Developing and Validating a Statistical Model for Travel Mode Identification on Smartphones

Behrang Assemi, Hamid Safi, Mahmoud Mesbah, and Luis Ferreira

***Abstract*—Smartphone travel surveys are able to capture accurate details about individuals' travel behavior. However, extracting the required information (e.g., travel mode and purpose) from the data captured by smartphone applications is relatively complex, particularly when relying on the computational power of smartphones and limiting the communications between these applications and third parties [e.g., geographic information systems (GIS)]. These limitations are mainly enforced to enable passive data collection through smartphones by automatically recognizing the mode and purpose of trips. Furthermore, limited data transfer between the application and third parties ensures the privacy protection of survey participants and facilitates real-world travel surveys with large sample sizes. Accordingly, the objective of this paper is to develop a model of travel mode identification, which can be integrated with smartphone travel surveys without using GIS data or interacting with participants. Most existing models and algorithms are either inaccurate or computationally complex, and require extensive processing power. A smartphone travel survey, namely, the Advanced Travel Logging Application for Smartphones II (ATLAS II), has been used to collect individuals' travel data across New Zealand and Queensland, Australia. A detailed algorithm is put forward to clean the captured data, segment trips into single modal trips, and develop multiple statistical models for comparison, using the data collected from New Zealand. The preferred approach, which is adapted for the integration with smartphone travel survey applications, is validated using the two separate data sets from New Zealand and Australia. The resulting mode identification model (i.e., a nested logit model with eight variables) could detect travel modes with the accuracy of 97% for New Zealand after preprocessing (i.e., data cleaning and trip segmentation) and 79.3% for Australia without any preprocessing.**

***Index Terms*—Advanced Travel Logging Application for Smartphones II (ATLAS II), discriminant analysis, multinomial logistic regression, nested logit, adaptive categorically structured Lasso (CATS Lasso), smartphone travel survey.**

I. Introduction

SMARTPHONE travel surveys are increasingly gaining the attention of both researchers and practitioners, mainly due to their ability to collect comprehensive, accurate data about individuals' travel behavior [1]. However, given the complexity and size of data collected by these applications, little is known about efficient, simple methods of extracting the required information, such as travel mode and purpose, from the captured data [2]. Such methods are especially underexplored when strict limitations are applied to the interactions between the smartphone applications and participants as well as third party information providers, such as geographic information systems (GIS).

Previous studies, which have used the data collected by smartphone travel surveys or handheld geographic positioning system (GPS) devices, often rely on complementary data sources (e.g., GIS information or follow-up surveys) [e.g., [3], [4]–[8]]. A few automated mode identification algorithms are proposed in the literature, relying only on the collected data [2]. However, without additional information, either a wide range of variables (which requires significant processing power to compute) or complex algorithms (e.g., machine learning and neural networks) have been used for the analysis and model estimation [9]–[12]. There is an increasing demand for simple, automated procedures of analysis due to growing sample sizes, while there are very few such procedures that have been tested on large, real-world samples without manual intervention [13]. The main advantages of such procedures include: 1) allowing development of standalone smartphone travel surveys that predict travel modes without unnecessary interactions with users, and 2) eliminating the need for costly communication with third-party service providers, such as GIS databases.

To fill this gap, this research aims at providing an accurate, automated algorithm for travel mode identification, using only smartphone travel survey data (i.e., not using GIS data). The resulting mode identification model should be simple, allowing it to be integrated with smartphone travel surveys and to rely on the limited processing power of smartphones. To address this objective, a data collection approach is undertaken using the Advanced Travel Logging Application for Smartphones II (ATLAS II) smartphone travel survey in a pilot survey in New Zealand (NZ) [14], as well as Queensland, Australia. An algorithm is developed to clean the collected data and extract trip segments, each with one single mode of transport for a more precise model development. The identified trip segments are applied to develop three different mode identification models, namely, a multinomial logistic regression (MNL) model, a nested logit model and a multiple discriminant analysis (MDA) model, to compare the models and to adapt the best model for smartphone integration.

The findings of this study contribute to the body of knowledge on travel mode identification using smartphone travel surveys. The main contribution of the study is the development of statistical models which can be simply integrated with smartphone travel surveys to reliably identify travel modes on

Manuscript received March 19, 2015; revised August 27, 2015 and November 23, 2015; accepted December 28, 2015. Date of publication February 11, 2016; date of current version June 24, 2016. This work was supported in part by Australian Research Council Grant DE130100205. The Associate Editor for this paper was Q. Zhang.

The authors are with the School of Civil Engineering, The University of Queensland, Brisbane, Qld. 4072, Australia (e-mail: b.assemi@uq.edu.au; h.safi@uq.edu.au; mahmoud.mesbah@uq.edu.au; l.ferreira@uq.edu.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TITS.2016.2516252

TABLE I
BASIC STATISTICS OF RECORDED TRIPS IN NEW ZEALAND (NZ)

Mode	Frequency	Total length (km)	Total time (hours)	Overall mode share based on travel time (%)	NZ household travel survey mode share based on travel time (%)
Walk/Run	199	311.3	70.8	13.4	13.0
Bicycle	33	406.8	22.2	4.2	2.2
Car	1,012	12,057.1	406.8	76.8	78.4
Bus	44	390.2	24.8	4.7	6.4 (public transport and other modes)
Train	10	87.1	3.9	0.7	
Plane	3	237.8	1.2	0.2	
Sum	**1,301**	**13,490.3**	**529.7**	**100**	**100**

smartphones. The models do not rely on other datasets (e.g., GIS data), which are not necessarily available for every region. Furthermore, this integration facilitates passive travel data collection, as the smartphone application accurately identifies the mode of each trip immediately after its completion. Finally, using a representative, countrywide sample to develop and validate the models significantly enhances the validity and the generalizability of the findings.

The paper is structured as follows. The data collection method is described in the next section. Then, the theory, including the algorithm development and the data analysis methods, is discussed. The results of the analysis are presented, and the major findings of the study are discussed. Finally, some conclusions, along with the implications of this research for both theory and practice, are provided.

II. DATA COLLECTION METHOD

A. ATLAS II: Advanced Travel Logging Application for Smartphones II

ATLAS II (Advanced Travel Logging Application for Smartphones II), a state-of-the-art smartphone travel survey [14], [15], was used in this research to collect data. ATLAS II is an automated prompted recall travel survey developed for conducting individual travel surveys through iOS-running devices (i.e., iPhones and iPads). When installed on a smartphone, this application continuously runs in the background without interfering with the normal phone usage. When the phone is carried by its user beyond a specific distance threshold (400 meters in this study), the application automatically starts recording a trip by locally logging accurate GPS data. The data include the longitude, latitude, location accuracy, instant speed, heading, and the timestamp at which each log is captured. The application automatically stops recording the trip, when the user remains stationary for a while (2.5 minutes in this study). Hence, the application automatically segments the recorded data into separate trips in real-time. However, these segments should be split or merged together based on the definition of a complete trip, if required. The application initially records two separate trips as a single trip, when the time interval between the two trips is less than the specified threshold (i.e., 2.5 minutes). In contrast, the application initially splits a single trip into two separate trips, when there is a stop during the trip longer than the specified threshold (i.e., 2.5 minutes).

The participants are asked to review their recorded trips on their smartphones, label each trip by specifying the purpose(s) and the travel mode(s), and upload their labeled trips on the application server through an easy-to-use menu in the application. To assist participants with labeling their trips, the application visualizes the trajectory of each trip on a map and shows the origin address, destination address, start time, finish time and total distance. However, there is always the possibility that participants make mistakes while specifying the mode and purpose of each trip. This is a major motivation for developing smartphone travel surveys with passive data collection capability, which can eliminate the burden of trip labeling for participants.

B. Data Collection Procedure

Extensive data should be collected to develop and validate a travel mode identification algorithm. The data should include different travel modes and different traffic conditions, as the variables used in common mode identification algorithms (e.g., speed) are highly affected by traffic flows [2]. To address these requirements, ATLAS II was deployed in a countrywide travel survey in New Zealand, and in a smaller survey in Queensland, Australia. The data cover both urban and rural areas. The New Zealand data were used for model development and validation, while the Australia data were used for further validation.

