Hierarchical Process of Travel Mode Imputation from GPS Data in a Motorcycle-Dependent Area

Minh-Hieu Nguyen, Jimmy Armoogum

Hierarchical Process of Travel Mode Imputation from GPS Data in a Motorcycle-Dependent Area

AUTHORS' INFORMATION

Minh Hieu Nguyen, Corresponding author
PhD
Université Gustave Eiffel / AME / DEST
14-20 Boulevard Newton, Champs sur Marne, F-77447 Marne la Vallée Cedex 2, France
Faculty of Transport Economics, University of Transport and Communications (UTC)
03 Cau Giay Street, Dong Da District, Hanoi, Vietnam
Email: hieunm@utc.edu.vn
ORCiD: https://orcid.org/0000-0003-4863-5446

Jimmy Armoogum
PhD, Senior researcher
Université Gustave Eiffel / AME / DEST
14-20 Boulevard Newton, Champs sur Marne, F-77447 Marne la Vallée Cedex 2, France
Email: jimmy.armoogum@univ-eiffel.fr

Hierarchical Process of Travel Mode Imputation from GPS Data in a Motorcycle-Dependent Area

ABSTRACT

This study presents attempts to impute modes from data collected via smartphones in Hanoi (Vietnam), where the dominant mode of travel is the motorcycle. The inclusion of the motorcycle mode and an imbalance in the modal share of the Hanoi data resulted in ineffective use of supervised-learning models to detect all modes simultaneously. For a high level of accuracy and reasonable interpretability, a hierarchical process was developed. Initially, walk, bicycle, and motorized modes were identified by a fuzzy logic-based algorithm. Subsequently, based on the distribution of bus stops and the operation of buses in practice, rules employing the average distance between stops, which a vehicle passed slowly or stopped at, were introduced to detect bus segments. Finally, a random forest model was built to distinguish the modes of motorcycle and car. The proposed hierarchical process achieved an accuracy of 89.1%. The bus detection, which required only the coordinates of the bus stops, demonstrated a recall of 87.2%. The motorcycle mode of travel was noted to be the main source of misclassification. This mode has contributed to the diversity of the mode detection field, which has previously only focused on walk, bicycles, cars, buses/trams, and trains. The hierarchy was developed and validated using a dataset that did not include travel by metro or train and would be biased toward persons working and studying at a university. These limitations emphasize the need to test the process on a more diverse sample with more travel options.

Keywords: mode imputation, GPS, travel survey, smartphone, hierarchical process, fuzzy logic

HIGHLIGHTS

- Inclusion of motorcycle raised the complexity of mode classification with the most significant conflict being between car and motorcycle.
- Besides acceleration and speed related variables, the use of heading change rate in a fuzzy-logic algorithm contributed to satisfactory detection of walk, bike and motorized modes.
- Consideration of both slowly passing and stopping at bus stops enhanced bus detection via addition of the average distance between bus stops.
- The proposed hierarchical process benefitted from rule-based, fuzzy logic-based and random-forest methods to achieve 89.1% accuracy.

1 INTRODUCTION

The Global Positioning System (GPS) has contributed significantly to the collection of data in travel behavior research (Chen et al., 2010; Burkhard et al., 2020; Nguyen et al., forthcoming). Compared with conventional data collection methods, such as face-to-face interviews or computer-assisted telephone interviews, GPS collects data continuously and passively, which are objective and accurate in terms of time and space (Armoogum et al., 2014; Chen et al., 2010; Forrest and Pearson, 2005; Thomas et al., 2018; Wolf et al., 2003). Such data are frequently referred to as big spatiotemporal data. Unfortunately, positioning data may not be directly appropriate for research goals owing to the lack of information about the travel mode—one of the most important trip characteristics. Consequently, many inference models have been developed to detect basic travel modes, including walk, bicycle, bus/tram, car, and train (Bohte and Maat, 2009; Dabiri and Heaslip, 2018; Feng and Timmermans, 2019, 2016; Gong et al., 2012, 2018; Marra et al., 2019; Nour et al., 2016; Rasmussen et al., 2015; Schuessler and Axhausen, 2009; Semanjski et al., 2017; Shafique and Hato, 2015; Stenneth et al., 2011; Stopher et al., 2008; Tsui and Shalaby, 2006; Xiao et al., 2015). However, the motorcycle, which is a major mode of transport in a number of cities in developing countries (Nguyen and Pojani, 2018; Huynh, 2020), has not been included in these models.

Mode detection methods can be divided into two categories: all-in-one and hierarchical processes. The former type, which uses only one model to infer all modes, has been preferred in recent times (Dabiri and Heaslip, 2018; Feng and Timmermans, 2019; Gong et al., 2018; Semanjski et al., 2017; Shafique and Hato, 2015; Xiao et al., 2015). However, such methods may fail to perform well in cases wherein the data contain modes used substantially more than other modes. In hierarchical processes, travel modes are classified from aggregate levels (e.g., motorized and non-motorized modes) to disaggregated levels (e.g., walk, bicycle, bus, car, and metro).

The aim of this research was to create a hierarchical process to impute travel modes from GPS data collected in a motorcycle-dependent city of a developing country to extend knowledge of the mode detection field. The remainder of this paper is organized as follows. Section 2 reviews recent studies on mode detection to determine the advantages and disadvantages of the methods. Section 3 provides detailed descriptions of data collection in Hanoi, and the steps of the hierarchy are proposed. Section 4 presents the results and discussion, and section 5 concludes the paper and suggests future research directions.

2 REVIEW OF STUDIES ON MODE DETECTION

2.1 Mode-detection methods

According to Gong et al. (2014), mode imputation methods can be divided into three main groups: deterministic, probabilistic, and machine-learning methods. Deterministic methods are based on predefined, ad-hoc rules of speed, acceleration, and distance to locations of bus stops and train stations (Bohte and Maat, 2009; Gong et al., 2012). They are simple and easy to interpret because they are based on transport practices. For example, a trip would be inferred to have been taken on foot if its nearly maximum (i.e., the 85th percentile of) speed does not exceed 10 km/h and its average speed does not exceed 6 km/h (Gong et al., 2012). Rules are useful for cases wherein the specific characteristics of travel modes are shown; however, these rules are insufficiently flexible to deal with the reality of travel, such as the slow movement of almost all modes on congested roads. The performance of deterministic methods depend largely on experts' experience and knowledge of the travel environment in the research area of interest. Furthermore, the use of a large number of variables is not practical, because this would result in an exponential increase in the number of rules, defined as the combinations of variables.

The second type of mode inference method (probabilistic) involves extensions of rules in fuzzy-logic-based models (Rasmussen et al., 2015; Schuessler and Axhausen, 2009; Tsui and Shalaby, 2006) and a probability matrix (Stopher et al., 2008). Instead of strictly making decisions, probabilistic methods consider the overlap of modes' behaviors to generate probabilities for each of the modes simultaneously. The mode that has the highest probability is attributed to the trip. Probabilistic approaches are flexible classifiers; however, they share the limitations of deterministic methods mentioned above owing to their reliance on devising rules. Most deterministic and probabilistic models use fewer variables than there are transportation modes, with accuracy levels of over 90% being attained in GPS-enabled tests (Rasmussen et al., 2015; Stopher et al., 2008; Tsui and Shalaby, 2006) but a level of only 70% being reached in a regional experiment (Bohte and Maat, 2009). Schuessler and Axhausen (2009) ignored the model accuracy owing to the absence of ground truth.

Machine-learning algorithms (e.g., support-vector machines, random forests, and artificial neural networks) are currently preferred owing to their ability to learn directly from big data to effectively classify modes (Dabiri and Heaslip, 2018; Feng and Timmermans, 2019; Gong et al., 2018; Semanjski et al., 2017; Shafique and Hato, 2015; Stenneth et al., 2011; Xiao et al., 2015). The advantage of machine-learning methods over deterministic and probabilistic methods is that they create powerful classifiers using various types of variables, the number of which exceeds the number of classes (Bzdok et al., 2018). For example, Feng and Timmermans (2016) used 17 variables related to movement, participants' information, and the quality of GPS points to detect 10 modes. The contribution and importance of variables in machine-learning models may be confusing to some extent, as their combination and interaction occur in black-box and mathematically complicated processes. Notably, mode use has not been balanced with the majority of trips belonging to some modes (e.g., walk, car) and the small minority of trips employing other modes, such as bus/tram and bicycle (Dabiri and Heaslip, 2018; Nour et al., 2016; Xiao et al., 2015). If big data are not collected, the imbalance in the mode shares would mean that relatively little data of minor modes would be used to train the model; thus, their detection would be significantly poorer than that of major modes. If adequate data are provided, the overall accuracy levels exceed 90% (Semanjski et al., 2017; Xiao et al., 2015) and can reach nearly 100% (Feng and Timmermans, 2016; Shafique and Hato, 2015).

Although the lists of transportation modes and the methods used vary across studies, all researchers have paid close attention to basic modes in developed countries and modern cities of China, including walk, bicycle, car, bus/tram, and metro (Bohte and Maat, 2009; Dabiri and Heaslip, 2018; Feng and Timmermans, 2019; Gong et al., 2012, 2018; Marra et al., 2019; Semanjski et al., 2017; Shafique and Hato, 2015; Stopher et al., 2008; Xiao et al., 2015).

