plans which could provide not realistic data base and
discourage people to take part in the survey. Afterwards, some researches (15 and 17) have
been conducted to optimize the process of smart phone-based data collection using GPS.
Despite these limitations, this method is still widely used and has several benefits such as being
able to use a consistent data collection approach across the sample and not allowing participant
smartphone biases interfere with data availability.

In terms of accuracy and prediction power, numerous machine learning methods have
been applied. Machine learning algorithms are well known in mining data and finding
underlying patterns across attributes in order to gain the ability of modeling and predicting.
Data mining is the process of data collection, data preparation, data exploration, data analysis,
and final inferences. So, the main task after data collection is how to analyze the data in a
general term. Several methodologies have been developed to process GPS traces for mode
recognition (10-12 and 22-23). The methodologies are divided in to two main categories:
procedural and machine learning methods (10). In procedural methods, the mode identification
is based on logical assumptions about passengers’ travel behavior such as using more public
transportation if a passenger is close to the stations (24). Or, according to passengers’ safety
concerns, they might less likely travel after mid-night (25) especially walking. In 2008, Stopher
et al. (26) presented a framework based on speed, acceleration, and some additional hierarchical
assumptions to distinguish between different modes. For example, availability of a bicycle at
home and low speed, and acceleration rate are hints to infer that the passenger used a bicycle
to complete his/her trip. This method overlooks the chance of presence of same attributes in
different modes such as velocity distribution and also needs more than just raw GPS (Global
Positioning System) data to assign a proper mode to a trip. Similarly, Chung and Shalaby in
2005 (24), and Liao, Fox, and Kautz in 2007 (25) investigated rule-based methods especially
with regards to environmental and situational factors for mode recognition.

Machine learning methods have been applied widely in this field of study because of
their extensive ability to analyze data through complex approaches. Machine learning methods
have the opportunity to find underlying rules in recondite data and utilize them to forecast new
observations with a high accuracy.

Other studies related to mode recognition have focused on using various available sensors and
methods to investigate the most conforming pattern for mode (12, 27-30). Each sensor has its’
own property that could lead to an acceptable outcomes. For example, GIS (Geographic
Information System) data are very useful information especially for some modes such as buses
or subways. Moreover, using GSM (Global System for Mobile) data need dense GSM cell
towers to have reliable data which they are not always available (12). Table 1 shows the list of
latest studies in which various machine learning methods have been used. This table also
expresses methodology accuracy, different modes detected, and applied sensors.
Table 1: Summary of Studies Related to Sensor-Based Data Collection

<table>
<thead>
<tr>
<th>First Author</th>
<th>Year</th>
<th>Methodology</th>
<th>Accuracy (%)</th>
<th>Modes</th>
<th>Sensors</th>
</tr>
</thead>
<tbody>
<tr>
<td>Feng (22)</td>
<td>2013</td>
<td>Bayesian Belief Network</td>
<td>96</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Feng (22)</td>
<td>2013</td>
<td>Bayesian Belief Network</td>
<td>81</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Feng (22)</td>
<td>2013</td>
<td>Bayesian Belief Network</td>
<td>96</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Stenneth (23)</td>
<td>2011</td>
<td>Bayesian Network</td>
<td>74.9</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Stenneth (23)</td>
<td>2011</td>
<td>Decision Tree</td>
<td>66.9</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Stenneth (23)</td>
<td>2011</td>
<td>Random Forest</td>
<td>75.4</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Stenneth (23)</td>
<td>2011</td>
<td>Naïve Bayesian</td>
<td>71.8</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Stenneth (23)</td>
<td>2011</td>
<td>Multilayer Perceptron</td>
<td>59.1</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Stenneth (23)</td>
<td>2011</td>
<td>Bayesian Network</td>
<td>92.5</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Stenneth (23)</td>
<td>2011</td>
<td>Decision Tree</td>
<td>92.2</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Stenneth (23)</td>
<td>2011</td>
<td>Random Forest</td>
<td>93.7</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Stenneth (23)</td>
<td>2011</td>
<td>Naïve Bayesian</td>
<td>91.6</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Stenneth (23)</td>
<td>2011</td>
<td>Multilayer Perceptron</td>
<td>83.3</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Manzoni (27)</td>
<td>2011</td>
<td>Decision Tree</td>
<td>82.14</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Xu (32)</td>
<td>2010</td>
<td>Fuzzy Logic</td>
<td>93.8</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Reddy (12)</td>
<td>2010</td>
<td>Decision Tree followed by first-order discrete Hidden Markov Model</td>
<td>93.6</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Zheng (33)</td>
<td>2008</td>
<td>Support Vector Machine</td>
<td>51.7</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Zheng (33)</td>
<td>2008</td>
<td>Decision Tree</td>
<td>72.1</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Zheng (33)</td>
<td>2008</td>
<td>Bayesian Net</td>
<td>57.7</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Zheng (33)</td>
<td>2008</td>
<td>Conditional Random Field</td>
<td>61.7</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Gonzalez (34)</td>
<td>2008</td>
<td>Neural Network</td>
<td>91.23</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
<tr>
<td>Mun (35)</td>
<td>2008</td>
<td>Decision Tree</td>
<td>83</td>
<td>Car: x, Walk: x, Bike: x, Bus: x, Train Subway: x, GPS: x, GIS: x, GSM: x, Wi-Fi: x, Accelerometer: x</td>
<td></td>
</tr>
</tbody>
</table>

