Journal of Transport Geography 78 (2019) 70–86

Contents lists available at ScienceDirect

Journal of Transport Geography

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

Determination of the influence factors on household vehicle ownership patterns in Phnom Penh using statistical and machine learning methods

Tran Vinh Ha[a,b], Takumi Asada[a], Mikiharu Arimura[a,*]

[a] Division of Sustainable and Environmental Engineering, Muroran Institute of Technology, T 050-8585, 27-1 Mizumoto-cho, Muroran, Hokkaido, Japan
[b] Faculty of Urban Environmental and Infrastructural Engineering, Hanoi Architectural University, km 10 Nguyen Trai, Thanh Xuan district, Hanoi, Viet Nam

ARTICLE INFO

Keywords:
Vehicle ownership
Phnom Penh
Features ranking
Multinomial logit model
Neural networks
Random forests

ABSTRACT

Vehicle ownership patterns and their determinants play an important role in transportation policy-making. This issue has been paid even greater attention in developing countries that aspire to reach sustainable transportation development goals in the era of urbanization and globalization. In this study, the multinomial logit model, neural networks and random forest were applied to examine the features' impact level and to also predict vehicle ownership patterns in Phnom Penh city. The empirical results indicate that household income is the most powerful variable affecting motorization in Phnom Penh. Supplementation of individual trip characteristics such as total number of trips made, number of trips made for work purposes and overall travel distance all make effective contributions as classifiers. Furthermore, it is acknowledged that the machine-learning approach outperformed not only in terms of predicting accuracy, but also in dealing with unbalanced categories when compared with the statistical approach. This detection supplies the advantages of applying machine learning techniques in terms of, but not limited to, the field of vehicle ownership.

1. Introduction

One of the goals of sustainable transportation development is to control the boom in ownership and usage of private motorized vehicles, which lies at the core of a variety of problematic issues facing urban development such as excessive gasoline consumption, traffic congestion and traffic accidents, and is also considered a huge source of air pollution in cities. To achieve this objective, it is important to fully grasp the need to understand and predict vehicle ownership patterns and the relevant influential factors that affect this matter. The attention of this topic is demonstrated by many studies that have examined vehicle ownership patterns and how their determinants have been utilized across both developed countries (Clark et al., 2016; Guo, 2013; Oakil et al., 2016; Ritter and Vance, 2013; Whelan, 2007; Yagi and Managi, 2016) and developing countries (Choudhary and Vasudevan, 2017; Guerra, 2015; He and Thogersen, 2017; Jou et al., 2012; Rahul and Verma, 2017; Soltani, 2017; Yang et al., 2017).

Nevertheless, the South-east Asian nations that comprise ASEAN as an active economic and fast-growth area have rarely been the focus of deeper analysis. Sillaparcharn (2007) produced a series of log-leaner models at the province and national scale to predict the vehicle type ownership trends of individuals in Thailand. Although these models served as useful statistical indexes, limitations of aggregate data and a shortage of explanatory variables remained as problematic issues. Moreover, splitting vehicle types in the independent models did not reveal any potential interactions between the variables and their influence on the alternatives. Another macro view on vehicle ownership can be found in the paper of Law et al. (2015) and Tuan (2011), which only concentrated on the relationships of two-wheel and four-wheel ownership rates in specific countries including eight representatives of ASEAN. In terms of a micro view, Tuan and Shimizu (2005) examined motorbike ownership in Hanoi, Vietnam as it represents the most popular means of commuting in the city. However, the analysis of Yamamoto (2009) covered all means of private vehicles in Osaka, Japan and Kuala Lumpur, Malaysia. Some note-worthy findings of these studies are that there was a lack of assessment in the features' impact; the analysis was done without consideration of the existing differences in terms of the sample's attributes by area; and finally, that the use of traditional methods seemed to be a backward approach in comparison with other fields, especially with regard to the context of big data mining.

