Developing and Optimizing a Transportation Mode Inference Model Utilizing Data from
GPS Embedded Smartphones

Akram Nour, corresponding author
Ph.D. Candidate
Department of Civil and Environmental Engineering
University of Waterloo
200 University Avenue West
Waterloo ON Canada N2L 3G1
anour@uwaterloo.ca
Tel: (519) 500 3453

Jeffrey Casello
Associate Professor
School of Planning and Department of Civil and Environmental Engineering
University of Waterloo
200 University Avenue West
Waterloo ON Canada N2L 3G1
jcasello@uwaterloo.ca

Bruce Hellinga
Professor
Department of Civil and Environmental Engineering
University of Waterloo
200 University Avenue West
Waterloo ON Canada N2L 3G1
bhellinga@uwaterloo.ca

Wordcount:
Abstract: 249
Text: 5259
Figures and tables (x250): 2000 (5 Figures and 3 tables)
Total: 7508
ABSTRACT

Advances in wireless communications and technologies provide the opportunity to collect detailed information on travel trajectory using smart-phones equipped with GPS and accelerometers. These types of smart-phones are ubiquitous and, as such, present an opportunity to conveniently collect spatial and temporal data at regular time intervals. This can be useful to utilize as a method to document travel behavior – origin, destination, departure time, route choice, trip purpose, and mode choice. One of the challenges that has been addressed in the literature is how to identify the transportation mode of travel.

The paper presents a data-driven classification model to infer transportation mode choice from data collected with GPS equipped smart phones. Rather than making a priori assumptions, we instead employ an optimization method to objectively produce the following classifier components and methods: a ranked feature vector based on the power of differentiation between different modes; the classification technique between the range of candidate classifiers; the number of ranked attributes to include in the feature vector; data formatting; and optimal model parameters.

The model is trained and tested using known transportation mode segments – limits of travel by a given mode. The calibrated model is evaluated by testing its ability to classify travel mode correctly for GPS data at a different level of disaggregation than the one used in the model training step. The model provides an accuracy of approximately 86% at the disaggregated level (e.g. Walk, Bike, Transit, and Private Automobile) and approximately 94% at aggregated level (e.g. Non-Motorized and Motorized.)
1. INTRODUCTION

Transportation planning and engineering rely on valid transportation data – origins, destinations, departure times, modes and paths – to inform decision making and design. Previously, these data were gathered through manual processes, often with limited accuracy. A growing field in transportation research is the collection, analysis and interpretation of automatically generated travel data from Global Positioning Systems (GPS). With GPS-enabled smart phones becoming nearly ubiquitous, transportation researchers now possess a very rich data set from which travel behavior can be gathered and be used to inform transportation infrastructure investments as well as policy analysis.

One common question addressed in the literature is how to use GPS data to determine the mode of travel. The development and implementation of a method to improve upon previous solutions to this problem is the focus of this research. Our approach builds upon previous work but introduces several novel methodological advancements. Generally, we attempt to limit the a priori assumptions about the data. Instead, we take a more holistic approach to the analysis that allows the data itself to determine the appropriate inputs and methods to determine travel mode.

More specifically, in this paper we relax the assumptions around the order of modes used for a given trip and the duration associated with each mode. Instead, we propose a technique that identifies travel segments – both moving and stationary – then attempts to classify the transportation modes used for the moving segments. Like others, we attempt to determine modes for each segment based on readily available data: maximum speed, average speed, acceleration rates, and jerk (rate of change of acceleration). Unlike others, however, we do not classify travel segments as belonging to a mode based on a single, weighted combination of travel characteristics. Instead, we utilize a classifying algorithm that iteratively introduces attributes from a feature vector such that those characteristics with the strongest differentiating powers between transportation modes are identified and used in the mode classification.

The proposed statistical method is based on a pattern recognition model. Several other researchers have attempted to identify the circumstances under which a given model produces the most accurate mode identification. In this research, multiple regression and factorial analysis have been used to investigate an exhaustive combination of factors to calibrate the classification model in order to optimize the model parameters. This allows our technique to produce more robust results for widely varying data sets.