2.2 Mode detection process

Mode detection processes can be divided into two types according to the number of steps taken to generate an outcome. The first type includes all-in-one processes, in which the modes of all trips are detected by only one model. The second uses a hierarchical process to build a multi-step procedure to infer modes at aggregate levels (e.g., non-motorized and motorized modes) prior to disaggregated levels (i.e., each mode). Machine-learning models are typically used for all-in-one processes, and so they have both the advantages and disadvantages of the learning methods discussed above.

Hierarchical processes are based on separating modes that are sufficiently different from one another; thus, the division can attain a very high rate of success. The simplest versions of hierarchical classification are rule-based methods. In Bohte and Maat (2009), for example, walking trips were detected first using maximum and average speed, as walking is the slowest mode of transport. With a rule-based hierarchical process, Gong et al. (2012) achieved an accuracy level of 82.6% for the case of New York. A complex hierarchical classification can be developed by combining methods. Rasmussen et al. (2015) successfully imputed 92.4% of trip segments. First, they distinguished rail segments from

others by examining the proximity of points on each segment to the rail network. They then developed a fuzzy-logic algorithm based on speed and acceleration to distinguish walking and cycling segments from car and bus segments. Finally, they resolved the confusion between car and bus segments by using map-matching algorithms. Nour et al. (2016) applied the k-nearest neighbor algorithm to categorize all data into aggregate levels (i.e., motorized and non-motorized modes) and then disaggregated levels (walk, bicycle, car, bus). To enhance the detection of bus and car, all segments of motorized modes were analyzed to determine whether they involved bus segments by estimating the average rate of stopping close to transit stations. The researchers found increases of 65% and 10% in recall and precision of bus, respectively, compared with those of k-nearest neighbor. Marra et al. (2019) introduced a process that involved first segmenting trips into walk and non-walk segments on the basis of speed- and time-based rules. Next, train and bus/tram segments were identified by probabilistic functions using actual operational data of public transport. Historically visited places and routes extracted from multi-day GPS data improved the detection of transfer points. Finally, a random forest model was developed to infer bicycle and car trips. The proposed system was tested in Zurich and Basel (Switzerland); an accuracy of 86.1% was attained regarding classifying the four modes, namely, walk, bus/tram, car or bicycle, and train, and an accuracy of 87% was attained regarding classifying the modes bicycle and car.

The advantage of the hierarchical process over the all-in-one process is noticeable in the variable selection. In an all-in-one process, each variable affects all trips unnecessarily, leading to an incorrect classification. Semanjski et al. (2017) reported that adding a speed variable to a support vector machine model, that had previously used only spatial variables, resulted in a small improvement of overall accuracy at the expense of increasing the misclassification of walk and bicycle segments. In each step in a hierarchical process, some unique variables for several mode groups are deployed to limit any confusion. Moreover, a hierarchy does not use numerous variables simultaneously but focuses on mode groups; therefore, it can fit small and imbalanced data with a low risk of overfitting. However, error propagation is a problem. As researchers consider more disaggregated levels, the accuracy will decrease. The accuracy of steps lower in the hierarchy is equal to or lower than that of the step above. If an observation is wrongly classified in the previous step, there is no way to correct it in the subsequent stage.

Thus, although all-in-one processes with machine-learning methods are currently preferred, hierarchical processes would be a better choice to overcome the problem of imbalanced data and to improve the interpretability (see Table 1). The inclusion of new travel modes in the classification list is interesting and contributes to the diversity of the mode detection field.

Table 1. *Comparison of hierarchical and all-in-one processes*

	Hierarchical process	**All-in-one process**
Definition	Detect modes from aggregate levels to disaggregated levels	Detect all modes simultaneously
Main algorithms used	Rule-based and probability-based	Learning-based
Frequently used in	Infancy of GPS-based travel surveys	Recent times
Appropriate data size	Both data in tests and big data	Big data
Interpretability of results	High	Low
No. of variables used	Usually several; fewer than modes	Many; usually more than modes
Main modes classified	Walk, bicycle, bus/tram, car, metro, and train	
Geographical research scope	Mainly in cities, metropolitan areas of developed countries, and well-structured cities in China	

3 DATA AND METHOD

3.1 Data

3.1.1 Data collection

For this research, the study area was Hanoi, the capital of Vietnam. In terms of population, it is the second largest city, with 7.32 million inhabitants in an area of 3,344.4 km^2. The central area of Hanoi comprises eight districts (Caugiay, Badinh, Hoankiem, Dongda, Haibatrung, Thanhxuan, Hoangmai, and Tayho) and is surrounded by suburban and rural areas. Daily mobility in Hanoi mainly involves the use of private vehicles, predominantly the motorcycle (Nguyen et al., 2019b).

The data used in this research were collected from mid-March through mid-April 2019. The recruitment was implemented through invitations sent to our colleagues at University of Transport and Communications in Hanoi and a call for participation posted on the author's Facebook page. Anyone who expressed their interest in the survey was contacted so that his/her personal information could be collected before (s)he received instructions (in Vietnamese) and installed TravelVU, a survey-dedicated smartphone application developed by Trivector (Sweden). Respondents easily found the TravelVU on the App Store and Google Play Store, the two most prevalent digital distribution platforms. The TravelVU app recorded the participant's longitude, latitude and timestamp every 1–3 seconds. When an internet connection was available, all of the raw GPS points were transferred to Trivector's servers and then analyzed by in-built inference algorithms. Next, trip legs (i.e., segments) and associated activities were re-sent to the smartphones. This enabled the participants to simply check and correct labels of segments (i.e., ground truth data).

Finally, of the more than 80 people who registered, 63 had validated data that were composed of both records and the corresponding ground truth. As the first metro line was not in operation at the time of the research, the dataset encompassed five modes, including walk, bicycle, motorcycle, car, and bus. A more detailed description of this recruitment process can be seen in Nguyen et al. (2019a).

3.1.2 Data filtering

The validated segments of 63 participants were used to develop a hierarchical mode-detection process. As the records included only timestamps and coordinates in the World Geodetic System 1984 format, the distance between two consecutive points was calculated using Vincenty's equations (Vincenty, 1975). With distances and timestamps, speed and acceleration profiles were easily gauged. The criteria used to filter the data were as follows:

- The speed limit for roads in Vietnam is 120 km/h; therefore, points with speeds over this threshold were removed[1].
- Any point that had the same coordinate or timestamp as its previous point was ignored.
- Any segment with a duration of under 60 seconds was ignored.
- Any segment whose points were all outside Hanoi was excluded because this research did not use GIS data of the public transport systems in other provinces or cities. Travel connecting Hanoi with other provinces, however, was still within the scope of this study.

After these filters were applied, 2,791 segments were eligible for further analysis. As can be seen in Figure 1, the data of the Hanoi survey were unbalanced, with major classes including motorcycle, car, and walk, and minor classes including bicycle and bus.

[1] It is important to note that a point with a speed over 120 km/h may not constitute noise or a bad record. Such points may represent an over-speeding situation of a car on an expressway. In Hanoi, the highest speeds of motorcycles and buses are around 70 km/h (the allowable level) because they are banned from running and do not operate on expressways, respectively. For this reason, disregarding points with speeds of over 120 km/h, enabling less computation, did not decrease the detection performance of segments by car, motorcycle and bus. However, if the comprehensive distribution of speed is desired in case of the speed limit at 120 km/h, the threshold to eliminate erroneous records may be 150 km/h, according to Wang et al. (2017).

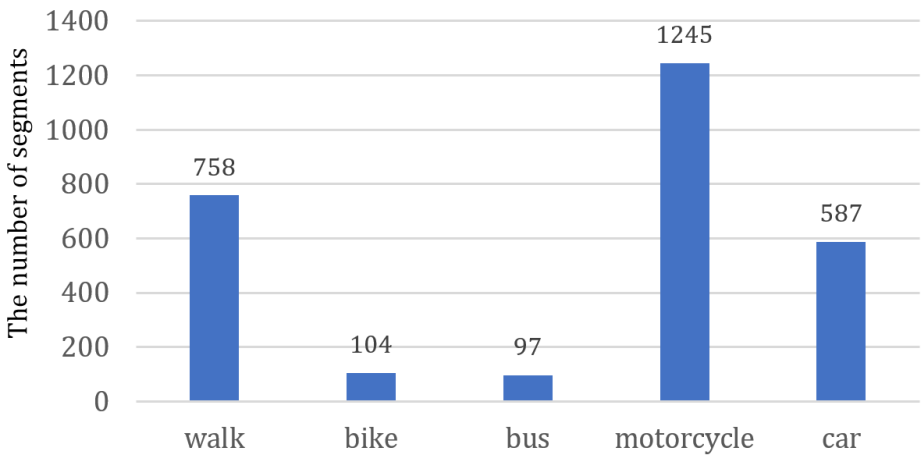


Figure 1. *Numbers of valid segments by modes in the Hanoi survey*

3.2 Hierarchical mode imputation process

The hierarchical mode inference process encompassed three steps (see Figure 2). The process began by distinguishing walk and bicycle segments from motorized ones using a fuzzy logic algorithm. In the second step, bus segments were separated from other motorized segments using stop-related rules. In the last step, car segments were distinguished from motorcycle segments using a random forest model.