The development of machine learning (ML) has become a revolution with regard to information technology, and its applications have widened to other specific fields including transportation. This is a new development when we consider how traditional statistical methods have been preferred for use in academic research studies. It has been

* Corresponding author.
E-mail addresses: 17096009@mmm.muroran-it.ac.jp (T.V. Ha), asada@mmm.muroran-it.ac.jp (T. Asada), arimura@mmm.muroran-it.ac.jp (M. Arimura).

https://doi.org/10.1016/j.jtrangeo.2019.05.015
Received 13 December 2018; Received in revised form 20 May 2019; Accepted 20 May 2019
Available online 28 May 2019

recognized that statistical methods are known to be valid and have been relied upon for their accuracy over an extended period of time. Moreover, these conventional methods have the advantage of being easily interpreted, evaluated and applied, while ML algorithms are thought of as a "black box" and it can be hard to understand their operational properties (Cantarella and de Luca, 2005). Nevertheless, recent studies have indicated the prominent superiorities of ML in transportation research studies. For example, in measuring the capability of neural networks (NN) and discrete choice models in mode-choice behavior, Hensher and Ton (2000) did not identify significant differences between the performance levels of different models, whereas in a later study NN was found to have outperformed in comparison to its competitors (Cantarella and de Luca, 2005). In another well-known ML algorithm, the support vectors machine (SVM) was marked as a superior method to predict the mode share among other ML and multinomial logit models (MNL) (Zhang and Xie, 2008). Lately, random forests (RF), a powerful ML algorithm, was assessed in terms of its performance in a study conducted by Hagenauer and Helbich (2017). It was implied that RF is the most accurate predictor, while MNL remained at the bottom of the list. These documents provide strong evidence that supports the potential applications of ML in transportation research studies.

Although there have been numerous studies that have used ML in the transportation field, the application for the vehicle ownership topic has not been considered. Karlaftis and Vlahogianni (2011) reviewed the comprehensive aspects of the two approaches, statistical and NN, used in transportation research studies with recommendations for choosing a suitable method for a specific case. Surprisingly, so far we have found only one study in the research of Karlaftis and among others that compare the performance of the nested logit model with NN in predicting household car-type owning decisions by Mohammadian and Miller (2002), but it did not assess all facets of the models other than the overall predicting capability. L. Cheng and co-authors listed several up-to-date studies that apply Random Forest (RF) in transportation, along with the researcher's intention of using this model in commuting mode choice analysis (Cheng et al., 2019). Notably. none of them was found to be relevant to vehicle ownership. The same situation has been identified for SVM, which can be realized in various articles (Allahviranloo and Recker, 2013; Sun and Park, 2017; Zhang and Xie, 2008). This shortcoming, on one hand, needs to be compensated for and on the other hand is wide open for further research.

In the context of the favorable economic, Phnom Penh, the capital of the Kingdom of Cambodia, one of the ASEAN members, has the opportunities to become a modern and developed city and also faces the challenge of motorization and its negative effects. The economic growth leads to the rapid rise of traffic demand including private vehicle ownership. The experiences from similar cities like Hanoi (Vietnam), Bangkok (Thailand), Jakarta (Indonesia) etc. are the worthwhile reference of the uncontrollable private vehicle rate which causes the serious problem for the city life. It is argued that the transport policy without the understanding of the commuters' behavior is one of the key factors resulting in this situation.

This study is aimed at determining the effective features that influence vehicle ownership in the city of Phnom Penh with regard to different contexts of building up the environment using advanced methods. This method of identification would be helpful in generating a given city's transportation policies for controlling the ownership of private vehicles. Additionally, the ability to forecast vehicle ownership is also evaluated in various scenarios, not only in terms of overall performance but also in terms of individual outcomes that would be used to better understand the structure of motorization of a given city. A favorable outcome would reinforce the ability to widen the applications of ML, not only in vehicle ownership investigations but also to other relevant transportation sectors.

The objective of our research is stated at three points. The first is the determination of the features that influence vehicle ownership in Phnom Penh city. The second is the exploration of the variation of the features' impact over the built-up environment in effecting a household decision of owning a vehicle. The third is the evaluation of the ability of prediction the vehicle ownership in the given city. There are also the research questions that will be addressed in the present study namely: In the category of socio-economic, built-up environment and other features that affect the vehicle ownership, which have a higher influence in comparison with the counterparts and whether they support the household from owning a private vehicle or not? Are the people live in the city's center and those who live in the suburban area affected by the mentioned features in the different magnitude and direction? How good predictor performance is when an overall measurement index is equal?

To address the questions are raised above, we propose to analyze the data collected from the questionnaire interview which covers the wide attributes of socio-economic, built-up environment and individual. Then we apply the conventional statistical and ML methods in various scenarios in order to examine the features' influence. The predictors' performance is then evaluated by the Kappa analysis to identify the outstanding model. The applied measurement method is trustworthy by satisfying the requirements of evaluation not only the overall accuracy but the ability to capture the diversity of vehicle ownership pattern.