The evaluation method used to assess the performance (and therefore transferability) of the model also differs from previous work. In our method, we train the model using the user-provided labels that define the beginning point, end point and mode that constitute a transportation mode segment. When we evaluate the model, we do not use these data. Instead, we begin by defining beginning and end points of a more disaggregate segment definition – moving (speed greater than a threshold) and stationary (speed less than a threshold). We then apply the mode classification model to infer the transportation mode of moving segments. In essence, this presents the classifier with a “new” data set – the same data but at a more disaggregate level – than what was used to train the model.

The remainder of the paper is organized as follows. The subsequent section reviews previous work on this topic. Section three describes conceptually the approach we take to mode identification, with detailed quantitative explanations as warranted. Section four demonstrates our model’s performance for data generated by travelers on multiple modes: walking, cycling, on-street bus transit and private automobile.
2. LITERATURE REVIEW

GPS data loggers have been utilized for collecting travel survey data, largely because of the lower burden on survey respondents and the higher accuracy compared to traditional trip diaries. Lee-Gosselin (8) classified the implementation of GPS travel surveys into three general categories. The first category imitates traditional travel (trip diary) surveys. The second category consists of methods that passively collect data though the GPS units, which is the focus of this research. The central idea of this approach is to collect travel data without requiring data inputs from the study participants. The challenge in this approach is inferring travel behavior and trip characteristics from raw GPS data. The third category is a combination of the first and second categories. GPS data are collected through either handheld or in-vehicle GPS units; then, participants are asked to review their GPS traces and annotate their travel components.

Chung and Shalaby (1) collected data using wearable GPS loggers and written detailed trip reports. They developed an algorithm to classify changes in mode in one of three categories: end-of-walk (EOW), start-of-walk (SOW), and end-of-gap (EOG), where the gap is the period when a loss of GPS signals occurred. They employ feature vectors to categorize SOW and EOW points such as speed and acceleration. A fuzzy logic-based model was used to classify the mode segments. This initial work correctly identified 55/60 mode segments. This work has become the foundation upon which many other researchers have built their classification models.

Moiseeva et al. (10) developed a system called “TraceAnnotator” that uses the Bayesian belief network (BBN) to estimate the transportation mode from GPS data from GPS loggers. The variables considered in the BBN also included speed and acceleration amongst others. The authors reported an overall accuracy of approximately 92%.

Gonzalez et al. (5) developed a smartphone application called TRAC-IT which collected GPS data and supplemental information from the user (e.g. number of occupants in the vehicle, transportation mode, trip purpose etc.). The data were analyzed using a Neural Network (NN) classifier to detect the transportation mode associated with the trip. The method was tested on a small data set. The reported accuracy of the system was quite high (91%) in predicting one of three modes.

Reddy et al. (11) developed a transportation classification system that uses GPS and 3-axis accelerometer. The classifier utilized a combined Decision Tree - Discrete Hidden Markov Model (DHMM) to classify five modes from a small and well-defined data set. The results presented indicate an overall accuracy of 93.6% in differentiating between motorized and non-motorized trips. The transition matrix was assumed to be constant over time, implying that transportation network changes or policy initiatives have no impact on these probabilities.

Zheng et al. (16) used a similar technique as proposed by Reddy et al. (11). They carried out an evaluation study in which they tested different discretization interval durations (i.e. for segmenting the GPS data) and found an overall accuracy of 75%. However, since their approach was based on the state transition matrix, it suffers from the same limitations identified previously.

Much of the previous work depends upon assumptions about trip constructs – e.g. all trips must begin and end with a walking segment that lasts longer than one minute. In addition, previous work has employed GPS data gathered at very short intervals (i.e. 1 second), primarily using data loggers. The literature reviewed has employed multiple classification techniques, parameters values, and data formats the performance of which is then compared. Ultimately, assessments are made based on the data sets used in the evaluation. Most importantly, previous
work has relied on what can be called a “static” feature set—a pre-determined set of travel attributes—to identify the transportation mode.