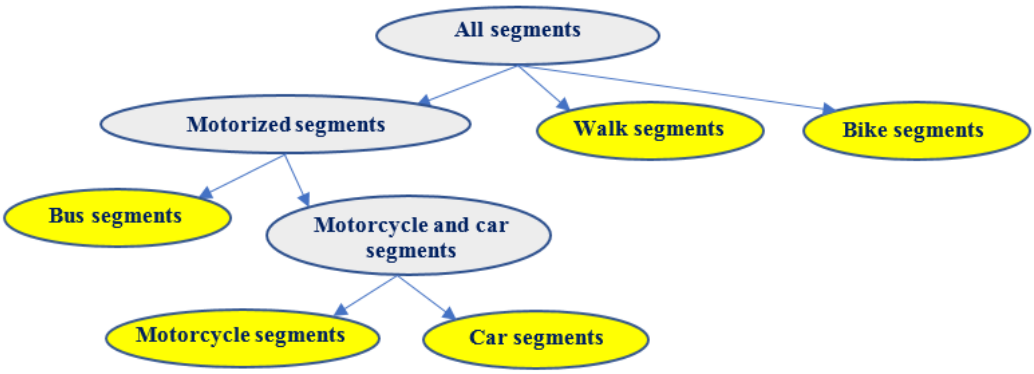


Figure 2*. Hierarchical mode detection process*

3.2.1 Step 1: Fuzzy logic model to classify walk, bicycle, and motorized segments

With fuzzy logic, speed and acceleration variables are generally sufficient for detecting walk, bicycle, and motorized segments (Rasmussen et al., 2015). However, the classification problem in Hanoi was more complex because motorcycle segments' acceleration and speed profiles were similar to those of walking and bike ones (see Figure 3). In particular, the overlapping of the 95th percentile of speed, median speed, and average acceleration between motorized and non-motorized modes was significant at low ranges (e.g., from 2 m/s to 8 m/s for the 95th percentile of speed); therefore, the heading change rate, which was the average change of heading per meter (Dabiri and Heaslip, 2018), was added. The heading between two points was calculated from their coordinates using equations presented by Dabiri and Heaslip (2018). As can be seen in Figure 3, the heading change rates were typically high for walking and typically low for motorized modes, which can be explained by the fact that motorized modes kept strictly to roads, but pedestrians did not always walk in a straight line from point to point. The heading change rates for bicycles were higher than those of motorized modes but lower than those of walking.

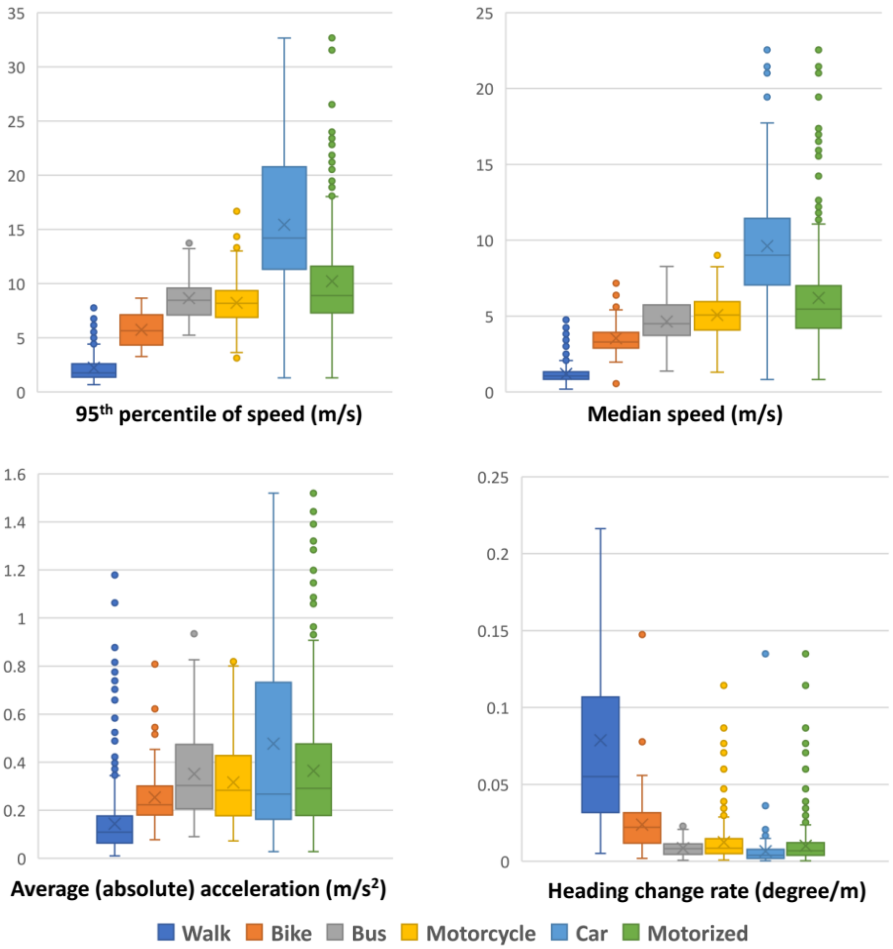


Figure 3. *Boxplots of variables by travel modes*

The operation of the fuzzy logic model was similar to that of Rasmussen et al. (2015) and Schuessler and Axhausen (2009). A detailed description can be found in Nguyen (2020). To limit the complexity of the rules in the fuzzy logic model, heading change rate was used as a supplement in case the speed and acceleration profiles between modes were ambiguous. On the basis of the trapezoidal membership functions (see Figure 4) and rules (see Table 2), each segment received three probabilities corresponding to walk, bicycle, and motorized modes. The mode with the highest probability was attributed to the segment.