This paper is arranged in four parts, the upcoming section will interpret the data preparation process and its applied methods. The data preparation process includes the description of the dataset and the explanation of the chosen features used for analyzing. Following this part is the introduction of the analysis scenarios and the interpretation of applied methods. The third section presents the principal results which begin with the models training process. The importance and the impact's direction of determinants and models' performance are also displayed in this section. In the last part, we will draw conclusions and include a discussion of this study.

2. Data and methodology

2.1. Data processing

2.1.1. Data set and variables

The data set used for Phnom Penh city was provided by Japan International Cooperation Agency (JICA). The data were acquired in 2012 with the original total records included 9,239 households. However, after rearranging the data and removing any missing data, we ended up with 8,842 records covering 96 traffic analysis zones (TAZ) within the city's boundaries. Fig. 1 presents Phnom Penh's traffic zone map and the boundaries of the city center and the suburban area. The data were enriched with additional features including population density and the length of the daily total trips made by the members of any given household.

Start from the idea of examining the vehicle ownership patterns in the comprehensive contexts of socioeconomic, demographic and transportation considerations. In this paper, the chosen features were grouped into three sectors including "household attributes", "built-up environment" and "personal trip characteristics".

a. Household attributes

Household income and size

Household income (Income) has a strong impact on vehicle ownership. It is assumed that in developing countries, wealthy families tend to upgrade their vehicles from two wheels to four wheels because of the attention to safety, convenience, and style. By contrast, a low-income family would be more motivated by the vehicle's basic functions that are affordable to them and that would still meet their daily commuting demands. In the present study, the household income was cataloged in 7 levels ranging from under 250 USD/month to over 2,000 USD/month.