In our work, we propose a method that classifies transportation modes using fewer and less restrictive assumptions on trip structure. We also develop the model based on data gathered at longer intervals—five seconds. The model that we develop is automated such that multiple combinations of classification techniques, parameter values, and data formats are concurrently employed.

3. METHOD

The approach we take to solve the transportation mode inference problem can be summarized into three main stages: (1) data collection and processing; (2) training and testing the transportation mode classifier; and (3) evaluation by applying the classifier model to unlabelled data to infer transportation mode.

3.1 Stage 1: Collecting and Processing Data

GPS travel data— spatial data (x,y), speed, and time (t)—are automatically uploaded from participants’ smartphones to a secure server using a custom software application (15). Simultaneously, users are asked to label the time and location when they switch between transportation modes—for example leaving a car and beginning to walk. The potential labels are stored and transmitted with the GPS data. After uploading the data, participants have online access that allows for reviewing, modifying and verifying the recorded trip information.

In addition to manual modifications by the users, an automatic algorithm was developed to review the data to ensure that the participants’ indicated the change in mode at logically correct times and locations—i.e. not while traveling at high speed, but rather while stationary. FIGURE 1 presents a hypothetical trip with velocity plotted as a function of time. The algorithm corrects suspicious transportation mode transfer points—labeled as points 2, 5, and 6 in the diagram—by shifting the mode transfer label forward or backward along the time axis to the beginning or end of the adjacent stationary segment.

FIGURE 1 Outcomes of mode transfer point correction
3.2 Stage 2: Training and Testing the Classifier

The result of the previous step is a data set of verified trips which are the input into the second stage. As mentioned earlier, transportation mode inference can be solved as a pattern recognition problem using a number of supervised learning models (classifiers). The main goal of any supervised learning model is to build (train) a classifier using a known set of input data (feature vector) and known responses (labels or classes). This “trained classifier” is then capable of generating reasonable predictions for travel mode in response to new data.

We disaggregate the training stage into four sequential sub problems and propose a solution method for each. The sub problems include: Transportation Mode Segmentation (TMS); Feature Estimation and Selection (FV); Training the Transportation Mode Classifier; and Selecting the Optimal set of Attributes and Type of Classification Model. Optimality is achieved when misclassification rate (MCR) is minimized. Each algorithm described in the following sections.

**Transportation Mode Segmentation**

We define a trip as a time series of GPS points bounded by two activities. Within each trip, the boundaries of travel by a given transportation mode are known from user inputs. We further define Transportation Mode Segments (TMS) as portions of a trip that contain sequential data points labeled as having been completed by the same mode. A trip may be comprised of a single mode, in which case the entire trip will consist of one TMS. Alternatively, a trip may involve multiple transportation modes (and therefore multiple TMS), in which case one or more Mode Transfer Segments (MTS) must exist.

**Feature Estimation and Selection**

For all trips in the data set, we extract the transportation mode segments belonging to mode \( m \) amongst a set of candidate modes \( M \). In our case, we begin with four modes: private auto, public transit, cycling and walking. To quantify the characteristics of travel by a given mode, we develop a Feature Vector \( (FV^m) \) that contains a set of attributes that can be directly computed from the GPS data. For example, average speed, maximum speed, maximum acceleration, and jerk (the rate of change of acceleration) can be calculated for each mode from the GPS data. Mathematically, we can define a feature vector of dimension \( n \) of attributes based on a single transportation mode segment \( s1 \):

\[
FV^m = \{a^m_{1,s1}, a^m_{2,s1}, a^m_{3,s1} \ldots a^m_{n,s1}\}
\]

Alternatively, distributions of values can be developed for each attribute \( a \) of a given mode \( m \) when observations are made across all segments, \( s \). It is also possible to calculate mean values for an attribute belonging to a given mode.