Table 1 shows the variety of methods used in the field of mode recognition. Most of the researchers have examined at least three different modes of walking, car, and bus which are more frequently used. Different levels of accuracy have been achieved which the highest one by using only a raw GPS data is 93.8%. The level of accuracy depends on the associated applied algorithm, the quality and the size of the data base, the defined attributes, and assumptions made in each methodology.

Another aspect which has not received enough attention to date is the concept of variable importance. Identifying influential variables assists to investigate more underlying factors which leads to interpret data comprehensively and to forecast new observation with higher accuracy.

Bolbol et al. (10) investigated important variables with analysis of variance (ANOVA) test to select the best features which separately describe each mode. They examined the role of four main attributes for each mode: speed, acceleration, distance, and heading differences. They stated that acceleration and heading changes are the most important attributes in the modes of...
car and train, respectively. Also, they generally declared that speed and acceleration are the best attributes for distinguishing between transportation modes.

Stenneth et al. (23) applied the algorithms of Chi squared and information gain. They concluded that the first ranking attributes were: average speed, average rail line closeness, average bus closeness, and average acceleration. The researchers used GIS information in addition to GPS data. The most important variables according to GPS were average speed and average acceleration.

**OBJECTIVE AND SCOPE**

This study aims to introduce a novel machine learning method called random forest to not only provide higher accuracy in mode classification but also to introduce two common indexes for feature selection based on a random forest model: a mean decrease index, and a Gini index which specifically determine important attributes in each mode separately. This study only uses data collected using smart phones via GPS and accelerometer sensors. It also focuses on detecting three transportation modes including car, bus, and walking.

**METHODOLOGY**

After a detailed literature review, the framework was created for this study. The first step was to collect travel data. For this aim, a group of participants were trained to record their travel data (GPS tracks) via an appropriate application. Then, data pre-processing was carried out leading to define attributes for each GPS track based on collected data such as acceleration and delta accuracy. After that, different machine learning methods were examined to accomplish the model specification phase in order to find an appropriate model resulting in mode classification and feature selection at a high level of performance. Finally, the accuracy of mode classification and attribute effectiveness for each mode were discussed.

In this study, a random forest formulation is applied to investigate underlying factors in a GPS data to not only classify each mode at a high accuracy but also determine influential attributes in describing each class.

**Random Forest Technique Review**

Random forest consists of N tree classifiers. Creating each tree required two types of information: a subset of training data and the related attributes. The procedure in building a tree is to split the attributes in each node and continue to a next level until a final leaf consists of an ultimate answer generated. In each node the best attribute is selected for decision tree induction based on the entropy formula (Eq. 1) (36).

\[
\text{Entropy} (S) = \sum_{i=1}^{c} -p_i \log_2 p_i
\]

Where S denotes the entropy for each attribute, c is the number of different values in each attribute, and \( p_i \) is the proportion of S belonging to class i.