The sampling and data collection in New Zealand were conducted by the Ministry of Transport, New Zealand using well-established sampling methods [16], in March-April 2014. In total, 76 participants uploaded their trips for an average duration of 4.9 days. After removing the "fake trips" (a fake trip is a series of "fake jumps" recorded by the smartphone application, and a fake jump is an unreal jump in the location detected by the smartphone's GPS due to signal inaccuracies), overall, 1,301 trips were collected, consisting of 13,490 km of trips, performed in almost 530 hours. The data include more than 757,000 GPS logs collected using an average sample rate of two seconds. The data include walk/run, bicycle, car, bus, train and plane trips (the train and plane trips recorded in the survey were excluded from the analysis, due to their limited occurrences, that is 10 train trips and three plane trips). Table I presents the basic statistics of the collected data and compares them with the latest New Zealand household travel survey results [17].

As shown in Table I, the car share is significantly higher than the share of the other modes, and the public transport share (i.e., bus and train in this study) is very small in the recorded trips. However, the mode shares are consistent with the results of the recent New Zealand household travel surveys [17]. This

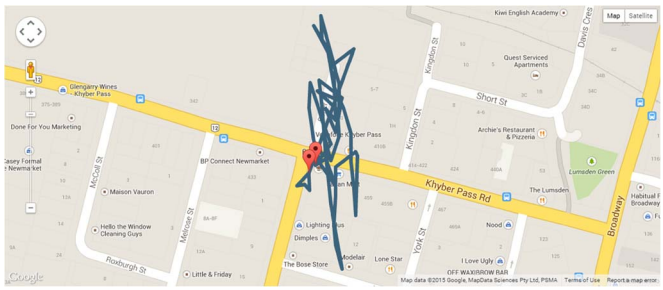


Fig. 1. Sample fake jumps detected by a smartphone's GPS sensor.

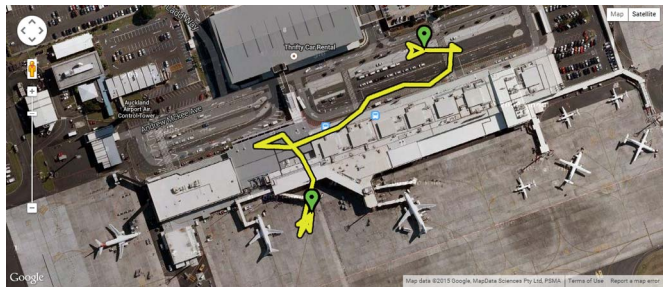


Fig. 2. Sample mistakenly motorized-labeled trip.

consistency supports the generalizability of our findings to the New Zealand context.

The data collected in Australia was used for model validation and transferability evaluation. The data include 193 trips uploaded by 99 participants. Of these trips, 3 are "bicycle," 39 are "bus," 114 are "car," and 37 are "walk/run."

III. Theory and Analysis

This section proposes an algorithm for travel mode identification based on the smartphone travel survey data collected in New Zealand. Although some rules and thresholds used in the algorithm are based on the findings of the existing literature, the algorithm as a whole is proposed and implemented for the first time in the current research. The data used to develop the algorithm only include the instantaneous movement attributes of survey participants recorded by ATLAS II in the form of tuples (timestamp, latitude, longitude, horizontal accuracy, instant speed, and heading). The algorithm consists of the following main steps, all implemented using R statistical software [18] and NLogit 5.0 [19], and explained in the subsequent subsections: cleaning data, identifying trip segments, identifying variables for mode detection, developing three initial mode identification models, and adapting the best mode identification model for smartphone integration. The algorithm is only used for model development relying on the data collected in New Zealand, and not for model validation/evaluation in Australia.

A. Data Cleaning

The data collected by ATLAS II may comprise inaccuracies (like any other smartphone travel survey) that should be cleaned before the data are used in mode identification model development. ATLAS II, however, performs the initial data cleaning while storing trip logs on a smartphone. The data cleaning within the application consists of two major steps. First, a location log captured by the smartphone is discarded while recording a trip, if the horizontal accuracy of the log is worse than 200 meters. Second, the logs of a finished trip are deleted if they are all located in a quadrilateral with diagonals smaller than 200 meters. It is very likely that such a trip is recorded by the application due to fake jumps mistakenly detected by the smartphone's GPS sensor, as shown in Fig. 1.

To remove the potentially inaccurate logs captured due to fake jumps and, thus, to ensure the data quality for mode identification, one further rule is applied to the collected data, as follows:

- Flag and remove every log in a trip when:
  - the distance between this log and its subsequent log is more than 50 m, and
  - the calculated average speed between the two logged points is greater than 30 m/s.

Furthermore, the mode of each recorded trip specified by the participants is examined in terms of its general correctness (i.e., whether it can be non-motorized or motorized). The trips which are potentially labeled incorrectly are excluded from the analysis. Fig. 2 demonstrates an example that is mistakenly labeled by a participant as a motorized trip, whereas it is obviously a walking trip. The general correctness of the travel modes is assessed, using the following rule:

- A trip can be non-motorized (i.e., bicycle or walk/run), if the average speed is less than or equal to 7.00 m/s (approximately 25 km/h), and the average acceleration is less than or equal to 0.23 m/s^2, given the average speed and mean acceleration of bicycle traffic in normal conditions [20]. Otherwise, the trip is potentially motorized (i.e., car or bus).

B. Identifying Trip Segments

The recorded trips should be split into single modal segments before the data can be used for mode identification model development. As ATLAS II automatically recorded the trips, it is likely that a recorded trip consists of multiple segments travelled using different modes of transport. However, a single modal segment is ideally a part of a trip that is travelled using only one single mode. To identify the single modal trip segments a procedure is proposed next (shown in Fig. 3). This procedure is developed based on the findings of the existing literature [2], [3], [5], [6], [8], [12], [13], [21], and the characteristics of ATLAS II travel survey application [14], [15].

By applying this procedure, the transfer points (where the mode is potentially changed) as well as the activity points (where a participant enters a building or stays still for a period of time) are detected for each trip. Then, the recorded trips are split at these points, because it is likely that the participants have changed their travel modes there. Although this procedure may split some single modal trips into several segments, oversplitting does not adversely impact on mode identification.