Naturally, there exist many attributes that may be computed for a given mode. To improve the effectiveness of the mode identification process, it is beneficial to understand which quantitative attributes of transportation mode segments have the greatest differentiating power between modes. For example, we expect that maximum speed can be effective for determining if a trip segment is made by a motorized versus a non-motorized mode.

To formalize the process of quantifying the differentiating power of an attribute we use the following approach. For each attribute, we use a student \( t \) test (Case 3: two samples with unequal sizes and unknown variances) to conduct comparisons between the means of distribution to determine if the two sets are significantly different from each other. When we conduct this test
at a 95% confidence interval, the resulting $p$ values quantify the probability that the mean value for mode 1, $\bar{a}_{m1}^\text{1}$, is equal to the mean value for mode 2, $\bar{a}_{m2}^\text{2}$. Mathematically, we compute the $t$ statistic for the means of attribute $a$ for modes 1 and 2: $t_{a}^{m1,m2}$. We then calculate the degrees of freedom for the problem, $df$, such that the comparison can be made between the calculated $t$ statistic and the critical value. In this comparison, if $|t_{a}^{m1,m2}| > |t_{crit}|$, we reject the null hypothesis ($H_0$: both samples come from populations with equal means) at $\alpha = 5\%$ confidence level. If not, the test fails to reject the null hypothesis, suggesting the selected attribute is suitable for differentiating between the two transportation modes.

Simultaneously, once a $t$ value is estimated, a $P$ value - a scalar value in the range $[0, 1]$ - can be also determined using the Student’s $t$-distribution table. The $P$ value is the probability of observing a test statistic as (or more) extreme than the observed value under the null hypothesis. We elect to use $(1 - P$ value) as the statistic that quantifies the differentiating power ($DP$) of a feature vector attribute. Mathematically:

$$DP_{a}^{m1,m2} = (1 - P_{a}^{m1,m2})$$

Given four transportation modes, we can calculate $\binom{4}{2}$ or six pairwise comparisons for a given attribute as shown in TABLE 1. Ultimately, we wish to compute a single attribute differentiating power ($ADP$) statistic for each attribute. A simple way to convert the six $DP_{a}$ values to a single $ADP_{a}$ value is to take an average. We believe, however, that some pairwise comparisons are more important than others. We particularly attempt to emphasize the differentiation between motorized and non-motorized modes. As such, we elect to compute a weighted average as:

$$ADP_{a} = \frac{\sum_{m1=1}^{M} \sum_{m2=1}^{M} (DP_{a}^{m1,m2} \cdot \beta_{m1,m2})}{\sum_{m1=1}^{M} \sum_{m2=1}^{M} (\beta_{m1,m2})}$$

Where:

$\beta_{m1,m2}$ is the weighting associated with the comparison between mode $m_1$ and $m_2$; the default value of $\beta_{m1,m2} = 1$

Note that $\beta_{m1,m2} = 0$ for all blank cells in TABLE 1.

**TABLE 1 Differentiating Power for all pairwise comparisons**

<table>
<thead>
<tr>
<th></th>
<th>Walk</th>
<th>Bike</th>
<th>Transit</th>
<th>Auto</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>$DP_{a}^{m1,m2}$</td>
<td>$DP_{a}^{m1,m3}$</td>
<td>$DP_{a}^{m1,m4}$</td>
<td></td>
</tr>
<tr>
<td>Bike</td>
<td></td>
<td>$DP_{a}^{m2,m3}$</td>
<td>$DP_{a}^{m2,m4}$</td>
<td></td>
</tr>
<tr>
<td>Transit</td>
<td></td>
<td></td>
<td>$DP_{a}^{m3,m4}$</td>
<td></td>
</tr>
<tr>
<td>Auto</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We compute $ADP_{a}$ for all attributes. We can then sort the features based on their $ADP$; the one with the largest value for $ADP$ has the strongest differentiating power, and is ranked as...
The feature with the lowest ADP will have the weakest differentiating power and will be ranked \( N \).