The novelty of ensemble methods is that they aggregate number of individual classifiers outcomes and report the most popular result as a final vote which is the category of the input variables (37). This approach is well performed in a random forest model based on randomization which makes the model more powerful for classification (38). The fundamental of random forest is based on two randomization approach: bagging and random attribute selections (38).

In the first step, for developing each tree in a forest a subset of data is randomly drawn with replacement from the original data to be used as a training data which is called bootstrapping. The portion of the data which is not selected is called out-of-bag data and used as a testing set. Similarly, the number of attributes is extracted randomly with replacement from all the existing ones. Then, a tree is generated using these training set and attribute selection.
Also, each tree is evaluated by its testing data. The number of misclassified observations is reported as the out-of-bag error rate and is used to evaluate variable importance in the final result. This type of testing provides fair estimation of accuracy of the model compared to the usual test approaches (39). Next, this procedure is undertaken to building N different trees in a forest.

In addition to the excellent ability of a random forest in classification and prediction, the model reports the measurement of influential attributes based on permutation process in the original data (8). Subsequently, to compute the effectiveness of each attribute in each class separately and measure the importance level of attributes in a whole data, the quantity of associated attribute changes randomly to incorrect values in the data. Then, the model is developed again using the permuted data set (8). After that, the assigned class to each observation is compared with the correct one. If the number of error in classification of the data is considerable, it is concluded that this attribute has a major effect in the process of modeling and predicting (8). The average importance of each attribute is calculated both in each class and for a whole data called mean decrease accuracy index. Variable importance of \( X_j \) is calculated using Equation 2:

\[
VI(X_j) = \frac{1}{\text{n} \text{tree}} \sum (\text{errorOOB}_f - \text{errorOOB}_{f_j})
\]  

Where \( VI(X_j) \) is the measurement of importance of attribute \( X_j \), the sigma is over all trees \( f \) of a forest and \( \text{n} \text{tree} \) is the number of built trees in a forest. \( \text{errorOOB}_f \) is the error rate before permutation and \( \text{errorOOB}_{f_j} \) is the error rate after the permutation of variable \( X_j \).

Another outcome of a random forest model is called mean Gini index (8). According to related studies (41-45), this measure recommended when all attributes are either numerical or categorical because it is sensitive to the kinds of attributes. This index is also an attribute selection measure similar to the mean decrease accuracy index. However, its amount calculated the impurity of an attribute regarding to the classes (8). In this procedure, at each node, the reduction in Gini impurity is reported for each variable which is used for splitting (40). Gini impurity \( \Delta GI(t) \) is calculated using Equation 3 (40):

\[
\Delta GI(t) = P_l GI(t) - P_l GI(t_L) - P_R GI(t_R)
\]  

Where, \( GI(t_L) \) is a Gini index on the left side of the node, \( GI(t_R) \) is a Gini index on the right side of the node, \( P_l \) is the number of observations before the split, \( P_l \) is the number of observations on the left side after the split, and \( P_R \) is the number of observations on the right side after the split and \( GI(t) \) is the Gini index defined by Equation 4:

\[
GI(t) = 1 - \sum_k p(k|t)^2
\]  

Where \( p(k|t) \) is the rate at which class \( k \) is distinguished correctly at node \( t \). The average of all reductions in the Gini index impurity used to calculate the Gini importance measure.

It should be noted that the domain of feature selection and variable importance measurement is a wide subject in machine learning studies (46-48). The methods are sensitive to the number of attributes and observations which is out of scope of this paper (49 and 50).