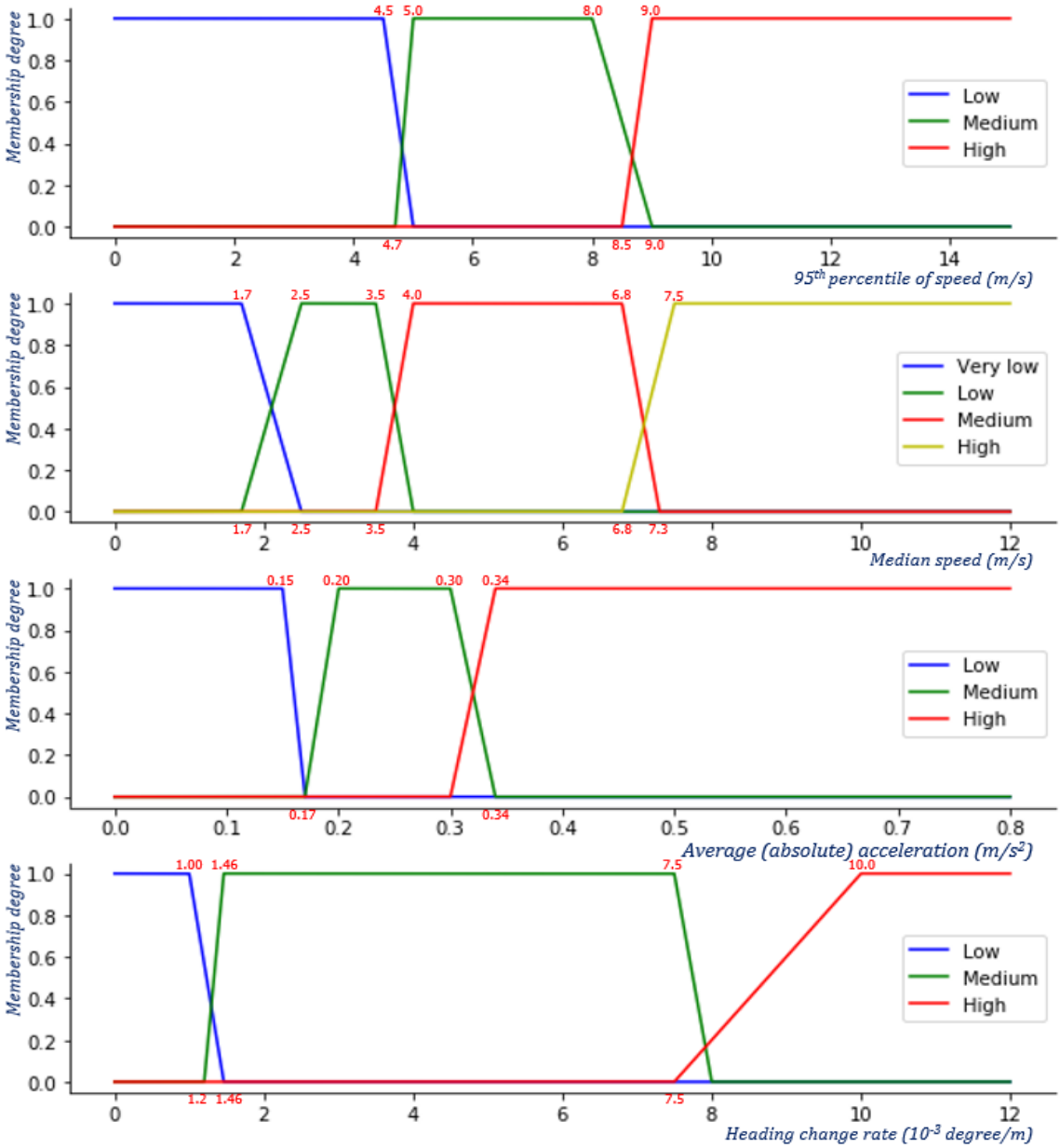


Figure 4. *Membership functions of the fuzzy logic model*

Table 2. *Rules of the fuzzy-logic algorithm*

Rule	95th percentile of speed	Median speed	Average (absolute) acceleration	Heading change rate	Mode
1	Low	Very low	-	-	WALK
2	Low	Low	Low	Low	MOTORIZED
3	Low	Low	Low	Medium	BICYCLE
4	Low	Low	Low	High	WALK
5	Low	Low	Medium	Low	BICYCLE
6	Low	Low	Medium	Medium	BICYCLE
7	Low	Low	Medium	High	WALK
8	Low	Low	High	Low	MOTORIZED
9	Low	Low	High	Medium	BICYCLE
10	Low	Low	High	High	WALK
11	Low	Medium	Low	-	BICYCLE
12	Low	Medium	Medium	-	MOTORIZED
13	Low	Medium	High	-	MOTORIZED
14	Low	High	Low	-	BICYCLE
15	Low	High	Medium	-	MOTORIZED
16	Low	High	High	-	MOTORIZED
17	Medium	Very low	Low	-	WALK
18	Medium	Very low	Medium	Low	BICYCLE
19	Medium	Very low	Medium	Medium	WALK
20	Medium	Very low	Medium	High	WALK
21	Medium	Very low	High	Low	MOTORIZED
22	Medium	Very low	High	Medium	WALK
23	Medium	Very low	High	High	WALK
24	Medium	Low	Low	Low	MOTORIZED
25	Medium	Low	Low	Medium	BICYCLE
26	Medium	Low	Low	High	WALK
27	Medium	Low	Medium	Low	MOTORIZED
28	Medium	Low	Medium	Medium	BICYCLE
29	Medium	Low	Medium	High	WALK
30	Medium	Low	High	Low	MOTORIZED
31	Medium	Low	High	Medium	BICYCLE
32	Medium	Low	High	High	MOTORIZED
33	Medium	Medium	Low	Low	MOTORIZED
34	Medium	Medium	Low	High	BICYCLE
35	Medium	Medium	Medium	Low	MOTORIZED
36	Medium	Medium	Medium	Medium	BICYCLE
37	Medium	Medium	Medium	High	MOTORIZED
38	Medium	Medium	High	-	MOTORIZED
39	Medium	High	Low	-	BICYCLE
40	Medium	High	Medium	-	MOTORIZED
41	Medium	High	High	-	MOTORIZED
42	High	-	-	-	MOTORIZED

3.2.2 Step 2: Rule-based bus detection

The outcomes of the first step were the walk, bicycle, and motorized segments. The second step analyzed motorized segments to detect bus segments.

Ambiguity between bus and car segments is a well-known challenge for mode detection. It can be addressed successfully with the use of GIS data along with actual or real-time operational data (Feng and Timmermans, 2019; Gong et al., 2012; Rasmussen et al., 2015; Semanjski et al., 2017; Stenneth et al., 2011; Nour et al., 2016; Marra et al., 2019). Bus detection in Hanoi was more complex than that in previous studies owing to four reasons as follows; (1) limited external data sources, (2) the distribution of bus stops, (3) the (fairly) frequent occurrence of bus bunching at peak times, and (4) the passing of bus stops where no boarding and alighting of passengers took place.

- *First*, the implementation of map-matching algorithms to evaluate the consistency between actual paths and bus routes (Rasmussen et al., 2015; Semanjski et al., 2017; Tsui and Shalaby, 2006) and the use of the buses' real-time or actual operational information (Marra et al., 2019; Stenneth et al., 2011) to enhance the detection were infeasible because only coordinates of stops were available.

- *Second*, to reduce the duration of bus trips, many bus stops are located near intersections so that the red-light phases can be used for boarding and alighting. However, this makes non-movement confusing as it can be owing to waiting for the traffic lights or to collecting passengers. Furthermore, bus stops are regularly distributed in front of points of interest, such as markets, universities, hospitals, and residential areas. Consequently, many non-bus segments also have either origins or destinations near bus stops.

- *Third*, over 70% of 120 bus routes provide connections to the central business districts, resulting in the overcrowding of buses there at peak hours. At a stop, there may be two, three, or even four buses in a row. Owing to bus bunching, boarding and alighting may not take place exactly at the stop, and a bus may pass the stop slowly. Figure 5 shows an example where three buses in a row allow passengers to get on and off at the same time. The boarding and alighting of Bus 3 are recorded quite far from the location of the bus stop. After boarding and alighting are completed, Bus 3 passes the bus stop slowly. Thus, there are no GPS points indicating that Bus 3 stopped at the bus stop.

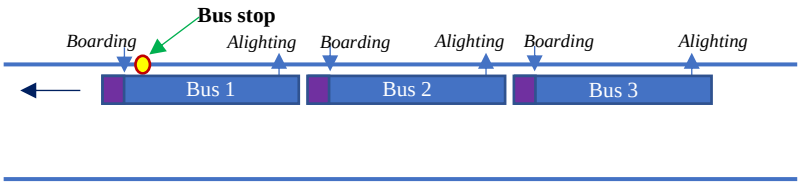


Figure 5. *Bus bunching with three vehicles at a station*

- *Fourth*, if there are no passengers waiting to board or alight, a bus tends to ignore a stop to save time and avoid blocking the traffic behind it. However, it is still able to stop immediately to satisfy any sudden requests.

Points two to four above show that it was unreliable to detect bus segments using the proximity of both the origin and the destination to bus stops. Points three and four showed that the stationary state of buses at stops would not be sufficient to identify the vast majority of bus segments.

The authors therefore decided to extend the approach introduced by Nour et al. (2016), which is based on the rate of stops adjacent to transit stations. First, the inverse of the average stop rate (i.e., the average distance between stops) was employed. Second, unlike Nour et al. (2016), to avoid the loss of the opportunity to detect bus segments, particularly short ones, bus stops in the proximity of intersections were retained. Third, stopping at every bus stop and passing of stops slowly were considered. Fourth, Nour et al. (2016) determined the threshold of the stop rate to detect bus segments

in their data; thus, the threshold may be valid for their sample only. In this study, with stopping at and slow movement past bus stops taken into consideration, almost all stops on a segment that a bus had moved on were detected and included. The average distance between bus stops of a bus segment should be compatible with a threshold of average distance between stops on the citywide network.

The proposed bus detection method had three stages (see Figure 6):

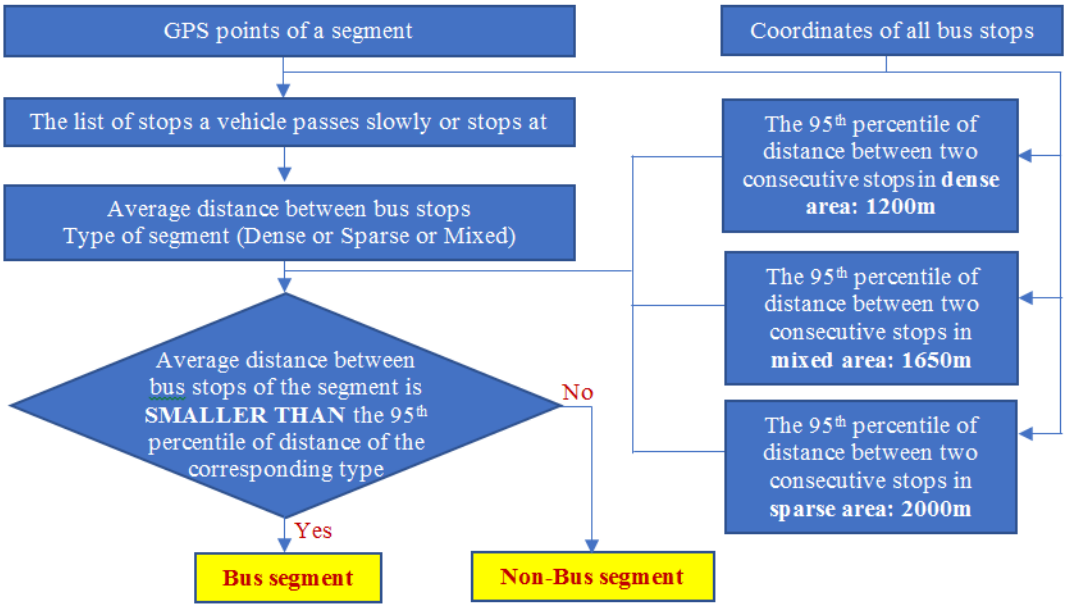


Figure 6. *Flowchart of bus-segment detection*

** Stage 1: Finding the list of stops a vehicle passed slowly or stopped at*

First, the radius from a bus stop in which to search for slow movement or stopping was defined as 100 m. To estimate the speed of slow movement, 15 bus segments were randomly selected. Each of them was plotted on a map so that all the stops the bus should stop at could be known. The speed threshold of slow movement was the median value of instantaneous speed levels of all points that were within the 100 m buffer from each bus stop. In this way, the threshold of 2.5 m/s was determined.