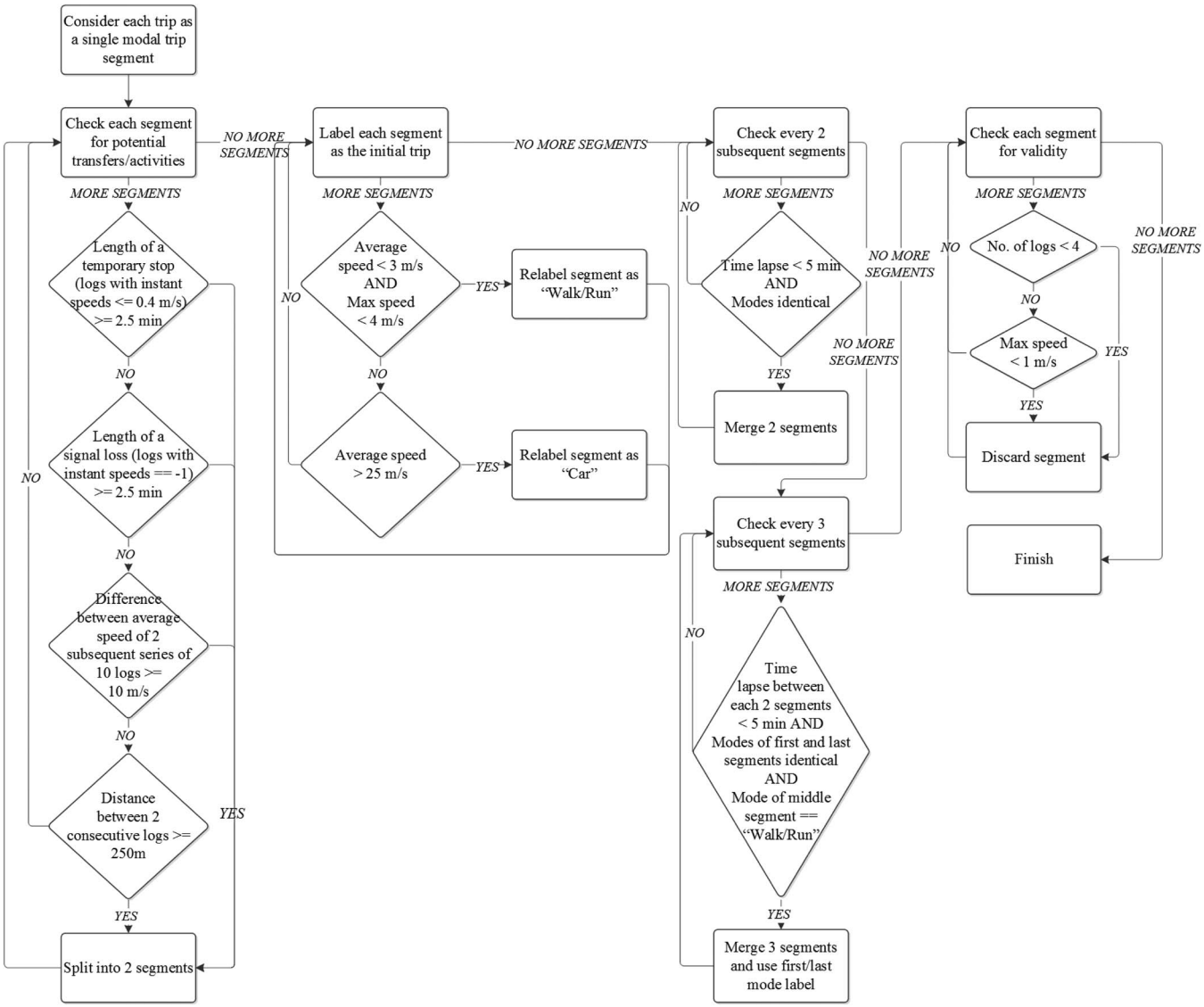


Fig. 3. Proposed trip segmentation algorithm for model development.

By contrast, not splitting the multi-modal trips into segments negatively influences the mode identification outcomes [2]. Over-split segments can be merged at a later stage, as elaborated next.

The trip segments' identification procedure proposed by this study is described as follows:

1. Create an initial set of trip segments:

   As discussed before, ATLAS II automatically splits a trip into two separate trips, when a stop occurs during the trip and lasts longer than 2.5 minutes. This threshold is shorter than the thresholds suggested in the literature [3], [7], [12], [13], mainly to avoid missing any parts of trips because of short stops during the trips. The recorded trips are considered as the initial set of trip segments, without any changes.
2. Split each trip segment by identifying transfer/activity points within the trip:

   A trip segment should be split into two segments at every potential transfer/activity point, since it is likely that the mode of transport changes at these points. Short stops and GPS signal losses are usually considered as the indicators of transfer/activity points. However, a significant change in the speed or a long distance between two consecutive logs can also indicate a transfer/activity point. A significant speed change may happen when the participant changes his/her mode of transport quickly (e.g., walking to a bus stop and immediately boarding a bus). A long distance between two consecutive logs indicates the participant's movements in an undercover area (e.g., a bus station), as the application has been unable to receive GPS signals during these movements. Accordingly, each trip segment is split into two, if either of the following conditions exists (the 2.5 minutes threshold in the first two conditions is to prevent over-splitting the trips due to stopping at traffic signals):

   - there is a stop that has not been detected by ATLAS II (logs with instant speeds less than or equal to 0.4 m/s) for at least 2.5 minutes, or

TABLE II
BASIC STATISTICS OF TRIP SEGMENTS FOR NEW ZEALAND (NZ) DATA

Mode	Number of trip segments	Total length (km)	Total time (hours)	Share of trip segments (%)	NZ household travel survey share of trip segments (%)
Walk/Run	316	331.1	83.8	23.8	16.1
Bicycle	17	163.4	10.9	1.3	1.8
Car	952	11,713.7	337.2	72.0	78.7
Bus	38	337.2	19.6	2.9	3.4 (public transport)
Sum	**1,323**	**12,545.4**	**451.5**	**100**	**100**

- there are continuous undefined speed (-1) logs (i.e., signal losses) for at least 2.5 minutes, or
- there is at least 10 m/s difference between the average moving speed of two subsequent series of 10 logs, or
- there is a distance of at least 250 meters between two consecutive logs.

The segmented trips are labeled with the same mode as the initial trip segments.

3. Re-label the trip segments:

While the trip segment modes may be different from the mode of the original trip (especially considering short walking trip segments before and after other types of trips that are extracted in the previous step), the identified trip segments may require re-labeling using two speed-related rules.

Re-label the trip segments with:

- the average moving speed less than 3 m/s, and the maximum speed less than 4 m/s as "walk/run," or
- the average moving speed greater than 25 m/s as "car" (the speed limits for buses and heavy vehicles are 22 m/s and 25 m/s respectively in New Zealand) [6], [22].

4. Merge over-split trip segments:

Splitting trips and re-labeling the resulting trip segments (as explained above) can be driven by some traffic incidents which cause short stops or signal losses (e.g., traffic congestion). Thus, the sequence of re-labeled trip segments should be reviewed to merge and re-label over-split trip segments. Accordingly,

- two trip segments are merged, when the time difference between them is less than five minutes, and their travel modes are identical (there is no need for re-labeling the resulting trip), or
- three trip segments are merged and the resulting trip is re-labeled, when the time difference between each two segments is less than five minutes, the modes of the first and the last segments are identical, and the mode of the middle segment is "walk/run." Use the first segment's mode for the resulting trip.

5. Delete invalid trips:

The segment identification procedure can form some short, invalid trips by separating the stationary parts of the recorded trips as trip segments. This usually happens for the initial as well as the last parts of trips, as the movements are often very slow and, thus, frequent logs are captured with instant speeds close to zero. To prevent potential inaccuracies in the resulting mode identification model, invalid trip segments are identified and removed, by applying either of the following rules:

- the number of logs in a trip segment is less than four, or
- the maximum speed of a trip is less than 1 m/s.

The result of trip segmentation for the New Zealand dataset entails 1,323 valid trajectories (about 12,545 km), travelled in 452 hours (some recorded trips were discarded in the data cleaning and trip segmentation procedures). These data include 666,017 valid GPS logs. The basic statistics of the trip segments along with the share of trip segments in the latest New Zealand household travel survey results [17] are presented in Table II.

C. Identifying Variables for Mode Detection

A wide range of variables was evaluated in this study to determine a minimal set that can accurately identify the travel mode, using the data collected by smartphone travel surveys only. The variables initially were used in three different statistical models to choose the most accurate modeling technique for mode identification. However, these variables were shortlisted to select the minimal set for the adopted modeling technique, namely nested logit analysis. Table III summarizes the list of 31 variables included in the initial mode identification models, along with the abbreviations used in this paper to refer to these variables.

A state-of-the-art variable selection method, namely adaptive categorically structured Lasso (CATS Lasso) [23], was used in this study to shortlist the variables. The standard forward or backward variable selection methods are not stable and cannot be applied for multinomial logit and nested logit analysis, because of the complex characteristics of this analysis method [23]. Adaptive CATS Lasso provides coefficients for the variables for each potential outcome, showing the impacts of these variables in determining the outcomes. We have used bootstrapping method [24] to determine the significance of the coefficients.