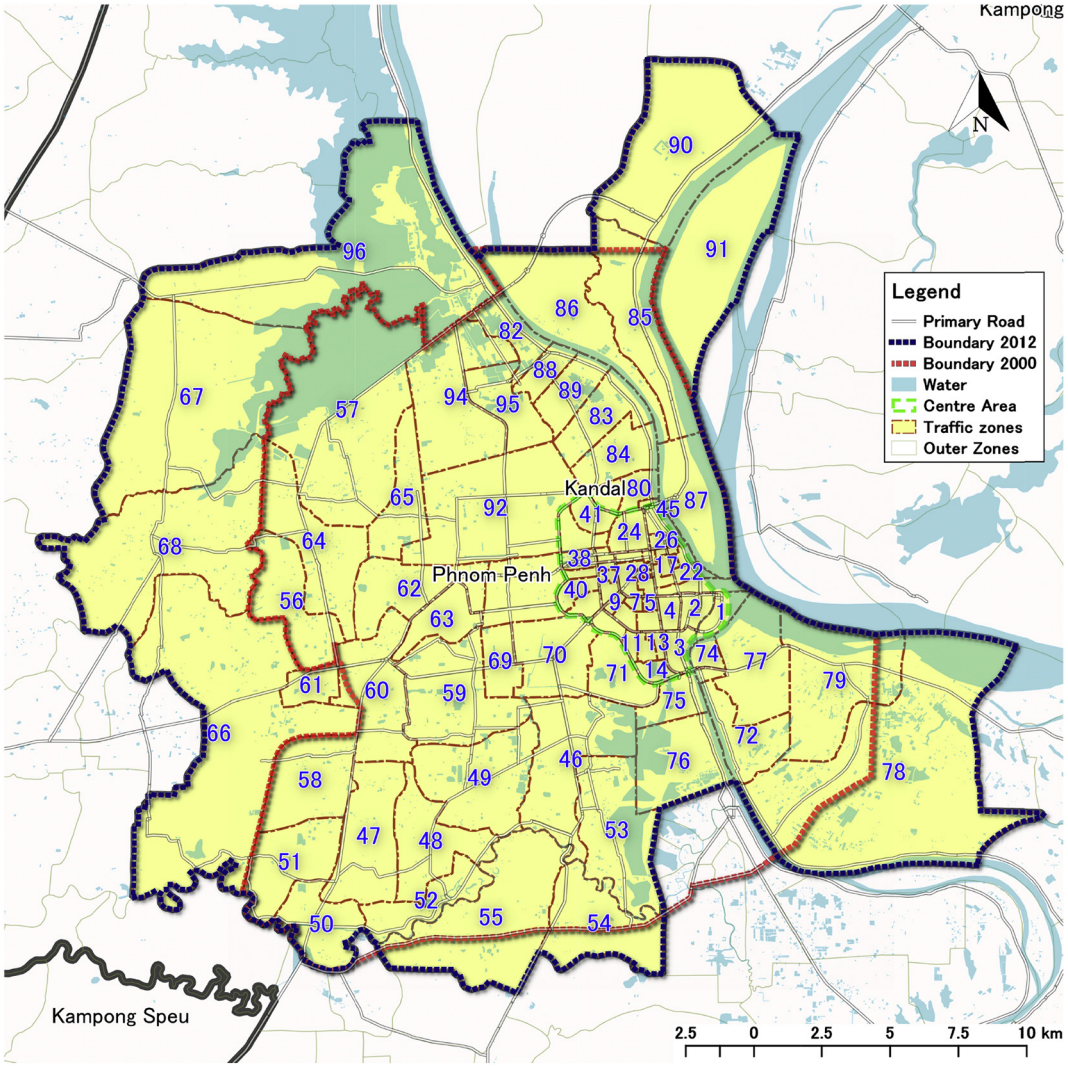

Fig. 1. Phnom Penh traffic analysis zones (TAZ) map.
Source: JICA survey data and Phnom Penh statistic book 2012.

The first 4 levels are indicated by an increase of 250 USD for each level. From level 5 to level 7, the amount of money in each level differed by 500 USD each.

The household size (HH.mems) followed household income as an important explanator in previous research studies. In the study of Maltha et al. (2017), "HH.mems" was the most contributive variable (with "Income") in explaining car ownership in the Netherlands in a positive relationship. Whereas, Ritter and Vance (2013) in their research found that "HH.mems" was one of the factors restraining car ownership in Germany. These results imply that the household size was not only a reliable variable but also had different influences on vehicle ownership. This variable in our study had a value ranging from 1 to 13 persons per household with a mean value of 4.9.

Household composition

In terms of the household composition, we considered four features including the number of persons over 16 years of age (O.16), the number of children under or equal to the age of 5 (Total.5), the number of people working in the house (TotalEm) and the total number of men in the family (Total.M). We assumed that the first feature would motivate motorization rate, while in Phnom Penh the average "O.16" reached a value of 3.7. Notably, the presence of children was suggested to impact on determining the number and type of vehicle that could meet the requirements of mobility, convenience, and safety. This led to members of the household would be concerned with having more vehicles as well as with their quality (by upgrading from motorbike to car for example).

The labor force plays an important role in contributing to the household's level of income. The higher the level of employment among members of the household would influence the number of vehicles owned by the members of that household. In the present study, the average "TotalEm" was 2.24, 2.14, and 2.31 in the greater city area, urban and suburban areas respectively. The "Total.M" feature was expected to reflect the travel behavior as well as vehicle ownership by gender in Phnom Penh city. The hypothesis was that men would travel more than women and would be more likely to commute by private vehicles, which can influence the possibility of owning more vehicles in the family. This variable did not vary much throughout the study area of Phnom Penh with the highest value of 9 and a mean of about 2.30.

b. Built-up environment.

In terms of the built-up environment features, certain variables were used such as the population density, land-use, employment density, etc. This paper intended to understand the relationship between the population density (Pop.Dens) and vehicle ownership in Phnom Penh city. The relationship will explain whether or not a high density of