**DATA COLLECTION**

Data collection was conducted in December 2013 over 2 weeks in Tehran, the capital of Iran. GPS traces were collected using smart phones equipped with GPS through an appropriate application. Participants were employed to run the application for all day long, between 6 a.m. till 9 p.m., and collect their trip data with their consent. A sample data collected by a participant from home to work in the morning and back from work to home in the afternoon is plotted on the map as shown in Figure 1. This sample travel tracking data encompasses all three desired modes including walking, bus, and car. The participants are 25 males and ten females. All of
them are students between 19 and 25 years of age. Prior to starting data collection, participants were trained. In the training session, the objective of the study, method of working with the smart phone application, procedure of filling complement data form for validation (data labeling) were discussed in details. In order to label data with correct classes, participants were asked to report the used modes in complement data form and attach to each GPS file. Finally, more than 245,418 lines of GPS traces were collected that contains three modes: Car, bus, and walking. The segments, time, date, instant speed, accuracy, bearing, altitude, latitude, and longitude were recorded per GPS track in each trip. Table 2 provides a descriptive statistics for each attribute in each class. As shown in Table 2, the bus average speed is larger than that of cars because especial separated lanes are dedicated to a large number of buses in Tehran called Bus Rapid Transit (BRT) moving without traffic interference. However, the car maximum speed is larger than that of buses. The amount of standard deviation in delta bearing for walking is much more than delta bearing of the other modes, since in walking a person has more freedom to change his/her direction. Moreover, the average accuracy in walking is less than the accuracy of other two modes. The reason could be because of the walking low speed which results in capturing more data point in a unit of distance than the other modes. So, the deviation (in a unit of distance) of GPS traces from the main path for walking is more than that of others.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Average</th>
<th>Std.Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delta Bearing (deg)</td>
<td>-0.447 E-01</td>
<td>29.324</td>
<td>-357</td>
<td>358</td>
</tr>
<tr>
<td>Accuracy (m)</td>
<td>6.610</td>
<td>2.201</td>
<td>3.194</td>
<td>47.994</td>
</tr>
<tr>
<td>Speed (m/s)</td>
<td>23.346</td>
<td>10.535</td>
<td>0.251</td>
<td>31.618</td>
</tr>
<tr>
<td>Delta Speed (m/s)</td>
<td>-0.670 E-01</td>
<td>1.187</td>
<td>0</td>
<td>12.005</td>
</tr>
<tr>
<td>Acceleration (m/s²)</td>
<td>-0.257 E-01</td>
<td>0.535</td>
<td>-19.78</td>
<td>4.729</td>
</tr>
<tr>
<td>Delta Acceleration (m/s²)</td>
<td>-0.359 E-02</td>
<td>2.257</td>
<td>-12.244</td>
<td>126.960</td>
</tr>
<tr>
<td>Delta Bearing (deg)</td>
<td>-0.661 E-01</td>
<td>43.316</td>
<td>-359</td>
<td>359</td>
</tr>
<tr>
<td>Accuracy (m)</td>
<td>6.413</td>
<td>4.913</td>
<td>1.478</td>
<td>49.999</td>
</tr>
<tr>
<td>Speed (m/s)</td>
<td>10.376</td>
<td>6.181</td>
<td>0.250</td>
<td>35.237</td>
</tr>
<tr>
<td>Delta Speed (m/s)</td>
<td>-0.503 E-02</td>
<td>0.968</td>
<td>0</td>
<td>29.088</td>
</tr>
<tr>
<td>Acceleration (m/s²)</td>
<td>0.395</td>
<td>0.718</td>
<td>-13.258</td>
<td>28.690</td>
</tr>
<tr>
<td>Delta Acceleration (m/s²)</td>
<td>-0.107 E-01</td>
<td>0.820</td>
<td>-39.120</td>
<td>28.762</td>
</tr>
<tr>
<td>Delta Bearing (deg)</td>
<td>-0.537</td>
<td>75.086</td>
<td>-359</td>
<td>359</td>
</tr>
<tr>
<td>Accuracy (m)</td>
<td>12.805</td>
<td>10.398</td>
<td>1.706</td>
<td>49.999</td>
</tr>
<tr>
<td>Speed (m/s)</td>
<td>1.572</td>
<td>0.933</td>
<td>0.25</td>
<td>5.490</td>
</tr>
<tr>
<td>Delta Speed (m/s)</td>
<td>-0.514 E-01</td>
<td>1.085</td>
<td>0</td>
<td>4.954</td>
</tr>
<tr>
<td>Acceleration (m/s²)</td>
<td>0.596</td>
<td>0.589</td>
<td>-8.435</td>
<td>4.408</td>
</tr>
<tr>
<td>Delta Acceleration (m/s²)</td>
<td>0.251 E-01</td>
<td>0.881</td>
<td>-19.539</td>
<td>10.561</td>
</tr>
</tbody>
</table>