For each segment, to count the number of stops a vehicle passed slowly or stopped at, the distance to the nearest bus stop of each point whose instantaneous speed was under 2.5 m/s was noted. If the distance was smaller than 100 m, the corresponding bus stop was retained. The result of searching for all points of a segment was a list of bus stop candidates. In the list, duplicates were deleted. Any stop, whose distance to its previous stop in the list was under 350 m (i.e., the minimum distance between two stops on a route in the bus network), was eliminated.

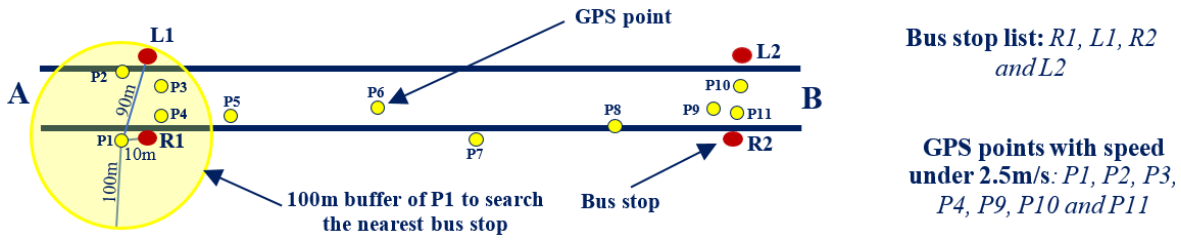


Figure 7. *An example of searching for stops a bus passed slowly or stopped at*

Figure 7 shows an example in which a person on a bus that moves from A to B and either stops at or passes slowly the two stops R1 and R2. In the list of potential stops the following criteria are applied;

a. R1 is added twice because GPS points P1 and P4 have instantaneous speeds under 2.5 m/s, and the distances to R1 are the smallest (under 100 m);
b. L1 is added twice (i.e., corresponding to P2 and P3);
c. L2 is added once (i.e., corresponding to P10); and
d. R2 is added twice (i.e., corresponding to P9 and P11.

Therefore, to count the number of bus stops on the A–B segment;

a. All duplicates of L1, R1, L2, and R2 are eliminated;
b. L1 is removed because its distance to R1 is under 350 m; and
c. L2 is removed because its distance to R1 is under 350 m.

As a result of the above, the final list comprises R1 and R2.

** Stage 2: Calculating the average distance between stops*

The average distance between bus stops for a segment was calculated using Equation 1 below. Hereafter, "bus stop" refers to those designated boarding and alighting points that a bus passed slowly or stopped at. The length of the segment was the sum of the distances between consecutive points that belong to the segment.

$$Average\ distance\ between\ bus\ stops = \frac{Length\ of\ segment}{Number\ of\ bus\ stops - 1} \quad (1)$$

** Stage 3: Determining whether a segment is a bus segment*

- Estimating the threshold of distance between two consecutive stops on the bus network

The task was to seek a threshold of average distance between stops on the whole bus network to determine whether a segment was travelled by bus. All distances between two consecutive stops on the same bus route and the same direction were noted.

For example, in Figure 7, the distances between L1 and L2 along with R1 and R2 were valid, whereas the distances of L1–R1, L2–R2, L1–R2, and L2–R1 were disregarded. To cover most cases, the 95th percentile value that is higher than 95% of the other distances was chosen. Because of the different distributions of bus stops in different areas, the 95th percentile of the distance between two consecutive stops varies. As indicated in section 3.1, Hanoi comprises two main areas. The first encompasses central business districts with a dense distribution of bus stops. In the other area, the stop distribution is sparser. To calculate the 95th percentile distance in the dense area, all distances between two consecutive stops of the same bus route were examined. The 95th percentile distance in the sparse area was calculated in the same way. The 95th percentile distance of mixed area (i.e., both areas together) was calculated by considering stops on the entire bus network. The values of the 95th percentile of distance between two consecutive stops in the dense area, mixed area, and sparse area were 1,200 m, 1,650 m, and 2,000 m, respectively.

- Labeling segments

On the basis of the list of stops, segments can be divided into three types: dense segments, sparse segments, or mixed segments (see Figure 8).

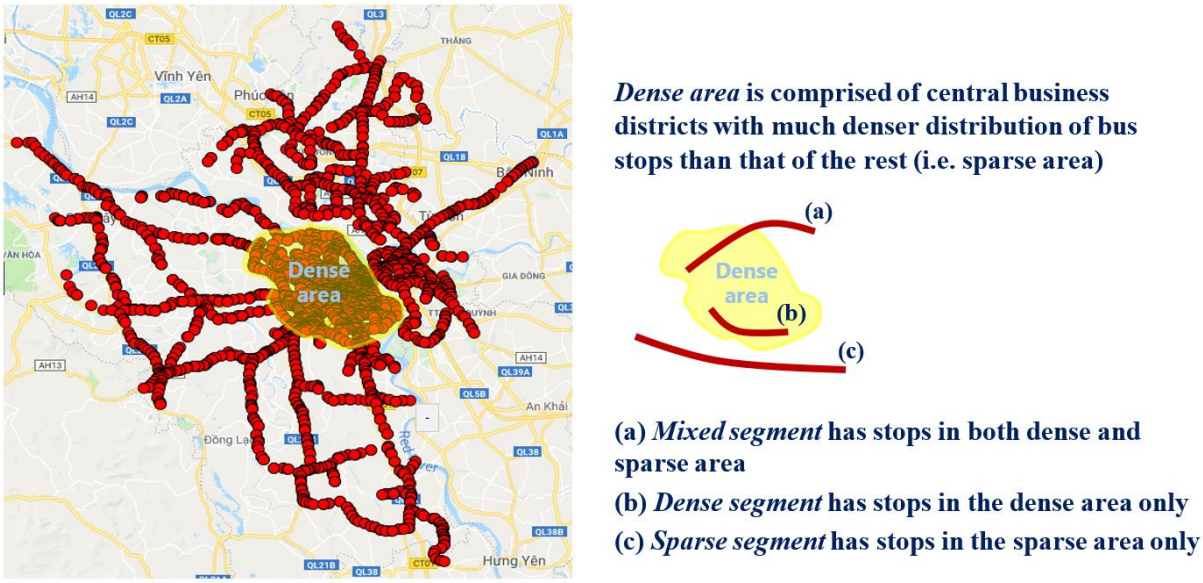


Figure 8. *Distribution of bus stops in Hanoi and classification of segments*

A dense segment was associated as a bus segment if the average distance between its stops (calculated by Equation 1) was smaller than the 95th percentile of the distance between two consecutive stops in the dense area (i.e., 1,200 m). If this condition was not met, it was associated as a non-bus segment. Similarly, sparse segments and mixed segments by bus were determined in the same way.

3.2.3 Step 3: Random forest algorithm to classify car and motorcycle

Confusion between the car and motorcycle modes was generally significant. There was lesser confusion where the car ran at high speed or accelerated/decelerated at a high rate (see Figure 3); however, in most situations, they showed very similar speed and acceleration profiles owing to moving in urban areas. To distinguish between them, the model must use many variables, which was not suitable for deterministic and probabilistic methods. Car—and especially motorcycle—segments accounted for significant percentages of the sample (see Figure 1), so a learning-based mode was developed.

A random forest algorithm is a standard and widely used non-parametric prediction tool introduced by Breiman (2001). It contains numerous decision trees and operates based on randomness. All the trees learn from samples selected randomly from the original data with replacement (see Figure 9). To split each node of a decision tree, a subset of randomly selected input features is used (Statnikov et al., 2008). In a classification problem, the decision trees' votes are aggregated to form the final prediction decision. By means of the randomness and the voting mechanism, random forests avoid overfitting and generate satisfactory prediction results.

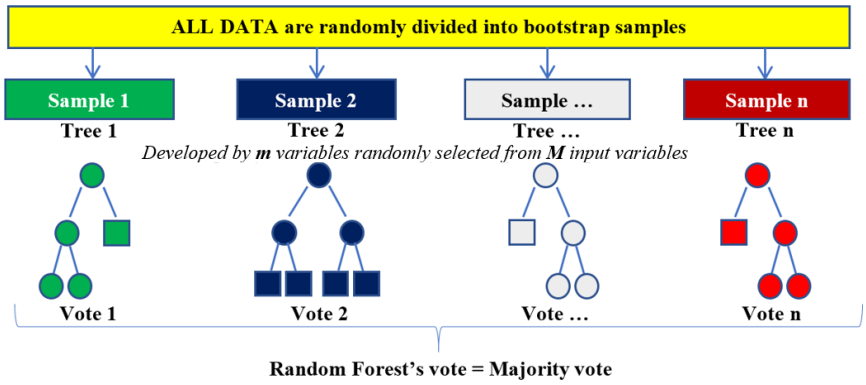


Figure 9. *A random forest structure*

Random forest models have been documented in a number of mode detection studies (Gong et al., 2018; Marra et al., 2019; Stenneth et al., 2011). In this study, a random forest model was implemented using Python and the scikit-learn library. In addition to the 95th percentile of speed, median speed, and average absolute acceleration, variables including the 95th percentile of absolute acceleration and segment length were used as inputs to the model. To train the model, 75% of the 1,245 motorcycle segments and 75% of the 587 car segments were used. Thus, the data used to evaluate the hierarchical process comprised 758 walk segments, 104 bicycle segments, 97 bus segments, 311 motorcycle segments, and 147 car segments.