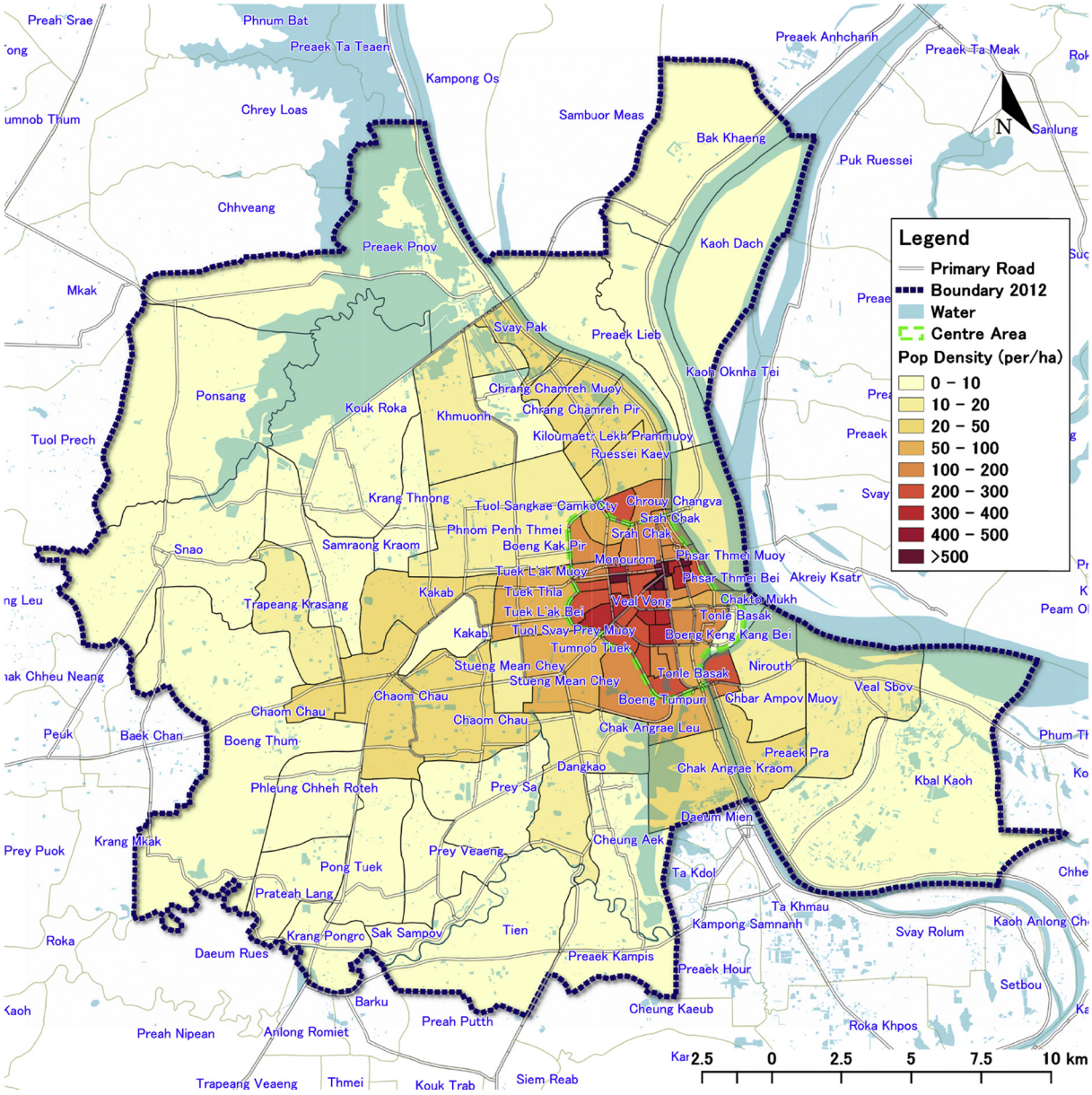

Fig. 2. Distribution of residents in Phnom Penh city in 2012.
Source: JICA survey data and Phnom Penh statistic book 2012.

population would stimulate the vehicle ownership rate. By integrating the statistical data and the TAZ plan, we utilized the distribution of the residents in Phnom Penh city in 2012, as is illustrated in the map in Fig. 2. The map denotes that the center area bears a significantly high capacity of population (over 100 persons per ha), while the density gradually declined with the increase in distance from the center area.

c. Personal trip characteristics

Personal trips, on one hand, represent the travel behavior of people and on the other hand account for traffic demand in the city. All of these characteristics were denoted to have a strong relationship with vehicle ownership in various studies (Guerra, 2015; Ritter and Vance, 2013; Soltani, 2017). In the case of Phnom Penh, there were three features associate with the personal trip attributes that were used to explain vehicle ownership namely: total number of trips made by household (TripNum), number of work-trips made daily (Work.trip) and total length of all trips (TotalLen).

The "TripNum" was calculated by summing up the number of trips made by all members of the household. The data denotes that members of households in suburban areas seemed to engage in more trips than those in urban areas by an average total number of trips of 9.41 and 9.35, respectively. This feature was believed to have a positive relationship with private vehicle ownership in Phnom Penh for the reasons of mobility demands in terms of daily travel.

"Work.trip" was a significant factor in terms of the purpose of taking trips ("to home", "to school", "to business", etc.). The increase in "Work.trip" was believed to have boosted private vehicle ownership. It can be argued that when we go to work there is a regulation to be on time. To avoid breaking the regulation because of dependence on the unreliable or inconsistent public transportation service, many residents choose to use a private vehicle to commute to their place of work or study. In the case of Phnom Penh, this kind of trip revealed an average number of trips of about 2.0 in all areas.