TABLE 2 Descriptive Statistics of Used Attributes in the Model
FIGURE 1 Plotted GPS Traces for a Sample Work Trip

MODELING
The first step in all modeling procedures is to clean the data set called data preparation. In the case of this study, especially, for the different ranges in accuracy of recording data and also because speed, which is an important attribute, is calculated based on the three geographical parameters: altitude, longitude, and latitude there are an indispensable errors. To overcome this problem, speed limit individually defined for different modes. For motorized modes this amount is assumed to be equal to 35 meter per second (125 kilometer per hour). For the mode of walking, the maximum speed reported 5 meter per second based on fuzzy member ship modeling studied by Zhang et al (11). However, in order to cover fast walking (based on local observations in this study), the maximum speed is assumed to be equal to 5.5 meter per second in the modelling process.

As mentioned before, a random forest model is applied in this study to present a new and advanced machine learning method in order to be used for mode classification from raw GPS data. New attributes are defined herein based on the raw data including acceleration (speed changes per second), acceleration differences per second, bearing changes per track, and accuracy changes per track. This new attributes are defined based on some logical rules for each mode. For example, acceleration and its’ changes per second cannot be very large for walking. Or, bearing changes in motorized modes is not the same as walking i.e. in walking a passenger has more freedom to suddenly change his/ her direction than a passenger in a motorized mode does. However, the attributes which are going to be used in a data mining should have same properties. For instance, the variables which make an observation very unique such as time, longitude, altitude, and latitude should not be used as an attribute in a model.

In developing a random forest model there are some parameters which should be rigorously selected in order to gain reliable and acceptable results (8). One of these parameters is the number of trees that should be built in a forest. The second parameter is the number of random variables used at each node. There has not paid enough attention to the best and efficient number of trees by researchers. Breiman claimed that the best number of trees is the number at which the out-of-bag error is almost stable (8). Moreover, the amount of this parameter is more relied on the number of observations, number of variables, and the desired accuracy which is unique in each problem (51). In this study, four different forests examined. Forest with 120, 140, 160, and 180 trees, and the forest with 160 trees was selected as a final model. The error rates for four abovementioned modeling were 4.18%, 4.19%, 4.16%, and 4.17%, respectively, (as the amount of the splitting attributes at each node was equal to three). Since almost the four applied models has equal error rate, the amount with the minimum error was selected.

The next parameter is the number of attribute which should be selected at each node for splitting. In classification problems, it is recommended to use the square root of numbers of attributes for this parameter (51). However, both numbers of second and third parameters are modeled to find the best fit. The error rates when two attributes used at each node are 4.46%,
4.43%, 4.43%, and 4.45% for 120, 140, 160, and 180 trees in a forest, respectively. Therefore, the smallest out-of-bag error is preferred and it was fixed to 3 in the final model.

RESULTS AND DISCUSSION
This section devotes to the results from the applied random forest model. As noted before, the main findings are divided into two parts: the accuracy of mode detection and the feature selection for each mode presented in the following section.