The initial feature ranking based on ADP provides a good, but imperfect ordered list of attributes due to feature-to-feature correlations. We begin by entering the highest ranking attribute into what we call a set of chosen features. For all remaining attributes, the average correlation between the attribute in the chosen set is calculated. The next attribute to move from the candidate set to the chosen set is identified as the attribute with the largest Adjusted Attribute Differentiating Power (AADP) Score, which is calculated as follows:

\[
AADP_a = ADP_a(1 - \alpha \rho_a)
\]

Where:
\( \alpha \) is a user-defined parameter to determine the sensitivity to correlation; and
\( \rho \) is the average correlation between each candidate feature and the features in the chosen set.

The process is repeated until all attributes are ranked from highest differentiating power to lowest differentiating power in the chosen set.

**Training the Transportation Mode Classifier**

Transportation mode inference is classified as a pattern recognition problem that can be solved through a number of supervised learning models. We have shortened the list of the wide range of models to be tested to solve our classification problem based on models’ flexibility and ease of interpretation of results. The models we considered are: Naïve Bayes (NB), Quadratic Discriminant Analysis (QDA), and k-Nearest Neighbor (k-NN). Each of these techniques is briefly described here.

NB classifier is a simple probabilistic method that provides classification results as a probability distribution (degree of certainty) over a set of classes \((3, 12)\). The Bayes rule is applied to compute the posterior probability of class \( c \in X_c \) given the feature vector for a particular instance of \( x_1, x_2, \ldots x_N \) as follows:

\[
P(X_c | x_i) = \prod_{i=1}^{N} P(x_i | X_c)
\]

The second candidate classifier is Quadratic Discriminate Analysis (QDA) \((7, 12)\), a multivariate statistical technique widely used in pattern recognition and machine learning. In QDA, a class discrimination model that separates different classes by a quadratic surface is built. In order to classify a new sample, the trained classifier finds class \( c \) which maximizes the quadratic discriminate function (equation 7).

\[
h(x_i) = \arg \max_c P(x_i | c)
\]

\[
P(x_i | c) = \frac{1}{(2\pi | \Sigma_c |)^{1/2}} \exp \left( -1/2 (x_i - \mu_c)^T \Sigma_c^{-1} (x_i - \mu_c) \right)
\]

where \( x_i \) is the feature matrix for the new case;
\( \Sigma_c \) is the covariance matrix for class \( c \)

The third candidate of the classification models is k-Nearest Neighbors (k-NN). k-NN is a non-parametric method that consistently achieves high performance among the various methods of supervised statistical pattern recognition \((2, 4)\). The simple principle behind k-NN is
to calculate the distance between the $k$ nearest neighbors of training cases to the new sample in
the feature space. While various distance metrics can be used, Mahalanobis distance may be the
most appropriate. Mahalanobis considers both the variance and covariance of the feature vector
variables in order to measure the degree of similarity between the feature vector of the current
data and the training data (9). The consideration of the variance reduces the effect of those
variables with high variations while the common effects of correlated variables can be excluded
by considering the covariance of the feature vector variables. The Mahalanobis Distance, $MD$, is
calculated as:

$$MD_k = \left( (x_h^k - x_i^k)^T \Sigma_x^{-1} (x_h^k - x_i^k) \right)^{1/2}$$

(8)

Where:

- $MD_k$ is the Mahalanobis distance between training cases $h$ and new instance $i$
- $x^k$ is the feature matrix, and
- $\Sigma_x^H$ is the covariance matrix of training feature variables.

After estimating the distance metric, the new data point is classified based on the class
that has the majority from $k$ neighbors. Therefore in order to avoid ties, $k$ is commonly selected
as an odd number. In addition, it is logical to discriminate between the $k$ nearest neighbors in the
prediction in respect to their relevant distances from the new point. Therefore, a weighting
function is introduced to increase the influence of the closer neighbors, as defined by Shepard
(13):

$$w_i = \frac{1}{MD_k^2}$$

(9)

In addition to these three statistical techniques, and the parameters that influence their
performance, the performance of classifying algorithms is also influenced by the data format.
Options here include using discretized or non-discretized data and subsequently applying
dimensionality reduction through Principle Components Analysis (PCA). PCA has been shown
to have mixed results – in some cases enhancing the performance of the classification model
while in other cases PCA produces poorer results.