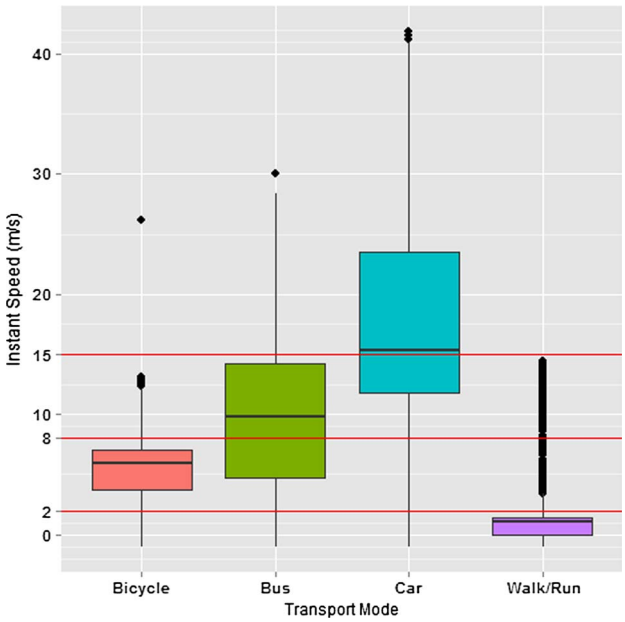
The variables evaluated in this study are proposed based on the findings of the existing literature [2], [5]–[8], [10], [12], [13], [25] and the nature of the data collected by ATLAS II [14]. These variables can be classified into seven major categories related to: 1) speed, 2) acceleration/deceleration, 3) orientation, 4) distance, 5) logging quality, 6) signal loss, and 7) temporary stop data (to evaluate their impacts).

Speed-related variables are the most obvious variables to distinguish different modes of transport, as frequently discussed in the literature. The maximum and 95 percentile of the speed are usually applied for mode identification, as the speed limits are different across various modes. For example, the maximum walking speed is considered to be 2.8 m/s [26], and the bus speed is limited to 22 m/s in New Zealand. Furthermore, it is shown in the literature that the median, mean, and standard deviation of the speed are also distinct for different modes. We suggest that the skewness and kurtosis of the speed distribution

TABLE III
INITIAL LIST OF VARIABLES EVALUATED FOR MODE DETECTION

Category	Variables[i]
Speed	Maximum (*SMAX*), 95 percentile (*SQ95*), median (*SMEDIAN*), mean (*SMEAN*), standard deviation (*SSTDDEV*), skewness (*SSKEW*) and kurtosis (*SKURT*) of speed distribution, and ratio of travel times with each of the following speed ranges to total travel time: below 2 m/s (*SBELOW2*), between 2 and 8 m/s (*SMID1*), between 8 and 15 m/s (*SMID2*), and more than 20 m/s (*SABOVE20*)
Acceleration/deceleration	95 percentile (*AQ95*), minimum (*AMIN*), maximum (*AMAX*), mean (*AMEAN*), median (*AMEDIAN*), variance (*AVAR*), standard deviation (*ASTDDEV*), skewness (*ASKEW*) and kurtosis (*AKURT*) of acceleration distribution, and mean of deceleration (*DMEAN)*
Orientation	Maximum change of orientation (*OMAXCH*) and average change in orientation (*OAVGCH*)
Distance	Total distance travelled (*DITOTAL*) and ratio of direct distance between origin and destination to travelled distance between the two points (*DIRATIO*)
Data quality	Average horizontal measurement error captured by the GPS sensor (*QAVGHACC*) and ratio of invalid records (i.e., the records with undetermined instant speeds) to total records (*QINVALID*)
Signal loss	Maximum signal loss duration (*SILOSSMT*) and distance travelled during the maximal signal loss (*SILOSSMD*)
Temporary stop	Number of temporary stops (*STCOUNT*) and average duration of temporary stops (*STAVGT*)

[i] *Variable abbreviations are shown in parentheses.*



Transport mode	Instant speed (m/s)			
	5 percentile	25 percentile	75 percentile	95 percentile
Bicycle	1.2	3.7	7.1	8.2
Bus	0.0	4.7	14.2	22.8
Car	2.3	11.8	23.6	28.4
Walk/Run	0.0	0.0	1.4	2.2

Fig. 4. Instant speed profiles for different travel modes in NZ dataset.

can also distinguish different modes, as these two variables indicate the speed pattern of trips.

Finally, the ratio of travel time with specific speed ranges to total travel time can distinguish different modes, as each mode has potentially a different speed profile. Considering the speed profiles for different modes of transport in New Zealand, extracted from the data and shown in Fig. 4, the ratio of travel times with the following speed ranges to total travel time were used in this study: below 2 m/s (for walk/run); between 2 and 8 m/s (for bicycle); between 8 and 15 m/s (for bus); and more than 15 m/s (for car).

Acceleration-related variables were also used by previous studies to distinguish different modes of transport. Given the literature findings, the 95 percentile, minimum, maximum, mean, median, variance, standard deviation, skewness and kurtosis of acceleration as well as the mean of deceleration were incorporated in the initial models to identify travel modes.

Orientation changes can also distinguish different modes, as the orientation may change frequently in some modes of transport (e.g., walk/run), while it hardly changes in the others (e.g., bus). Thus, two orientation-related variables were also included in the initial models: the maximum change of orientation and the average change in orientation.

Distance is another indicator of mode, as the normal distances travelled by non-motorized modes (i.e., walk/run and bicycle) are usually much less than the distances travelled by motorized modes. Accordingly, two main distance-related variables were included in the initial models: the total distance travelled and the ratio of direct distance between origin and destination to the travelled distance between the two points. The second variable was included in the analysis, because in some modes of transport (e.g., walk/run), a traveler is more likely to use more direct paths.

Data quality can also distinguish different modes because it is affected by different patterns of exposure to GPS signals depending on the mode. Generally, data quality is more likely to be lower in public transport trips, as the larger metallic body of these vehicles traps more GPS signals. Thus, two variables related to data quality were included in the initial models: the average horizontal measurement error (m) captured by the GPS sensor and the ratio of invalid records (i.e., the records without determined instant speeds) to total collected records.

Signal loss patterns were also shown by the literature as being distinctive for different modes. Two variables related to signal loss were included in the initial models: the maximum signal loss duration, and the distance travelled during the maximal signal loss.

Finally, temporary stop-related variables can also distinguish different modes of transport. The stop/start characteristics of bus trips are potentially different from other modes of transport.

Considering 30 seconds of immobility to be a temporary stop, the following two variables were also included in the initial models: the number of temporary stops, and the average duration of temporary stops.

D. Developing Initial Mode Identification Models

Three statistical modeling techniques were used in this study to find the most accurate approach for mode identification model development. These techniques include: multinomial logistic regression (MNL), nested logit, and multiple discriminant analysis (MDA). These techniques are discussed in this section, while the results of the model development and validation are presented in Section IV.

1) Multinomial Logistic Regression Model: The multinomial logistic regression (MNL) analysis is one of the most prevalent methods used to analyze the occurrences of discrete outcomes (i.e., modes of transport in this research) [27], [28]. The widespread use of this method in different areas of research is mainly attributed to its capabilities in simply providing the probabilities of each outcome, based on the values of independent variables (i.e., the above discussed variables in this study) [2]. The model presumes that the probability of each outcome is directly related to the variables characterizing that outcome [28]. Therefore, the model applies these probabilities to estimate the impact of each variable on the occurrences of discrete outcomes [29].

While applying the multinomial logistic regression analysis, the probability of outcome j (i.e., the occurrence of the jth mode) in observation i (i.e., $\pi_{ij} = \Pr\{Y_i = j\}$), can be specified by the following formula [30], [31]:

$$\pi_{ij} = e^{V_{ij}} / \sum_{j' \in \mathrm{J}} e^{V_{ij'}} \tag{1}$$

in which, J is the possible modes' set (including bicycle, bus, car, and walk/run), and V_{ij} is the utility of outcome j in the observation i, estimated by a linear combination of trip-related variables (discussed in the previous subsection) [32]:

$$V_{ij} = \alpha_j + \beta_{1j} x_{i1} + \cdots + \beta_{Kj} x_{iK} \tag{2}$$

in which, α_j is the general mean (i.e., a constant), β_{kj} is the impact of the variable x_k on the occurrence of outcome j compared to a base alternative, and x_{ik} is the value of the kth variable for observation i.