Mode Classification
The first outcome a random forest model is a confusion matrix (Table 3). In this matrix, the number of misclassified observations from out-of-bag data is reported across the forest in each class. Breiman (8) stated that in a random forest model it is not necessary to test the model on a separate test subset, instead the out-of-bag error rate could be reported. This amount is equal to 4.16% in this study. In other word, the model accuracy is about 95.84%. As shown in Table 3, the accuracies in bus, car, and walking separately are 87.93%, 97.68%, and 90.33%, respectively. As discussed before, the accuracies depends on the applied algorithm, quality and quantity of the sample size in each class, and the defined attributes (52). Since, the number of observations in mode of bus was less than the other classes, the process of learning performs weaker than the other mode which leads to the smaller accuracy rate. Overall, the average accuracy rate over all modes is higher compared to previous studies. The highest achieved accuracy in the recent studies was 93.8 for modes of walking, bike, and bus through fuzzy logic method conducted by Xu et al. (32) (without considering car as a mode).

Table 3 Confusion Matrix

<table>
<thead>
<tr>
<th>Mode</th>
<th>Bus</th>
<th>Car</th>
<th>Walk</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus</td>
<td>2967</td>
<td>367</td>
<td>40</td>
<td>87.93</td>
</tr>
<tr>
<td>Car</td>
<td>12</td>
<td>37176</td>
<td>871</td>
<td>97.68</td>
</tr>
<tr>
<td>Walk</td>
<td>3</td>
<td>757</td>
<td>7100</td>
<td>90.33</td>
</tr>
</tbody>
</table>

Table 4 Test Results

<table>
<thead>
<tr>
<th>Mode</th>
<th>Bus</th>
<th>Car</th>
<th>Walk</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bus</td>
<td>870</td>
<td>124</td>
<td>16</td>
<td>86.14</td>
</tr>
<tr>
<td>Car</td>
<td>6</td>
<td>11200</td>
<td>295</td>
<td>97.38</td>
</tr>
<tr>
<td>Walk</td>
<td>1</td>
<td>210</td>
<td>2071</td>
<td>90.79</td>
</tr>
</tbody>
</table>

Attribute Effectiveness
As discussed before, one of the important results of a random forest model is the measure of attribute effectiveness in describing each class in the both process of modeling and predicting. The calculated measures are sorted in a decent order for each class separately and for the whole
data. Figure 2 depicts the mean decrease accuracy index. The attribute that have the largest quantity of the index is the most influential factor in describing the data and in the process of prediction.

![Image of Figure 2](image_url)

**FIGURE 2 Results of Mean Decrease in Accuracy (a) All modes (b) Walking (c) Car (d) Bus**

As expected, in consistent with last studies (10 and 23), speed has the most important role in each class. The second effective factor is the accuracy of tacking GPS data. This attribute has not received enough attention in recent studies (10 and 23). Figure 1, also, illustrates the comparative relationships between attributes in each class. The effect of speed is much higher than other attributes in using bus. On the other hand, the next attribute (accuracy of tracking GPS) in other modes has a relatively large impact on describing the data.

Another attribute importance measurement which is mentioned before is Gini index. Similar to mean decrease accuracy, the attribute which has the highest amount of the Gini index is reported in Figure 3.
Kawakubo et al. (40) claimed that the Gini index and mean decrease accuracy is not significantly different, whereas the first important factors are almost the same. Similarly, in our model both attribute importance indexes state that speed is the most influential attribute in the modeling procedure. And, the second influential attribute is the accuracy of tracking GPS.

CONCLUSION
The major aim of this study was to present a new machine learning method and the critical role of defining attributes of classification algorithms for mode recognition and influential factors determination using GPS data. Three modes were considered including car, bus, and walking. A random forest model was applied. The outcomes of this model, not only provided higher accuracy in classification compared to previous studies, but also determined the most important feature in classifying each mode. The following conclusions have been derived:

1. The transportation mode detection was accomplished using a machine learning method (forest model) with an overall accuracy of almost 96% (more than previous studies).
2. In consistency with last studies, the first most important attribute to detect modes was instant speed in the process of modeling and prediction.
3. The second most effective attribute was related to the accuracy in GPS tracking data. Although GPS tracking accuracy has not been examined enough to date, it performed well in the random forest model.

FUTURE WORK
The future studies shall be concentrated on the modes which have the same attributes. For example, the speed ranges are very similar in walking and car in peak hours. Another investigation should be conducted on recognizing different motorized modes.

REFERENCES