Given the breadth of options in the classification process, and the lack of a priori
knowledge of what may produce optimal results, the classification process was automated to
iteratively test all permutations. More specifically, the classification model includes and
optimizes over five variables:

1. **Alpha**: the degree to which we consider correlation in ranking Feature Vector attributes.
   This parameter varies from no consideration to full consideration over six levels
   \{0,0.2,0.4,0.6,0.8,1.0\}; (6 Levels)
2. **NF**: the number of features used to classify modes. This set of features may contain
   \{1,3,5,6,7,8,10,11\} variables; (8 Levels)
3. **Disc**: feature Discretization \{0,1\} binary variable (0 $\rightarrow$ Continuous, 1 $\rightarrow$ discrete); (2
   Levels);
4. **PCA**: the use of PCA \{0,1\} binary variable (0 $\rightarrow$ when original data is used, 1 $\rightarrow$ when
   PCA is applied ; (2 Levels);
5. **CL**: the classifier model: \{1: NB, 2: k-Nearest Neighbor, 3: DA\}; (3 Levels).
Selecting the Optimal Set of Attributes and Type of Classification Model

An experimental design method is adopted to determine the optimal value for each parameter. In order to have a better estimate for the variance of the error, we train the classifier model with five partitions using the stratified cross-validation technique, and test the quality of prediction on the remaining partition. We repeat this process six times to produce six independent prediction results. For each iteration, the MCR is estimated. Using the range of results, a linear regression model is calibrated. Essentially, we regress MCR against binary variables representing each parameter level. This approach allows us to determine a posteriori the combination of parameter estimation, data transformation and statistical classification technique that optimizes the performance of the model. This gives our modeling method significantly more flexibility than previous work, and increases the model’s performance.

3.3 Stage 3: Model Evaluation

The previous steps have developed an optimized classifying model that specifies which Feature Vector attributes should be used to identify modes; the degree of correlation to be considered amongst these attributes; whether data should be considered as discrete or continuous; whether the use of PCA improves the predictive power of the model; and whether k-NN, QDA or NB is the appropriate classification method.

To apply the trained classifier, we reset the level of aggregation of GPS data. More specifically, all transportation mode labels are assumed to be removed from the data set and we do not assume that the data have already been divided into the transportation mode segments. The only information maintained from the travel diary is the identification of the beginning and end of one trip. Because TMS are not known, we begin by identifying moving segments for which a transportation mode can be identified. A moving segment is defined as a series of points bound by at least one point with speed, $V$, less than a speed threshold, $V_0$ for duration of time between five and 120 seconds. FIGURE 2 shows the same hypothetical trip as FIGURE 1. In this case, the diagram shows the boundaries of moving and stationary segments.

FV Estimation and Treatment

The next step in the analysis is to apply the trained classification model and identify the most likely transportation mode for each moving segment. As illustrated in FIGURE 2, the classification model has labeled moving segments 1, 6, 7 and 11 as walking segments; moving segments 2, 3, 4 and 5 have all been labeled as auto segments; and moving segments 8, 9 and 11 are labeled as transit segments.
4. MODEL APPLICATION AND RESULTS

In order to evaluate the performance of the proposed model, we gathered GPS data from cell phones. In total, 658 trips were conducted with a total of 457,945 data points. FIGURE 3 illustrates a sample trip trajectory that includes four activities and three different transportation modes (Walk, Auto, and Bus).

4.1 Results

The result of each step in the proposed methodology is explained in this section following the same sequence in which the steps were described in Section 3.
Transportation Mode Segmentation

As explained earlier, the first step in the proposed methodology is to interpret the user-supplied information to determine the beginning and end points of the transportation mode segments (TMS). The results from this step, shown in TABLE 2, provide a statistical overview of the TMS for each mode.