Note that the trip's length was not available in the data set. Consequently, we used an alternative method to calculate the trip

length using the QGIS program. For inter-zone trips, based on the location of the TAZ of the start and finish points, we hypothesized that the average trip length is the shortest route from the center of the original zone to the center of arrival zone. By using the QGIS program and integrating it with the road network of Phnom Penh city, we calculated the length of each trip. Additionally, the distance from the center of the traffic zones to the nearest road was also determined by QGIS and this value was added to the total trip length. For intra-zone trips with the same departure and arrival zone, we used another hypothesis and calculated the trip length using Equation (1), which was introduced by Fotheringham (1988).

$$D = 0.846 * \sqrt{A/\pi} \quad (1)$$

In this equation A is the area of the zone.

2.1.2. Correlation analysis

Before inserting the variables into the models, it is highly recommended to make an analysis of correlation. A strong relationship between predictor variables could result in an erroneous type of interaction and lead to bias in the model's performance. There are three types of methods, namely “Pearson”, “Polyserial” and “Polychoric”, that are used to determine the correlation between numeric variables and logical variables. In this study, we attempt to apply the two first methods when it is appropriate to the type of the variables. When the correlation value is significantly high, another method can be used to deal with this problem. One of the methods is to combine two related variables into one. For example, instead of listing the total income and the total number of household members, we can make a new variable by dividing the total income by the total members of the household. Even so, we need to check the type and attribute of the new feature whether they change or not.

As the results shown in Table 1, we found that the maximum correlation value was 0.65 (between the “TotalEm” and “Work.trip”). This value can then be used in the model.

2.2. Methodologies

2.2.1. Setting up scenarios

To understand the overall distribution of vehicle ownership throughout the city, an analysis was applied by area. Based on the differences in the population density and the level of income, we decided to analyze the data in three sets. The first set was made up of a “mixed” area which included the data of the whole survey area. The second and third sets consisted of the “urban” and “suburban” areas, respectively. While the “urban” group only included households that were located in the center of the city, the “suburban” group examined the vast number of households that resided in the surrounding and new areas of the city.

For a deeper examination of the structure of vehicle ownership in the city, we focused on the number and type of vehicles owned by each household. It is recognizable that the motorbike seems to be affordable for people who live in developing countries with medium to low levels of income. Moreover, when the public transport service is not convenient or is found to be insufficient for one's daily commuting purposes. Consequently, the ability of a household to buy more than one motorbike becomes feasible. For this reason, we created two outcome sets that focus on the number of motorbikes owned by a family. The first set focused on the motorized level of transportation, so it classified the household into three categories; ones which had no vehicle, ones which had only motorbikes (one or more) and the ones which had a car (s) with or without a motorbike. The second set was further widened by splitting household owned motorbike(s) into two outcomes; those owning only one motorbike and those with more than one motorbike. Combining the two groups of areas and the outcomes above, we arrived at six scenarios that we could use for our analysis as is shown in Table 2.

From a cursory look at the data in Table 2, we can recognize that the motorized index of Phnom Penh was quite high and most people used motorbike as a means of transportation. This also indicated that the percentage of the households that owned a car was at 18.93%. Notably, while the number of households whose members did not own motorized vehicles or only owned a motorbike increased from the urban to the suburban areas, the share of families that owned a car actually decreased.

2.2.2. Predictors and prediction process

For a comprehensive examination of the objectives, there were three algorithms that were applied including the multinomial logit model (MNL), a representative model for a statistical algorithm, the feedforward neural network (NN) and the random forest (RF) stand for the machine learning algorithm.

MNL, in terms of a disaggregate model, it has more advantages when compared with the aggregate model by its ability to reveal the causal relationships that exist between explanatory variables and the level of the outcomes (Bhat and Pulugurta, 1998). It also avoids bias from correlations that arise between aggregate units, which can be a serious problem (De Dios Ortúzar and Willumsen, 2011). MNL is also mentioned as an un-order response mechanism model that is based on the random utility maximization approach. Bhat and Pulugurta (1998) and Potoglou and Susilo (2008) gave detailed examples to demonstrate the outperformance of MNL with other models using an order response mechanism in examining vehicle ownership. For the above reasons, the application of the MNL model in this study was found to be suitable. The building of the MNL model process was based on the approach of neural networks that were introduced by Ripley (2007) and Venables and Ripley (2002) using accuracy as an index of the model's performance by changing the penalty of the sum of the square of the connection weights named *weight decay* that consisted of a value from 0 to 0.1.