The probability of each travel mode in the observations is divided by the probability of a base alternative (usually the one with the highest occurrence, i.e., car in this study) to form the odds ratios (i.e., π_{ij}/π_{ij^*}, where j^* is the base alternative). Assuming the utility of the base alternative $V_{ij^*} = 0$, the logit transformation can be specified as [30], [33], [34]:

$$\mathrm{logit}(\pi_{ij}) = \ln\left(\frac{\pi_{ij}}{\pi_{ij^*}}\right) = \alpha_j + \beta_{1j} x_{i1} + \cdots + \beta_{Kj} x_{iK} \tag{3}$$

The optimal model is estimated by maximizing the following log likelihood function [35]:

$$\sum_i [Y_i - \pi_i] = 0 \tag{4}$$

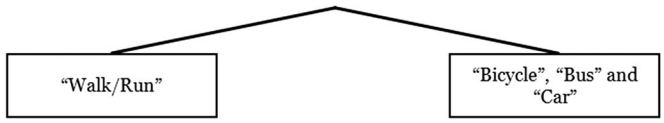


Fig. 5. Example nested structure of travel modes.

and

$$\sum_i x_{ik}[Y_i - \pi_i] = 0 \qquad \text{for} \quad k = 1, 2, \ldots, K \tag{5}$$

The results of the model estimation indicate the likelihood of the occurrence of each outcome, based on the value of variables X_k. In fact, the estimated coefficients indicate the impact of the relevant variables on the occurrence of a certain outcome.

2) Nested Logit Model: Applying the multinomial logit modeling, T_{in}, a linear function which determines the probability of outcome i for observation n, is considered as $T_{in} = \beta_i X_{in} + \varepsilon_{in}$, where β_i is a vector of estimable parameters for outcome i, X_{in} is a vector of the attributes that determine outcomes for observation n, and ε_{in} is the disturbance term. Thus, the probability of outcome i for the observation n, $P_n(i)$, is:

$$\begin{aligned} P_n(i) &= P(T_{in} \geq T_{In}) \\ &= P(\beta_i X_{in} + \varepsilon_{in} \geq \beta_I X_{In} + \varepsilon_{In}),\ \forall I \neq i. \end{aligned} \tag{6}$$

Assuming a random extreme value distribution [36] for the disturbance term, while all ε_{In}'s are independently and identically distributed, the previous formula can be revised to [37]:

$$\begin{aligned} P_n(i) &= P\left(\beta_i X_{in} + \varepsilon_{in} \geq \max_{\forall I \neq i} (\beta_I X_{In} + \varepsilon_{In})\right) \\ &= \frac{\mathrm{EXP}[\beta_i X_{in}]}{\sum_{\forall I} [\beta_I X_{In}]} \end{aligned} \tag{7}$$

which is the standard multinomial logit formulation. The model is estimated by maximizing the log likelihood of the outcomes [28]. The following log likelihood function is applied to estimate the parameters (i.e., β's):

$$LL = \sum_{n=1}^{N} \left(\sum_{i=1}^{I} \delta_{in} \left[\beta_i X_{in} - \ln \sum_{\forall I} \mathrm{EXP}(\beta_I X_{In}) \right] \right) \tag{8}$$

where δ_{in} is 1, if the outcome i occurs for observation n, and zero otherwise.

However, the assumption of independence of disturbance terms (ε_{in}'s) among alternatives, which is known as the independence of irrelevant alternatives (IIA), is very limiting. The nested logit model is a widely used modeling alternative for the standard multinomial logit that overcomes the IIA limitation [37]. The model groups different outcomes that potentially share unobserved effects into nests to address the issue of correlations among their disturbance terms [37]. To illustrate, Fig. 5 shows an example of the nested structure of the travel modes used for the analysis of the data in this study.

Applying the nested logit model, McFadden [38] has shown that the probability of outcome i in observation n can be

TABLE IV
PREDICTION ACCURACY OF THREE MODE IDENTIFICATION MODELS (%)

Mode	Walk/Run	Bicycle	Car	Bus	Overall
MNL model	100	100	99.6	26.3	**97.6**
Nested logit model	99.7	100	100	26.3	**98.4**
MDA model	98.7	100	86.3	55.3	**88.6**

estimated using the following formulas, given the disturbance terms are extreme value distributed:

$$P_n(i) = \frac{\text{EXP}[\beta_i X_{in} + \Phi_i LS_{in}]}{\sum_{\forall I} \text{EXP}[\beta_I X_{In} + \Phi_I LS_{In}]} \tag{9}$$

$$P_n(j|i) = \frac{\text{EXP}[\beta_{j|i} X_n]}{\sum_{\forall J} \text{EXP}[\beta_{J|i} X_{Jn}]} \tag{10}$$

$$LS_{in} = \ln \left[\sum_{\forall J} \text{EXP}[\beta_{J|i} X_{Jn}] \right] \tag{11}$$

where $P_n(i)$ is the unconditional probability of outcome i for observation n, $P_n(j|i)$ is the probability of outcome j for observation n conditioned on the outcome being in the nest of outcomes i, J is the conditional set of outcomes (conditioned on i), I is the unconditional set of nests, LS_{in} is the inclusive value (logsum) and Φ_i is an estimable parameter. According to McFadden [38], Φ_i's should be greater than 0 and less than 1 to conform to the model derivation. The difference between each Φ_i and 1 shows the significance of the assumed shared unobserved effects in the corresponding nest [37]. As recommended by the literature, we used advanced software packages to estimate all nests simultaneously, using full-information maximum likelihood [37].

3) Multiple Discriminant Analysis Model: Multiple discriminant analysis (MDA) is another multivariate analysis method which is used when the dependent variable (mode in this study) is categorical (with more than two categories) and the independent variables (variables shown in Table III) are numerical. The objective of MDA is to develop multiple linear combinations of the independent variables (i.e., discriminant functions) that differentiate best the objects in different groups (i.e., belonging to different categories of the dependent variable). The number of discriminant functions depends on the number of categories of the dependent variable: $NG - 1$, where NG is the number of categories [39]. Each discriminant function Y_i is specified by the following formula:

$$Y_i = \beta_0 + \beta_1 x_1 + \cdots + \beta_K x_K \tag{12}$$

in which, β_0 is the intercept, β_k is the discriminant coefficient for the independent variable k, and x_k is the independent variable k [39], [40].

The discriminant coefficients are calculated for each discrimination function to maximize the differences between the group means (i.e., centroids that are the means of the discriminant scores obtained by applying the discriminant function in each group) [39], [40]. In fact, MDA maximizes the between-group variance, which is different to multinomial logistic regression and nested logit analysis. The latter methods minimize the within-group variance by maximizing the log-likelihood in each group, as discussed above.

E. Model Validation

To develop and validate the model, the pre-processed New Zealand dataset was randomly divided into two separate datasets, which were used for model development and validation consecutively. To ensure the inclusion of the existing travel modes in each dataset, the random sampling was applied on the trips' data for each mode separately. 90% of the trips' data for each travel mode (overall 1,192 trips) were randomly selected for the model development and the remaining 10% (overall 131 trips) were dedicated for model validation.

For further validation and evaluation of the model, the developed model was applied to the data collected in Australia. As the model is intended to be integrated with smartphone travel survey applications, none of the data cleaning and trip segmentation procedures discussed in Section III were applied on this dataset. Indeed, the raw data were used to validate the model and to test its transferability to another geographical area. The results are discussed in Section IV.

IV. RESULTS AND DISCUSSION

After cleaning the New Zealand dataset, the data used to develop the mode identification model were reported by 75 participants, comprising 25 males and 50 females. The participants were recruited from different parts of New Zealand (Auckland: 29, Christchurch: 7, Wellington: 7, Other North Island Cities: 22, and Other South Island Cities: 10). The recruitment procedure also covered a wide range of age groups as well as household incomes.