### TABLE 2 TMS General Statistical Information

<table>
<thead>
<tr>
<th></th>
<th># TMS</th>
<th>Total Duration (min)</th>
<th>Max Duration (min)</th>
<th>Total #Points</th>
<th>Max #Points/Segment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>226</td>
<td>186.73</td>
<td>11.57</td>
<td>19993</td>
<td>874</td>
</tr>
<tr>
<td>Bike</td>
<td>109</td>
<td>166.17</td>
<td>8.72</td>
<td>14407</td>
<td>396</td>
</tr>
<tr>
<td>Transit</td>
<td>68</td>
<td>75.68</td>
<td>5.30</td>
<td>9002</td>
<td>527</td>
</tr>
<tr>
<td>Auto</td>
<td>454</td>
<td>1617.05</td>
<td>111.70</td>
<td>62292</td>
<td>1316</td>
</tr>
</tbody>
</table>

Feature Estimation and Selection

Next, statistical properties of attributes were computed. We calculated: average, maximum, minimum speeds, acceleration and jerk; we also computed the standard deviations, the 98th percentiles and the difference between the 98th and 50th percentile values for these three parameters of motion. We then ranked these attributes based on their ADP as shown in Equation 3. Regarding the weighting factor $\beta^{m_1,m_2}$, a sensitivity analysis has been done for the range 0.5 to 1; however, the impact on the ranking process was not significant. Therefore, the default value of 1 has been used.

The features with ADP less than 80% were eliminated, as the inclusion of weakly differentiating attributes actually diminished the classifiers performance. For instance, the minimum speeds for all transportation modes were almost equal – nearly zero – which created a common attribute and precluded differentiation. FIGURE 4 illustrates the ranked features and their ADP; those shaded in grey with values less than 80 were eliminated from the model.

![FIGURE 4 ADP of the Ranked Features](image)

Given this list of attributes to be included in the classifier model, we next attempt to identify the optimal parameters for the model: the degree of correlation to be applied in calculating $AADP$;
the data format; the classification technique and the parameters that define them; and the number of features (of the 11 brought forward from the previous step). The approach we took was to regress the values of each of these permutations against the misclassification rate \((MCR)\) – to determine the relative contribution of optimizing each parameter in the model. For the k-NN results, the number of nearest neighbors to be considered has been optimized in a separate but concurrent process. The results are shown graphically in FIGURE 5. The model produces optimal results with:
- The k-NN \((N=11)\) classification technique;
- All 11 attributes included in the model;
- PCA applied with 98% variation in the data; and
- Non-discretized data.

The regression results also indicate that \(\alpha\) is not statistically significant – that is correlation amongst the attributes does not meaningfully impact the model’s performance.

\[\begin{array}{c}
\text{PCA} = 0 \quad \text{DISC} = 0 \\
\text{PCA} = 1 \quad \text{DISC} = 0 \\
\text{PCA} = 0 \quad \text{DISC} = 1 \\
\text{PCA} = 1 \quad \text{DISC} = 1
\end{array}\]

**FIGURE 5** Results of the Relative Contribution of Optimizing Model’s Parameters

Although PCA is usually used for the purpose of dimensionality reduction, in this case PCA transforms the original data into linearly uncorrelated components which provides the best results, perhaps due to the strong correlation amongst the attributes in our data set. The inclusion of an increasing number of components in PCA helped explain an increasing amount of the variance in the data. Using a threshold of acquired retained variance of 98%, the first five ranked principal components have been selected. We have noted that the first three principal components explain almost 95% of the total variance in the data.

Based on the model structure determined in the previous step, we now calibrate the parameters of the classification model. Using the entire data set (at the transportation mode segment level), we apply a cross validation with 10 folds to train and test the final \(k-NN\)
classification model. The average results of this training and testing coincidentally produces an MCR of 9%.