Distinguished from MNL, NN applies the fixing error and pattern association approach for its algorithm (Hensher and Ton, 2000). In this paper, we used the common NN approach as it is a multi-feedforward layered neural network with three layers that have been presented in

Table 1
Results of correlation analysis among variables.

Variables	HH.mems	Total.5	O.16	Total.M	TotalEm	Income	Pop.Dens	TripNum	TotalLen	Work.trip
HH.mems	-									
Total.5	0.55	-								
O.16	0.64	0.09	-							
Total.M	0.62	0.33	0.38	-						
TotalEm	0.43	0.08	0.6	0.26	-					
Income	0.25	0.08	0.33	0.17	0.38	-				
Pop.Dens	0	0.04	0.06	-0.02	-0.04	0.18	-			
TripNum	0.49	0.01	0.38	0.34	0.35	0.21	-0.03	-		
TotalLen	0.25	-0.01	0.28	0.19	0.29	0.07	-0.32	0.45	-	
Work.trip	0.33	0.08	0.45	0.22	0.65	0.26	-0.02	0.58	0.33	-

Table 2
Proportion of vehicle ownership in 6 scenarios.

Area	4 classes				3 classes			Total
	NoVeh	X1Bike	X2Bikes	Car.Bike	NoVeh	Bike	Car.Bike	
Mixed	694	3057	3417	1674	694	6474	1674	8842
	7.85%	34.57%	38.65%	18.93%	7.85%	73.22%	18.93%	100.00%
Urban	221	902	1345	907	221	2247	907	3375
	6.55%	26.73%	39.85%	26.87%	6.55%	66.58%	26.87%	100.00%
Suburban	473	2155	2072	767	473	4227	767	5467
	8.65%	39.42%	37.90%	14.03%	8.65%	77.32%	14.03%	100.00%

Fig. 3. Each layer contains a number of units that represent certain features, hidden units and the outcomes, respectively. A hidden unit receives all of the signals from the input nodes that are multiplied by the connection weights (W_{ih}). These signals are summed by the product and then sent to the nonlinear function (also known as the activation function) before being passed onto the output unit. The usual activation functions include steps, logistics and the hyperbolic tangent function (Mohammadian and Miller, 2002). After emerging from the hidden unit, the signals are weighed again by their connection weights (W_{ho}). The arrive output node is then used to determine the outcomes. The model is trained by adjusting the connection weights beginning with W_{ho} in advance to minimize the errors between the actual outcomes and the response outcomes. This process is called back-propagation learning. When the number of hidden layers is fixed (one layer in this study) the other parameters are used to ascertain the number of units in this layer. By applying the same approach that is employed in the MNL model as introduced by Ripley (2007), the number of hidden units is determined based on the model's performance.

In the field of strengthening models, the ensemble technique is used to verify its power using two trends namely boosting (Schapire and Freund, 2012) and bagging (Breiman, 1996). Notably, RF is one of the members in the bagging family. Breiman described RF in his research paper (Breiman, 2001a, 2001b) as a combination of the decision trees, and its process is run by following the designated framework. Importantly, *n* trees are grown without pruning from different data sets that have been sampled using the bootstrap method (with replacement) on the original data. While the tree is forming, a constant number representing a subset of input variables (*mtry*) is identified. Variables are then randomly chosen for this subset and used for splitting the data at each node in the tree. Note that the tree is grown using about 63% of the data set, while the vast amount of remaining data is referred to as the out-of-bag data and is used to estimate the tree's error rate and also each variable's contribution. To reach an optimal RF model, the suitable number of random variables used in the single tree must be determined. An increase in this value influences the RF on both sides, one strengthens the single tree leading to an improvement in RF performance and the other increases the correlation from tree to tree resulting in a loss in the RF's level of performance (Breiman, 2001a). In this paper, the number of trees remained at 500 and the value of the *mtry* was received from 2 to 10 (equal to the maximum variables) over the process of identifying the best *mtry*.

For each algorithm, the training process had the responsibility of finding the outperforming model by determining the best-fit parameters and by avoiding any overfitting. The two traditional methods are bootstrap and k-fold cross-validation (CV) which could be applied in this case. While CV is known for having the advantage of being unbiased and for using less computing time, the bootstrap method performs better with a small sample set and a lower variable estimator. Recent research studies have found that the repeated 10-fold CV displayed better performance ability than the bootstrap method (Borra and Di Ciaccio, 2010; Kim, 2009). This supports the applications that have been proposed in this present paper. The final model with suitable parameters was captured by the 10-fold CV that was repeated 5 times on the R program platform with related packages.