4 RESULTS AND DISCUSSION

4.1 Hierarchical mode detection process

The evaluation of the mode imputation was based on four metrics: precision, recall, F-score, and accuracy (see Equations 2, 3, 4, 5). The closer to the value 1 the metrics are, the better the classifier performs.

$$precision = \frac{the\ number\ of\ segments\ correctly\ classified\ as\ mode\ i}{the\ number\ of\ segments\ classified\ as\ mode\ i} \quad (2)$$

$$recall = \frac{the\ number\ of\ segments\ correctly\ classified\ as\ mode\ i}{the\ number\ of\ segments\ actually\ being\ mode\ i} \quad (3)$$

$$F - score = \frac{2 * precision * recall}{precision + recall} \quad (4)$$

$$accuracy = \frac{the\ number\ of\ segments\ correctly\ classified}{the\ total\ number\ of\ segments} \quad (5)$$

In terms of accuracy, 89.1% of all segments were correctly inferred (see Table 3). This accuracy level is comparable to that of many previous studies using either all-in-one or hierarchical processes (Dabiri and Heaslip, 2018; Feng and Timmermans, 2019; Marra et al., 2019; Nour et al., 2016; Rasmussen et al., 2015; Stenneth et al., 2011; Tsui and Shalaby, 2006; Xiao et al., 2015).

Regarding Step 1 (see sub-section 3.2.1), the fuzzy logic algorithm functioned well, resulting in excellent classification of the walk and motorized modes. The classification of walk segments obtained the highest F-score at 96.3%, and only 18 out of 555 motorized segments were incorrectly identified. The detection of bicycle segments was not as accurate as that of other segments, with both precision and recall levels of approximately 75%.

Table 3. *Confusion matrix of mode detection results*

		Detected						***Recall***	***F-score***
		Walk	Bicycle	Bus	Motorcycle	Car	*Total*		
Reported	Walk	717	18	3	20	0	*758*	***94.6%***	***96.3%***
	Bicycle	4	79	2	19	0	*104*	***76.0%***	***75.6%***
	Bus	0	0	92	4	1	*97*	***94.8%***	***87.2%***
	Motorcycle	7	8	12	271	13	*311*	***87.1%***	***82.1%***
	Car	3	0	5	35	104	*147*	***70.7%***	***78.5%***
	Total	*731*	*105*	*114*	*349*	*118*	*1417*	-	-
Precision		***98.1%***	***75.2%***	***80.7%***	***77.7%***	***88.1%***	-	Accuracy: **89.1%**	

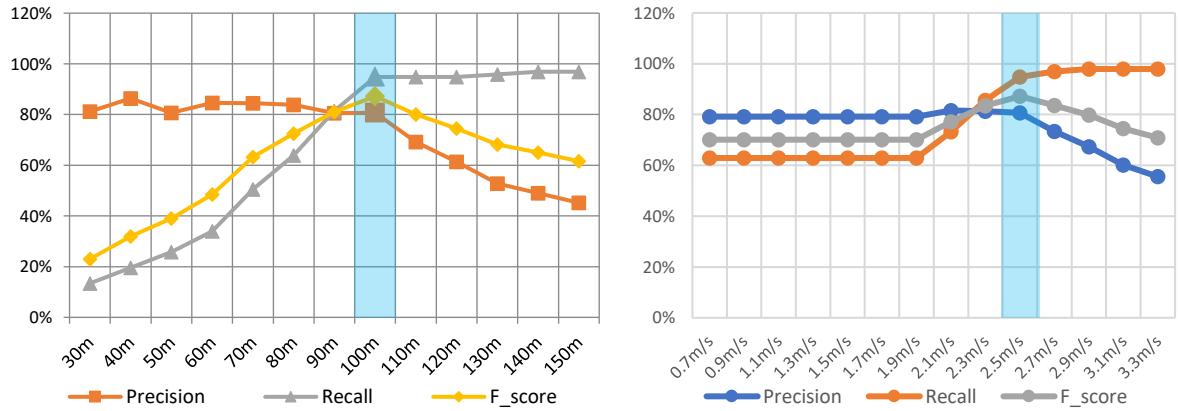
For bus detection, the high recall level of 94.8% showed that bus-detection rules identified almost all of the actual bus segments (92 of 97 cases). However, a number of other segments were falsely detected as bus segments—possibly owing to taking passing of bus stops slowly into consideration—resulting in a precision level of 80.7% corresponding to the bus. Despite this limitation, however, these precision and recall levels were comparable with those of studies using real-time information or high-quality GIS data (Feng and Timmermans, 2019; Gong et al., 2018; Marra et al., 2019; Rasmussen et al., 2015; Semanjski et al., 2017; Tsui and Shalaby, 2006). In comparison with the model of Nour et al. (2016), the bus-detection method of this study achieved a much higher recall (94.8% vs. 84.7%) and thereby a larger F-score (87% vs. 84%).

Bus identification was greatly affected by both the threshold of distance to search for the nearest bus stop and the threshold of speed representing the slow movement. The sensitivity of the inferences when (1) fixing the speed at 2.5 m/s coupled with changing the distance, and (2) fixing the distance at 100 m coupled with changing the speed were tested, respectively.

In the case of a 2.5 m/s speed threshold (Figure 10a), at the smallest level of 30 m, the ability to detect bus segments was unsatisfactory, with a recall of only 18% yielded. Between 30 m and 100 m, the higher the distance was, the higher the recall was. This emphasized that buses would not stop exactly at bus stops. In this range, the precision values nearly levelled out. Between 110 m and 150 m, the increase in recall levels was insignificant, and the precision decreased dramatically with distance. In fact, at 100 m, the recall reached the near-maximum level of approximately 95%, with 92 of the 97 bus segments being correctly detected. The changes in precision and recall were reflected by changes in the F-score. The highest F-score was achieved at 100 m, from which we concluded that 100 m was the best distance threshold when the speed was 2.5 m/s. The 100 m distance is suitable for the 2.5 m/s speed possibly because a bus tended to slow down before each bus stop.

In the case of a 100 m distance threshold, Figure 10b reveals that the consideration of stopping (i.e., very low speeds, such as 0.7 m/s) did not lead to accurate detection of bus segments, with a recall of 63% yielded. As the speed increased from 2 m/s to 2.7m/s, more bus segments were successfully recognized, albeit with nearly unchanged precision. From 2.9 m/s, precision dropped once recall was nearly stable, leading to a decrease in the F-score. The F-score reached its maximum value of 87.2% for a speed of 2.5 m/s.

These analyses have demonstrated that it was appropriate to consider stopping at and passing bus stops slowly together with the thresholds chosen (2.5 m/s and 100 m).



(a) The speed fixes of 2.5 m/s *(b) The distance fixes of 100 m*

Figure 10. *Sensitivity of bus detection to changes in distance and speed thresholds*

Motorcycles were shown to be the primary source of misclassification. Of 118 segments labeled as cars, 13 were actually motorcycle. Apart from walk, motorcycle was the only mode misclassified as bicycle, (8 segments). Nineteen bicycle segments (18% of the total bicycle segments) were falsely

classified as motorcycle segments. The ambiguity of the motorcycle mode with other modes reduced the correct number of inferences, which was reflected in its precision level of only 77.7%, about 10% lower than its recall level of 87.1%.

Car segments were also confused with motorcycle segments. Of the 147 car segments, 35 were labeled as motorcycle segments, resulting in a recall of 70.7%, the lowest value for all the modes. However, the confusion between car and bus, despite being a well-known issue in literature, was minor.

4.2 Comparing the proposed hierarchical process with other processes

A simple hierarchical process and an all-in-one process (see Table 4) were developed to assess the performance of the proposed hierarchical process. The former used rules related to speed and distance to the nearest bus stops. The latter was based on a random forest model.

The accuracy level of the simple hierarchical process was low at 61.3%, compared with 79.1% for the all-in-one process and 89.1% for the process proposed in this study. The poor performance of the rule-based process was anticipated because such a process was too simple to deal with the challenge of classifying five modes.

The all-in-one process had very low F-score values for bus and bicycle segments because it had a significant bias against modes having minor percentages of the dataset. In other words, the all-in-one process failed to deal with the imbalanced data problem. In contrast, the process proposed in this work generated comparable F-score values for all modes, and these values were higher than those of the random forest model.