The results of the evaluation of the initial MNL, nested logit and MDA models (with all variables) are provided in Table IV. As shown in the table, the overall mode detection accuracy is considerably higher in the MNL and nested logit models compared to the MDA model (i.e., 97.6% and 98.4% vs. 88.6% respectively). As shown "walk/run" and "bicycle" trips are identified very accurately with all models, "car" trips are detected very accurately by the MNL and nested logit models, and the "bus" trips are not detected very well by any of the models. The accuracy of the MDA model in identifying bus trips is higher than the accuracy of the other two models. This can be attributed to the nature of the models, as the MDA model maximizes the between-group variance for different modes, while the MNL and nested logit models minimize the within-group variance for each mode.

Although low levels of accuracy in identifying bus trips is the case with most of the existing models/algorithms [e.g., [2], [5], [6]], they can be partially attributed to the small number of public transport trips in the studied sample (i.e., 38 trips). The inclusion of GIS data in the analysis can improve the model's accuracy in detecting bus trips, as the accurate trajectories of trips can be compared with public transport

TABLE V
ADAPTIVE CATS LASSO RESULTS FOR VARIABLE SELECTION

Variable	SSKEW	SMID1	SMID2	SMAX	AQ95	AVAR	AMAX	DIRATIO
Mode	Coefficient (Bootstrap p-value)							
Walk/Run	0	-8.17**	-1.63	-0.22	-2.56***	0	0	0
Bicycle	0	9.81***	0	0	0	0	0	0
Car	-0.44	0	6.99**	0.51***	1.31*	5.66***	-0.53	-4.23**
Bus	0.96**	0	0	0.12	0	-0.49	0.74***	2.24**
Table description: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$								

TABLE VI
NESTED LOGIT MODEL FOR SMARTPHONE INTEGRATION (BASE ALTERNATIVE: "CAR")

Mode	Walk/Run		Bicycle		Bus	
Variable	Coefficient ($t-value$)	Max Elasticity	Coefficient ($t-value$)	Max Elasticity	Coefficient ($t-value$)	Max Elasticity
Intercept	20.67*** (2.91)	---	-10.20*** (-3.91)	---	-0.76 (-1.33)	---
SMID1	-16.49** (-2.10)	-1.90	13.39*** (3.99)	2.02	---	---
SMID2	-7.62* (-1.74)	-2.74	---	---	---	---
SSKEW	---	---	---	---	1.09*** (4.95)	-0.67
SMAX	-1.03*** (-2.59)	-16.28	---	---	-0.13*** (-4.37)	-2.68
AVAR	---	---	---	---	-0.95** (-2.11)	-0.41
AMAX	---	---	---	---	0.20*** (3.85)	0.80
AQ95	-5.51** (2.56)	-5.80	---	---	---	---
DIRATIO	---	---	---	---	1.47** (2.00)	0.95

Table description: * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

routes. However, for the integration of the proposed model with smartphone applications, GIS variables were not included in this study.

Finally, the nested logit model was selected for integration with smartphone travel surveys, because of its higher overall accuracy and advantages over the other two models. Although the results of the MNL and nested logit model are very similar, the nested logit technique is preferred, because of relaxing the IIA assumption [37].

As discussed in Section III, adaptive CATS Lasso method [23] was used to shortlist the variables. The results are shown in Table V. The variables which could not distinguish at least two different outcomes were discarded (i.e., the variables with three or four coefficients equal to zero). As shown, each of the remaining eight variables could distinguish at least two different outcomes and was significant for at least one of the outcomes. These variables were used to specify the utility functions (V's) for the final nested logit model, as below:

$$\begin{aligned} V(\text{Walk/Run}) &= \beta_{w0} + \beta_{w1} \times (SMID1) + \beta_{w2} \times (SMID2) \\ &\quad + \beta_{w3} \times (SMAX) + \beta_{w4} \times (AQ95) \\ V(\text{Bicycle}) &= \beta_{b0} + \beta_{b1} \times (SMID1) \\ V(\text{Bus}) &= \beta_{u0} + \beta_{u1} \times (SSKEW) + \beta_{u2} \\ &\quad \times (SMAX) + \beta_{u3} \times (AVAR) + \beta_{u4} \\ &\quad \times (AMAX) + \beta_{u5} \times (DIRATIO) \end{aligned}$$

The results of the nested logit model using the development dataset from New Zealand are summarized in Table VI, and discussed in the remainder of this section. The maximum elasticities are provided for each variable, which are the elasticities when the actual and estimated outcomes are the same in this model. The model fits well taking into consideration its satisfactory goodness-of-fit parameters:

- The log likelihood of the model is Log(Likelihood) $= -136.38$, $\chi^2_{14} = 3{,}370.4$, and $p < 0.001$.
- The adjusted McFadden R-square of the model is $R^2 = 0.925$.
- $\Phi = 0.794$(Std.Error $= 0.058$), $t = 3.55$ for the nest with bicycle, bus, and car trips.

The final nested logit model consists of eight variables: 1) skewness of speed distribution, 2) share of travel time with the speed ranging from 2 to 8 m/s, 3) share of travel time with the speed ranging from 8 to 15 m/s, 4) maximum speed, 5) acceleration 95 percentile, 6) acceleration variance, 7) maximum acceleration, and 8) ratio of direct distance to travelled distance between origin and destination. As shown in Table VI, the coefficients are all significant.

While all four variables of the utility function of the "walk/run" mode have significant impacts on identifying this travel mode, the maximum speed has the largest impact (given its maximum elasticity of -16.28). Given "car" as the base

TABLE VII
PREDICTION ACCURACY OF INTEGRATION MODEL FOR NEW ZEALAND DATASET

Mode / Labeled	Detected as	Walk/Run	Bicycle	Car	Bus	Total trajectories	Detected correctly (%)
Walk/Run	*Development*	**285**	0	1	0	**286**	**99.7**
	Validation	**29**	1	0	0	**30**	**96.7**
Bicycle	*Development*	0	**12**	3	0	**15**	**80.0**
	Validation	0	**1**	1	0	**2**	**50.0**
Car	*Development*	1	4	**852**	0	**857**	**99.4**
	Validation	1	0	**94**	0	**95**	**98.9**
Bus	*Development*	0	0	31	**3**	**34**	**8.8**
	Validation	0	0	4	**0**	**4**	**0**
Sum	*Development*	**286**	**16**	**887**	**3**	**1,192**	**96.6**
	Validation	**30**	**2**	**97**	**0**	**131**	**94.7**

TABLE VIII
PREDICTION ACCURACY OF INTEGRATION MODEL FOR RAW DATA FROM AUSTRALIA

Mode / Labeled	Detected as	Walk/Run	Bicycle	Car	Bus	Total trajectories	Detected correctly (%)
Walk/Run		**36**	0	1	0	**37**	**97.3**
Bicycle		0	**3**	0	0	**3**	**100**
Car		0	1	**113**	0	**114**	**99.1**
Bus		2	0	36	**1**	**39**	**2.6**
Sum		**38**	**4**	**150**	**1**	**193**	**79.3**

alternative in the model, when the maximum speed increases by 1 m/s, the chance of the walk/run mode decreases by a maximum of 16.28%. This result is consistent with the speed profiles for walk/run and car modes, shown in Fig. 4. Similarly, when the acceleration 95 percentile increases by 1 m/s^2, the likelihood of the walk/run mode decreases by a maximum of 5.80% compared to car.

With bicycle utility function having only one variable (i.e., SMID1), this variable appears to have a significant impact on determining the corresponding travel mode. Consistent with the speed profiles presented in Fig. 4, when the share of travel time with speed ranging from 8 to 15 m/s increases by 1%, the probability of the bicycle mode increases by a maximum of 2.02% compared to car (given the maximum elasticities).