4.2 Model Evaluation

As noted above, the evaluation methodology is not done at the transportation mode segment level. Instead, the trained, calibrated model is evaluated on a segment-by-segment basis - each moving segment in the full data set is classified as one of the four candidate modes. Given this approach, we are able to quantify all segments for which the classifier model predicts the correct transportation mode. In addition to computing overall performance, we also calculate recall - the number of segments correctly classified as mode \( m \) divided by the total number of segments that actually belong to mode \( m \). We also calculate recall at the point level. If a segment is determined to belong to mode \( m \), then all points contained in that segment are assigned mode \( m \). Using this approach, we can measure the impact of misclassification on long segments, with many points, and short segments with fewer points.

The classifier assigned the correct aggregate label – motorized (transit or auto) and non-motorized (walking or cycling) – to 94% of segments and 84.5% of points. The classifier assigned the correct mode to 70.3% of the segments and 85.9% of the points in the data set. TABLE 3 shows the classifier’s performance in terms of the confusion matrix for all four modes at the segment and point levels.

### TABLE 3 Confusion Matrix of Segments

#### Table 3a. Segment Level

<table>
<thead>
<tr>
<th>Reality</th>
<th>Walk</th>
<th>Bike</th>
<th>Transit</th>
<th>Auto</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>1379</td>
<td>8</td>
<td>6</td>
<td>6</td>
<td>98.57%</td>
</tr>
<tr>
<td>Bike</td>
<td>255</td>
<td>337</td>
<td>1</td>
<td>11</td>
<td>55.79%</td>
</tr>
<tr>
<td>Transit</td>
<td>116</td>
<td>59</td>
<td>53</td>
<td>349</td>
<td>9.19%</td>
</tr>
<tr>
<td>Auto</td>
<td>332</td>
<td>175</td>
<td>36</td>
<td>1434</td>
<td>72.53%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Reality</th>
<th>Walk</th>
<th>Bike</th>
<th>Transit</th>
<th>Auto</th>
<th>Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Walk</td>
<td>17126</td>
<td>61</td>
<td>132</td>
<td>153</td>
<td>98.02%</td>
</tr>
<tr>
<td>Bike</td>
<td>2601</td>
<td>9687</td>
<td>5</td>
<td>216</td>
<td>77.44%</td>
</tr>
<tr>
<td>Transit</td>
<td>1450</td>
<td>501</td>
<td>1542</td>
<td>4477</td>
<td>19.35%</td>
</tr>
<tr>
<td>Auto</td>
<td>2052</td>
<td>1417</td>
<td>377</td>
<td>53393</td>
<td>93.28%</td>
</tr>
</tbody>
</table>

The classifier performs very well in identifying walking segments, with nearly 96% recall. The model was able to correctly classify 55.8% of all bike segments; interestingly, the model’s performance was significantly improved for biking at the point level. This implies that errors in bicycle classification are a result of many short segments that are misclassified. A similar observation can be made for auto segments. The recall improves from 72.5% at the segment level to 93.3% at the point level. Further work is necessary to improve the classifier’s performance on these short segments.
Significant challenges were experienced in the classification of transit segments. The attributes of transit vehicle movements – in our case study buses operating in mixed traffic – are particularly difficult to distinguish from auto movements and, in some congested cases, walking segments. As with the recall for bike and auto modes, transit recall improved at the point level relative to the segment level. Further work is required to improve the classifier’s performance on both short segments and for transit.

5 CONCLUSIONS AND RECOMMENDATIONS

This study has made two main contributions. First, we have proposed an approach which can be used to objectively optimize the selection of the type of classification model, the number of feature vector attributes, the type of attributes to include in the feature vector, and whether or not to apply transformation techniques (e.g. discrete versus continuous and PCA) to the feature vector attributes. Second we have proposed a classification model and demonstrated its performance using field data. The proposed model requires fewer and less restrictive assumptions about the trip structure than most existing models and unlike many previous studies, the proposed model was evaluated under the more realistic conditions that the segmentation of trip data is part of the transportation mode estimation problem.

Our results indicate that overall the proposed model performs quite well. Many of the misclassification errors are associated with short segments and with distinguishing between transit and auto modes. We recommend that further work be conducted to improve the model in these two areas.

REFERENCES