The F-score values of bus segments for the rule-based process (0.17) and the random forest-based process (0.16) were much lower than that of the proposed process (0.87), which highlighted how much bus detection could be improved by considering the average distance between bus stops that a bus passed slowly or stopped at. Furthermore, the rule-based and random forest processes did not detect cars and motorcycles as well as the hierarchical process proposed here did, which emphasized the confusion between motorcycles and cars.

Table 4. *Simple hierarchical process and all-in-one process*

Process	**Description**				**F-score**	**Accuracy**
RULE-BASED (SIMPLE HIERARCHICAL)	*95th percentile of speed*	*Median speed*	*Proximity to bus stops*	*Mode*	*Foot: 0.89* *Bicycle: 0.27* *Bus: 0.17* *Motorcycle: 0.62* *Car: 0.64*	*61.3%*
Step 1	< 3.5	< 2.0	-	Foot		
Step 2	< 6.0	< 4.0	-	Bicycle		
Step 3	< 15.0	≥ 3.5	Yes	Bus		
Step 4	> 12.0	≥ 6.0	-	Car		
Step 5	The remainder of segments			Motorcycle		
RANDOM FORESTS (ALL-IN-ONE)	*Features:* 95th percentile of speed, median speed, proximity to bus stops (0 if no and 1 if yes), heading change rate, low speed rate, 95th percentile of acceleration, average (absolute) acceleration. *Splitting data*: at the rate of 75% vs. 25%				*Foot: 0.93* *Bicycle: 0.25* *Bus: 0.16* *Motorcycle: 0.80* *Car: 0.69*	*79.1%*
Proximity to bus stops refers to the distances from both the origin and the destination of a segment to the nearest stops within 75 m						

5 CONCLUSIONS

Making transportation mode inferences has been a common goal of GPS-data-based research owing to the absence of trip characteristics in logs. This study has addressed a difficult challenge, with the inclusion of motorcycles, a major travel mode, in data collected in Hanoi.

First, a hierarchical process was developed to classify walk, bicycle, and motorized modes using a fuzzy logic algorithm. In addition to acceleration and speed specific variables, heading change

rate was used to enhance the classifier's power. Bus segments were then distinguished from other motorized segments upon extension of the work of Nour et al. (2016). Specifically, an average distance between stops at which the bus passed slowly or stopped at was compared with those estimated from the bus network. The advantages of this method are that it could detect almost all of the bus segments by the coordinates of stops only, and this was easily understandable because it originated from the actual operation of bus services. To limit other modes being misclassified as bus, it was necessary to carefully choose thresholds of speed and distance. The distance should be determined before the speed is estimated from the sample. Finally, a random forest model was developed to detect motorcycle and car segments.

The proposed hierarchical process performed well, with an accuracy of 89.1%. The main source of confusion was the mode of motorcycle. The most frequent ambiguities were between motorcycles and cars, not between cars and buses. This is typical not only for Hanoi but also for cities in a number of developing countries where travel depends heavily on two-wheeled motorized vehicles. This study was an effort to extend the list of modes and the geographical scope of research into imputing travel modes from GPS data.

Although thoroughly developed, the process was validated only on a sample that was biased toward persons working and studying at a university in Hanoi. Thus, the inferences would, to some extent, benefit from the homogeneity of travel patterns of the participants. Moreover, the Hanoi urban transport system does not include any metro lines at present. These limitations emphasize the need to adapt and test the process using a more diverse sample with more travel options. Future research could be conducted to enhance identification of the motorcycle mode.

Appendix 1. Synthesis of mode detection studies

Authors and studies	Modes	Methods	Variables	Overall Accuracy	Process types
Tsui and Shalaby, (2006)[T]	Walk, cycle, bus, auto, streetcar, subway, off-road	Fuzzy-logic and map-matching	Speed, acceleration, data quality, spatial information	94%	Complex hierarchical
Stopher et al., (2008)[T]	Walk, bicycle, car, bus, tram	Probability matrix	Speed, spatial information	95%	Simple hierarchical
Bohte and Maat, (2009)[E]	Car, train, bicycle, foot, other	Rule-based	Speed, spatial information	70%	Simple hierarchical
Gong et al., (2012)[T]	Walk, subway, rail, car, bus	Rule-based	Speed, acceleration, spatial information	82.6%	Simple hierarchical
Rasmussen et al., (2015)[E]	Walk, bicycle, bus, car, rail, other	Fuzzy-logic and map-matching	Acceleration, speed, spatial information	92.4%	Complex hierarchical
Nour et al., (2016)[T]	Walk, bicycle, transit, auto	KNN and rule-based	Speed, acceleration, jerk, spatial information (i.e. transit stop rate)	92.5% (1)	Complex hierarchical
Marra et al., (2019)[E]	Walk, bus/tram, train, car, bicycle	Rule-based, probability-based and RF	Speed, acceleration, heading, actual operational data of public transport, historical travel data	86.1%(2) and 87%(3)	Complex hierarchical
Schuessler and Axhausen, (2009)[E]	Walk, cycle, car, urban public transport, train	Fuzzy-logic	Acceleration, speed	Not available	All-in-one
Stenneth et al., (2011)[T]	Train, bus, walk, car, bicycle, stationary	**RF**, NB, BN, DT, MLP	Acceleration, speed, spatial information and real-time data	92.8% and 92.9% (4)	All-in-one
Shafique and Hato, (2015)[E]	Walk, bicycle, car, train	**RF**, SVM, AdaBoost, DT	Acceleration values for 3 directions	99.8%	All-in-one
Xiao et al., (2015)[E]	Walk, bicycle, E-bike, bus, car	**BN**, SVM, MNL, ANN	Speed, acceleration, average heading change, distance	92.7%	All-in-one
Feng and Timmermans, (2016)[E]	Walk, bicycle, bus, car, motorbike, running, tram, metro, train, activity	**BN**, NB, LR, MP, DT, SVM, C4.5	Speed, acceleration, distance, data quality, spatial information	99.8%	All-in-one
Semanjski et al., (2017)[E]	Walk, bus, car, foot, train	SVM	Spatial information	94%	All-in-one
Dabiri and Heaslip, (2018)[E]	Walk, bicycle, bus, driving, train	**CNN**, KNN, SVM, DT, RF, MLP	Speed, acceleration, heading change rate, jerk	84.8%	All-in-one
Gong et al., (2018)[E]	Walk, bus, tram, auto	RF	Travel time, length, speed, participant's information, spatial information	Not available	All-in-one
Feng and Timmermans, (2019)[E]	Car, bus, bicycle, walk, train	BN	Speed, acceleration, distance, data quality, spatial information	88.1%	All-in-one

(1) Estimated based on the confusion matrix in (Nour et al., 2016).
(2) For classifying walk, bus/tram, train and private modes (i.e. car and bicycle) in Basel data;
(3) For classifying bicycle and car in Basel data;
(4) Refer to overall precision and overall recall, respectively.
RF: Random Forests; DT: Decision Tree; SVM: Support Vector Machine; KNN: K-Nearest Neighbors; MLP: Multilayer Perceptron, CNN: Convolutional Neural Network, LR: Linear Regression, NB: Naïve Bayes, BN: Bayesian Network, ANN: Artificial Neural Network.
In a study reporting a number of methods, the main method is in bold and the accuracy presented in the table is its.
[E] Refers to an experiment whose valid data are comprised of either at least 50 persons or at least 350 days (equivalent to 50 persons * 7 days) or at least 1400 trips (equivalent to 350 days * 4 trips/day).
[T] Refers to a test at small scale and thus fails to meet the experiment-specific criteria.

Appendix 2. Prediction results of the hierarchical process at fuzzy logic step and bus detection step

Prediction result at fuzzy logic step

		Predicted			Total	Recall	F-score
		Walk	*Bicycle*	*Motorized*			
Reported	*Walk*	717	18	23	*758*	*94.6%*	*96.3%*
	Bicycle	4	79	21	*104*	*76.0%*	*75.6%*
	Motorized	10	8	537	*555*	*96.8%*	*94.5%*
Total		*731*	*105*	*581*	*1417*	-	-
Precision		*98.1%*	*75.2%*	*92.4%*	-	**Accuracy**	**94.1%**

Prediction result at bus detection step

		Predicted				Total	Recall	F-score
		Walk	*Bicycle*	*Bus*	*Car/Motorcycle*			
Reported	*Walk*	717	18	3	20	*758*	*94.6%*	*96.3%*
	Bicycle	4	79	2	19	*104*	*76.0%*	*75.6%*
	Bus	0	0	92	5	*97*	*94.8%*	*87.2%*
	Car/Motorcycle	10	8	17	423	*458*	*92.4%*	*91.5%*
Total		*731*	*105*	*114*	*467*	*1417*	-	-
Precision		*98.1%*	*75.2%*	*80.7%*	*90.6%*	-	**Accuracy**	**92.5%**

REFERENCES

Armoogum, J., Bonsall, P., Browne, M., Christensen, L., Cools, M., Cornelis, E., Diana, M., Guilloux, T., Harder, H., Hegner Reinau, K., Hubert, J.-P., Kagerbauer, M., Kuhnimhof, T., Madre, J.-L., Moiseeva, A., Polak, J., Schulz, A., Tébar, M., Vidalakis, L., 2014. Survey Harmonisation with New Technologies Improvement (SHANTI). IFSTTAR.