For bus trips, the variables are all significant too. The maximum speed has the largest impact, given a 1 m/s increase of this variable decreases the probability of bus as the travel mode by a maximum of 2.68% compared to car. This is also consistent with the speed profiles shown in Fig. 4. An interesting finding is the positive impact of distance ratio on the likelihood of bus mode. This finding indicates that buses use a more direct route from an origin to a destination, compared to cars. The positive impact of acceleration maximum on the likelihood of bus mode seems counterintuitive, as it is generally expected that cars have higher acceleration compared to buses. This impact is, however, due to the inaccuracy of the estimation of acceleration, based on the instant speeds recorded by smartphones. This inaccuracy is often higher for bus trips, because of lower GPS signal reception in buses and the resulting jumps in the instant speeds recorded by smartphones. This issue can be addressed by using smartphones' accelerometer to record acceleration more accurately.

Table VII illustrates the mode detection accuracy of the model for both New Zealand development and validation datasets. As shown in the table, the overall accuracy of the final model (i.e., 96.6% for the development sample and 94.7% for the validation sample) is very close to the overall accuracy of the nested logit model with all variables (i.e., 98.4%). Applying the model on the validation sample results in relatively similar accuracy levels. Given the satisfactory level of estimation accuracy of the adopted model with eight variables, the model is a reasonable alternative to the initial model with 31 variables, especially for integration with smartphone travel surveys.

Table VIII illustrates the results of model validation on the "raw" data collected in Queensland, Australia. As shown, the results are very similar to the results of the model validation using the "pre-processed" New Zealand dataset. The accuracy of the model is very high for identifying "walk/run," "bicycle" and "car" trips, while it is very low for identifying "bus" trips. As this was the case for the pre-processed New Zealand dataset, these results suggest that the model can be reliably applied to another geographical location (i.e., Australia) and without any requirements for data processing (i.e., data cleaning and trip segmentation).

V. CONCLUSION

Three multivariate statistical models were developed to identify travel modes based on the data collected by smartphone travel surveys. The best model was then adopted to enable its integration with smartphone travel surveys. Four prevalent travel modes in New Zealand were considered in the model: walk/run, bicycle, car, and bus. Of the initial list of 31 variables, eight variables were selected for the final model to detect travel modes, using the adaptive categorically structured Lasso (CATS Lasso) method with bootstrapping. These variables include: the skewness of speed distribution, share of travel time with the speed ranging from 2 to 8 m/s, share of travel time with the speed ranging from 8 to 15 m/s, maximum speed, acceleration 95 percentile, acceleration variance, maximum acceleration, and ratio of direct distance to travelled distance between origin and destination. These variables can be simply recorded or calculated by smartphone applications (e.g., ATLAS II) while

recording participants' trips, without any interactions with third party information providers. Furthermore, given the model's high accuracy in identifying travel modes without requiring data cleaning or trip segmentation, it can be integrated with smartphone travel surveys to suggest the mode used for each trip, instead of asking participants to label their travel modes.

The findings of this research have theoretical and practical implications. Firstly, a model of travel mode identification was developed based on real data from a sample of travelers across New Zealand. The representativeness of the data used to develop the model ensures the reliability of the outcomes. The promising results of the model validation in Australia ensure its transferability to other geographical areas.

Secondly, the model can be integrated with smartphone applications. The model's independent variables can all be captured or calculated by smartphones, without any interactions with third parties. The results of the model validation in Australia show that the data cleaning and trip segmentation are not required for mode identification, using the proposed model. The integration of a mode identification model with smartphone travel surveys is an important prerequisite for the development of passive data collection applications.

Thirdly, as the mode identification model does not rely on third party information providers (e.g., GIS databases), the integration of the model with smartphone travel surveys improves the users' experience in data collection. Such an integration provides participants with automatically detected modes to validate them, instead of asking the participants to label their travel modes. Moreover, smartphone travel surveys (e.g., ATLAS II) that are equipped with the proposed model do not require participants' travel data to be transferred to any server for mode identification. This makes the data collection process more reliable, easier and cheaper, while ensuring the participants' privacy.

Some potential avenues for future research have also been identified. The accuracy of mode identification needs to be improved for public transport trips. On the one hand, other potential variables and modeling techniques should be evaluated to develop more accurate models of mode identification. On the other hand, larger samples with more public transport trips should be collected to develop more accurate mode identification models, as the inaccuracy of the proposed model in identifying public transport trips can be due to the limited number of relevant trips in the collected dataset (i.e., 38 bus trips). Finally, the inclusion of more sophisticated modes (e.g., taxi) in the model is an important gap to address by future research.

ACKNOWLEDGMENT

The authors are sincerely grateful to the Ministry of Transport, New Zealand, and especially to Mr. Matt Jones, for their support. The New Zealand data used in this research were collected by TNS Global for the 2014 Household Travel Survey pilot, on behalf of the Ministry of Transport. The authors would also like to express their appreciation to Professor Mark Hickman, Professor Fred Mannering and Professor Simon Washington, for their valuable advice.

REFERENCES

[1] L. Shen and P. R. Stopher, "Review of GPS travel survey and GPS data-processing methods," *Transp. Rev.*, vol. 34, no. 3, pp. 316–334, May 2014.

[2] C. Rudloff and M. Ray, "Detecting travel modes and profiling commuter habits solely based on GPS data," presented at the 89th Annual Meet. Transportation Research Board, Washington, DC, USA, 2010.

[3] W. Bohte and K. Maat, "Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in The Netherlands," *Transp. Res. C, Emerging Technol.*, vol. 17, no. 3, pp. 285–297, Jun. 2009.

[4] J. Wolf, R. Guensler, and W. Bachman, "Elimination of the travel diary: Experiment to derive trip purpose from Global Positioning System travel data," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1768, pp. 125–134, 2001.

[5] C. Chen, H. Gong, C. Lawson, and E. Bialostozky, "Evaluating the feasibility of a passive travel survey collection in a complex urban environment: Lessons learned from the New York City case study," *Transp. Res. A, Pol. Pract.*, vol. 44, no. 10, pp. 830–840, Dec. 2010.

[6] H. Gong, C. Chen, E. Bialostozky, and C. T. Lawson, "A GPS/GIS method for travel mode detection in New York City," *Comput., Environ. Urban Syst.*, vol. 36, no. 2, pp. 131–139, Mar. 2012.

[7] S. Tsui and A. Shalaby, "Enhanced system for link and mode identification for personal travel surveys based on Global Positioning Systems," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 1972, pp. 38–45, 2006.

[8] F. Biljecki, H. Ledoux, and P. van Oosterom, "Transportation mode-based segmentation and classification of movement trajectories," *Int. J. Geogr. Inf. Sci.*, vol. 27, no. 2, pp. 385–407, Feb. 2012.

[9] P. A. Gonzalez *et al.*, "Automating mode detection for travel behaviour analysis by using Global Positioning Systems enabled mobile phones and neural networks," *IET Intell. Transp. Syst.*, vol. 4, no. 1, pp. 37–49, Mar. 2010.

[10] C. Xu, M. Ji, W. Chen, and Z. Zhang, "Identifying travel mode from GPS trajectories through fuzzy pattern recognition," in *Proc. 7th Int. Conf. FSKD*, 2010, pp. 889–893.

[11] R. Brunauer, M. Hufnagl, K. Rehrl, and A. Wagner, "Motion pattern analysis enabling accurate travel mode detection from GPS data only," in *Proc. 16th IEEE ITSC*, 2013, pp. 404–411.

[12] A. Bolbol, T. Cheng, I. Tsapakis, and J. Haworth, "Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification," *Comput., Environ. Urban Syst.*, vol. 36, no. 6, pp. 526–537, Nov. 2012.

[13] N. Schuessler and K. Axhausen, "Processing raw data from Global Positioning Systems without additional information," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2105, pp. 28–36, 2009.

[14] H. Safi, B. Assemi, M. Mesbah, L. Fereira, and M. Hickman, "Design and implementation of a smartphone-based system for personal travel survey: Case study from New Zealand," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2526, pp. 99–107, 2015.

[15] H. Safi, B. Assemi, M. Mesbah, and L. Ferreira, "A framework for smartphone-based travel surveys: An empirical comparison with alternative methods in New Zealand," in *Proc. 10th Int. Conf. Transp. Survey Methods*, Leura, NSW, Australia, 2014. [Online]. Available: https://www.researchgate.net/publication/277249402_A_Framework_for_Smartphone-Based_Travel_Surveys_An_Empirical_Comparison_with_Alternative_Methods_in_New_Zealand

[16] Detailed Travel Survey Information, Ministry of Transport, Wellington, New Zealand, Mar. 7, 2014. [Online]. Available: http://www.transport.govt.nz/research/travelsurvey/detailedtravelsurveyinformation/

[17] "New Zealand household travel survey 2009–2012: Comparing travel modes," Ministry Transp., Wellington, New Zealand, 2013, Accessed: Mar. 7, 2014. [Online]. Available: http://www.transport.govt.nz/assets/Import/Documents/NZ-household-travel-survey-Comparing-travel-modes-April-2013.pdf

[18] R Core Team, R: A Language And Environment for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing, 2014.

[19] W. H. Greene, *NLOGIT Version 5.0: Reference Guide*. Plainview, NY, USA: Econometric Software, Inc., 2012.

[20] J. Parkin and J. Rotheram, "Design speeds and acceleration characteristics of bicycle traffic for use in planning, design and appraisal," *Transp. Pol.*, vol. 17, no. 5, pp. 335–341, Sep. 2010.

[21] E. H. Chung and A. Shalaby, "A trip reconstruction tool for GPS-based personal travel surveys," *Transp. Plan. Technol.*, vol. 28, no. 5, pp. 381–401, Oct. 2005.

[22] Speed Limits, New Zealand Transport Agency, Wellington, New Zealand, 2013. [Online]. Available: http://www.nzta.govt.nz/resources/roadcode/about-limits/speed-limits.html
[23] G. Tutz, W. Pößnecker, and L. Uhlmann, "Variable selection in general multinomial logit models," *Comput. Stat. Data Anal.*, vol. 82, pp. 207–222, Feb. 2015.
[24] P. Bühlmann and T. Hothorn, "Boosting algorithms: Regularization, prediction and model fitting," *Stat. Sci.*, vol. 22, no. 4, pp. 477–505, Nov. 2007.
[25] T. Feng and H. J. P. Timmermans, "Transportation mode recognition using GPS and accelerometer data," *Transp. Res. C, Emerging Technol.*, vol. 37, pp. 118–130, Dec. 2013.
[26] P. R. Stopher, Q. Jiang, and C. FitzGerald, "Processing GPS data from travel surveys," presented at the 2nd Int. Colloq. Behavioural Foundations Integrated Land-Use Transportation Models, Frameworks, Models Applications, Toronto, ON, Canada, 2005.
[27] J. Chen and P. Chitturi, "Choice experiments for estimating main effects and interactions," *J. Stat. Plan. Inference*, vol. 142, no. 2, pp. 390–396, Feb. 2012.
[28] D. Raghavarao, J. B. Wiley, and P. Chitturi, *Choice-Based Conjoint Analysis: Models and Designs*. London, U.K.: Chapman & Hall, 2011.
[29] W. G. Zikmund, S. Ward, B. Lowe, and H. Winzar, "Measuring consumer utility," in *Marketing Research: Asia Pacific Edition*. Thomson, Vic., Australia: Cengage Learning, 2007, pp. 467–484.
[30] M. E. Ben-Akiva and S. R. Lerman, *Discrete Choice Analysis: Theory and Application to Travel Demand*. Cambridge, MA, USA: MIT Press, 1985.
[31] W. Marshall and N. Garrick, "Effect of street network design on walking and biking," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2198, pp. 103–115, 2010.
[32] M. Roorda and B. Andre, "Stated adaptation survey of activity rescheduling: Empirical and preliminary model results," *Transp. Res. Rec., J. Transp. Res. Board*, vol. 2021, pp. 45–54, 2007.
[33] G. Rodríguez, "Multinomial response models," in *Lecture Notes on Generalized Linear Models*. Princeton, NJ, USA: Princeton Univ. Press, 2007.
[34] G. Bham, B. Javvadi, and U. Manepalli, "Multinomial logistic regression model for single-vehicle and multivehicle collisions on urban U.S. highways in Arkansas," *J. Transp. Eng.*, vol. 138, no. 6, pp. 786–797, Jun. 2012.
[35] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Hoboken, NJ, USA: Wiley, 2013.
[36] E. Gumbel, *Statistics of Extremes*. New York, NY, USA: Columbia Univ. Press, 1958.
[37] S. P. Washington, M. G. Karlaftis, and F. L. Mannering, *Statistical and Econometric Methods for Transportation Data Analysis*. Boca Raton, FL, USA: CRC Press, 2010.
[38] D. McFadden, "Econometrics models of probabilistic choice," in *Structural Analysis of Discrete Data With Econometric Applications*, C. F. Manski and D. McFadden, Eds. Cambridge, MA, USA: MIT Press, 1981.
[39] J. F. Hair, R. Anderson, B. Black, B. Babin, and W. C. Black, *Multivariate Data Analysis: A Global Perspective*. Upper Saddle River, NJ, USA: Pearson Education, 2010.
[40] M. Vetrivel Sezhian, C. Muralidharan, T. Nambirajan, and S. G. Deshmukh, "Attribute-based perceptual mapping using discriminant analysis in a public sector passenger bus transport company: A case study," *J. Adv. Transp.*, vol. 48, no. 1, pp. 32–47, Jan. 2014.

Behrang Assemi received the B.Eng. degree in software engineering from Amirkabir University of Technology (Tehran Polytechnic), Tehran, Iran, in 2003, the M.Sc. degree in computer science from Sharif University of Technology, Tehran, in 2006, and the Ph.D. degree in information systems from the University of New South Wales, Sydney, Australia.

Since 2013, he has been a Research Fellow with the School of Civil Engineering, The University of Queensland, Brisbane, Australia. He has published his research at peer-reviewed information systems and transportation journals and conferences, including Decision Support Systems and Transportation Research Record. His current research interests include road safety, travel surveys, live traffic information systems, route choice modeling, and crowdsourced data collection. Dr. Assemi is a Section Editor of the Australian Journal of Information Systems.

Hamid Safi received the M.Sc. degree in transport planning and engineering from Sharif University of Technology, Tehran, Iran, in 2005 and the Ph.D. degree in transportation engineering from the University of Queensland, Brisbane, Australia, in 2016. He has multiple papers published and presented in peer-reviewed transport and traffic-related journals and conferences. His main research focus is on proposing a comprehensive framework for smartphone-based travel data collection methods. His research interests include travel surveys, travel behavior analysis, and machine learning.

Mahmoud Mesbah received the B.Sc. degree in civil engineering from the University of Tehran, Tehran, Iran, the M.Sc. degree in transportation planning from Iran University of Science and Technology, Tehran, and the Ph.D. degree in transportation engineering from Monash University, Melbourne, Australia. In 2011, he joined the School of Civil Engineering, The University of Queensland, Brisbane, Australia, as a Lecturer. His research interests are transport modeling and planning, optimization of transport systems, public transport, planning for cycling facilities, and application of advanced technologies in transport.

Luis Ferreira received the M.Sc. degree in transport planning and management from the University of Westminster, London, U.K., in 1976, and the Ph.D. degree in transport modeling from the University of Leeds, Leeds, U.K., in 1984.

He is currently a Professor with the School of Civil Engineering, The University of Queensland, Brisbane, Australia. He has a strong multimodal research and consulting background encompassing road and rail, freight, and passenger transport. He has worked for 30 years in technical and managerial roles covering transport planning, research, management, and consultancy. He has been closely involved with transport and traffic planning and modeling, evaluation, and performance measurement of transport programs and projects as a practitioner, researcher, and trainer. His research interests include transport evaluation, intelligent transportation systems, safety, incident management, and driver behavior. He has an extensive and distinguished road and rail related publications record in international journals and conferences.