Bohte, W., Maat, K., 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. Transp. Res. Part C Emerg. Technol. 17, 285–297. https://doi.org/10.1016/j.trc.2008.11.004

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Burkhard, O., Becker, H., Weibel, R., Axhausen, K.W., 2020. On the requirements on spatial accuracy and sampling rate for transport mode detection in view of a shift to passive signalling data. Transportation Research Part C: Emerging Technologies 114, 99–117. https://doi.org/10.1016/j.trc.2020.01.021

Bzdok, D., Altman, N., Krzywinski, M., 2018. Statistics versus machine learning. Nat. Methods 15, 233–234. https://doi.org/10.1038/nmeth.4642

Chen, C., Gong, H., Lawson, C., Bialostozky, E., 2010. 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. Part Policy Pract. 44, 830–840. https://doi.org/10.1016/j.tra.2010.08.004

Dabiri, S., Heaslip, K., 2018. Inferring transportation modes from GPS trajectories using a convolutional neural network. Transp. Res. Part C Emerg. Technol. 86, 360–371. https://doi.org/10.1016/j.trc.2017.11.021

Feng, T., Timmermans, H.J.P., 2019. Integrated imputation of activity-travel diaries incorporating the measurement of uncertainty. Transp. Plan. Technol. 42, 274–292. https://doi.org/10.1080/03081060.2019.1576384

Feng, T., Timmermans, H.J.P., 2016. Comparison of advanced imputation algorithms for detection of transportation mode and activity episode using GPS data. Transp. Plan. Technol. 39, 180–194. https://doi.org/10.1080/03081060.2015.1127540

Forrest, T., Pearson, D., 2005. Comparison of Trip Determination Methods in Household Travel Surveys Enhanced by a Global Positioning System. Transp. Res. Rec. J. Transp. Res. Board 1917, 63–71. https://doi.org/10.3141/1917-08

Gong, H., Chen, C., Bialostozky, E., Lawson, C.T., 2012. A GPS/GIS method for travel mode detection in New York City. Comput. Environ. Urban Syst., Special Issue: Geoinformatics 2010 36, 131–139. https://doi.org/10.1016/j.compenvurbsys.2011.05.003

Gong, L., Kanamori, R., Yamamoto, T., 2018. Data selection in machine learning for identifying trip purposes and travel modes from longitudinal GPS data collection lasting for seasons. Travel Behav. Soc. https://doi.org/10.1016/j.tbs.2017.03.004

Gong, L., Morikawa, T., Yamamoto, T., Sato, H., 2014. Deriving Personal Trip Data from GPS Data: A Literature Review on the Existing Methodologies. Procedia - Soc. Behav. Sci., The 9th International Conference on Traffic and Transportation Studies (ICTTS 2014) 138, 557–565. https://doi.org/10.1016/j.sbspro.2014.07.239

Huynh, D., 2020. Making Megacities in Asia: Comparing National Economic Development Trajectories, SpringerBriefs in Regional Science. Springer Singapore, Singapore. https://doi.org/10.1007/978-981-15-0660-4

Marra, A.D., Becker, H., Axhausen, K.W., Corman, F., 2019. Developing a passive GPS tracking system to study long-term travel behavior. Transp. Res. Part C Emerg. Technol. 104, 348–368. https://doi.org/10.1016/j.trc.2019.05.006

Nguyen, M.H., Armoogum, J., Madre, J-L., Garcia, C., forthcoming. Reviewing Trip Purpose Imputation in GPS-based Travel Surveys. Journal of Traffic and Transportation Engineering (English Edition).

Nguyen, M.H., 2020. Imputations des modes et des motifs de transport pour les enquêtes “GPS” Université Paris-Est, France. https://www.dest.ifsttar.fr/linstitut/ame/laboratoires/dest-ifsttar/formation-a-la-recherche/details-theses/?nom=NGUYEN&prenom=Minh%20Hieu

Nguyen, M.H., Armoogum, J., Garcia, C., 2019a. Experiment on mobility survey using smartphone in Hanoi, Vietnam. Presented at the Transportation for A Better Life: Smart Mobility for Now and Then, Bangkok, Thailand.

Nguyen, M.H., Ha, T.T., Tu, S.S., Nguyen, T.C., 2019b. Impediments to the bus rapid transit implementation in developing countries – a typical evidence from Hanoi. Int. J. Urban Sci. 23(4), 464–483. https://doi.org/10.1080/12265934.2019.1577747

Nguyen, M.H., Pojani, D., 2018. Chapter Two - Why Do Some BRT Systems in the Global South Fail to Perform Or Expand?, in: Shiftan, Y., Kamargianni, M. (Eds.), Preparing for the New Era of Transport Policies: Learning from Experience, Advances in Transport Policy and Planning. Elsevier Academic Press, pp. 35–61. https://doi.org/10.1016/bs.atpp.2018.07.005

Nour, A., Hellinga, B., Casello, J., 2016. Classification of automobile and transit trips from Smartphone data: Enhancing accuracy using spatial statistics and GIS. J. Transp. Geogr. 51, 36–44. https://doi.org/10.1016/j.jtrangeo.2015.11.005

Rasmussen, T.K., Ingvardson, J.B., Halldórsdóttir, K., Nielsen, O.A., 2015. Improved methods to deduct trip legs and mode from travel surveys using wearable GPS devices: A case study from the Greater Copenhagen area. Comput. Environ. Urban Syst. 54, 301–313. https://doi.org/10.1016/j.compenvurbsys.2015.04.001

Schuessler, N., Axhausen, K., 2009. Processing Raw Data from Global Positioning Systems Without Additional Information. Transp. Res. Rec. J. Transp. Res. Board 2105, 28–36. https://doi.org/10.3141/2105-04

Semanjski, I., Gautama, S., Ahas, R., Witlox, F., 2017. Spatial context mining approach for transport mode recognition from mobile sensed big data. Comput. Environ. Urban Syst. 66, 38–52. https://doi.org/10.1016/j.compenvurbsys.2017.07.004

Shafique, M.A., Hato, E., 2015. Use of acceleration data for transportation mode prediction. Transportation 42, 163–188. https://doi.org/10.1007/s11116-014-9541-6

Statnikov, A., Wang, L., Aliferis, C.F., 2008. A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. BMC Bioinformatics 9, 319. https://doi.org/10.1186/1471-2105-9-319

Stenneth, L., Wolfson, O., Yu, P.S., Xu, B., 2011. Transportation mode detection using mobile phones and GIS information, in: Proceedings of the 19th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems - GIS '11. Presented at the 19th ACM SIGSPATIAL International Conference, ACM Press, Chicago, Illinois, p. 54. https://doi.org/10.1145/2093973.2093982

Stopher, P., FitzGerald, C., Zhang, J., 2008. Search for a global positioning system device to measure person travel. Transp. Res. Part C Emerg. Technol., Emerging Commercial Technologies 16, 350–369. https://doi.org/10.1016/j.trc.2007.10.002

Thomas, T., Geurs, K.T., Koolwaaij, J., Bijlsma, M., 2018. Automatic Trip Detection with the Dutch Mobile Mobility Panel: Towards Reliable Multiple-Week Trip Registration for Large Samples. J. Urban Technol. 25, 143–161. https://doi.org/10.1080/10630732.2018.1471874

Tsui, S., Shalaby, A., 2006. Enhanced System for Link and Mode Identification for Personal Travel Surveys Based on Global Positioning Systems. Transp. Res. Rec. J. Transp. Res. Board 1972, 38–45. https://doi.org/10.3141/1972-07

Vincenty, T., 1975. Direct and Inverse Solutions of Geodesics on the Ellipsoid with Application of Nested Equations. Surv. Rev. 23, 88–93. https://doi.org/10.1179/sre.1975.23.176.88

Wang, B., Gao, L., Juan, Z., 2017. A trip detection model for individual smartphone-based GPS records with a novel evaluation method. Advances in Mechanical Engineering 9, 168781401770506. https://doi.org/10.1177/1687814017705066

Wolf, J., Oliveira, M., Thompson, M., 2003. Impact of Underreporting on Mileage and Travel Time Estimates: Results from Global Positioning System-Enhanced Household Travel Survey. Transp. Res. Rec. J. Transp. Res. Board 1854, 189–198. https://doi.org/10.3141/1854-21

Xiao, G., Juan, Z., Zhang, C., 2015. Travel mode detection based on GPS track data and Bayesian networks. Comput. Environ. Urban Syst. 54, 14–22. https://doi.org/10.1016/j.compenvurbsys.2015.05.005

Acknowledgements

The authors highly appreciate (1) constructive comments of anonymous reviewers and the editor, (2) volunteer support of participants in the Hanoi survey and (3) useful discussion with Dr. Jean-Loup Madre, a senior researcher at IFSTTAR/AME/DEST.

Authors' contributions

The authors confirm contribution to the paper as follows: study conception and design: MHN and JA; data collection: MHN; analysis and interpretation of results: MHN; draft manuscript preparation: MHN and JA. All authors reviewed the results and approved the final version of the manuscript.

Conflicts of interest

The authors declare that there is no conflict of interest regarding the publication of this paper.