2.2.3. Evaluating features' contribution and effect trend

Features' contribution refers to the effect's strength of the independent variables in terms of the results of the dependent variables.

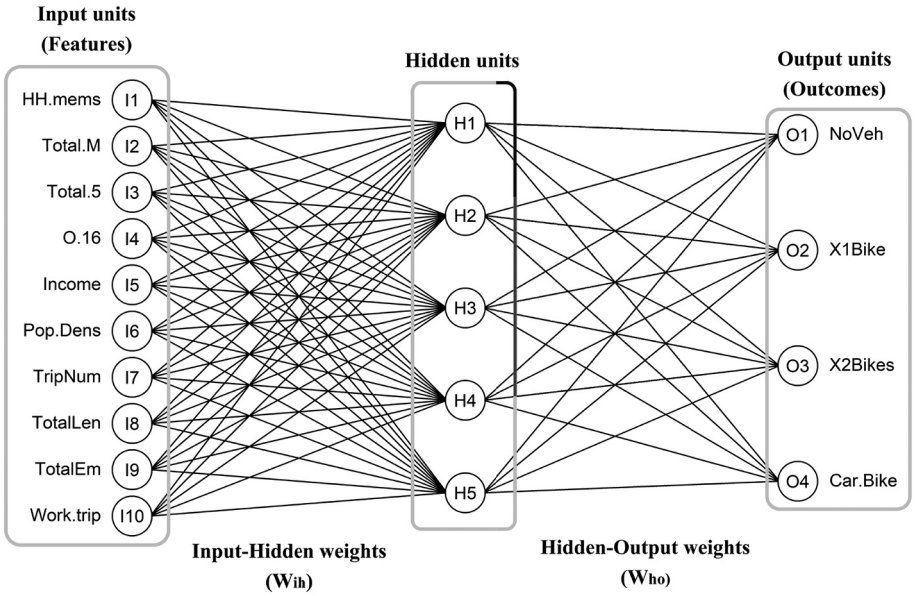

Fig. 3. Diagram of neural network model.

To put it another way, we strive to understand how sensitive the outcomes are to the variables. The task that answers the first research question of this paper was to determine how much of an influence occurs from the comprehensive elements such as any socio-economic variables, and just how much the aspects of the built-up environment affect the patterns of motorized vehicle usage in the city of Phnom Penh. The other major factor that needs to be considered is the influence trend of a variable on the outcomes. The trend can be represented in two ways; as positive in which the increasing value of the variable leads to an increase in the outcomes. Or it is more likely that the outcomes will occur and that this result is a reversal and is considered a negative trend.

With regard to the output of MNL, while the statistical indexes serve in the role of providing evidence which proves a real connection between the variables and the responses, the coefficients' magnitudes can not stand as the variables' representatives because of the notable differences in the measurements and the structures. To manage this issue, we used the standardized coefficients approach that was introduced in the article of Scott Menard (Menard, 2004) and again in a later paper (Menard, 2011). This approach employs the six methods applied for the logistic regression model. According to these articles, even with the employment of the six methods, the level's sequence of variables did not change. Thus, we decided to use the simplest method but with slight modifications in using only coefficients with significant values that are similar to the "relative importance" statistics, as was done in other noteworthy research studies (Levine, 1998; Zegras, 2010). The expression of this method is as follows: the effective index of variable i (E_i) is obtained from the summed absolute of the unstandardized coefficient (β_i) with the significant value multiplied by standard deviation (SD_i) over the outcomes (O), as is simulated in Equation (2).

$$E_i = \sum_{o=1}^{O} |\beta_i SD_i| \tag{2}$$

Differences between statistical models and ML models are complicated and it is not easy to perceive the variable importance due to the "black box" characteristic. Gevrey et al. (2003) interpreted and evaluated seven methods used to determine the variable's contribution in the NN model and noted that the "Perturb" and "Weights" are also simple and effective methods. In several later articles (Olden et al., 2004; Olden and Jackson, 2002), another approach named the "Connection Weight Approach" was presented by J.D. Olden and partners. This method was thought to be robust by the advantage of obtaining the inputs' effect in terms of both trend and strength over multiple hidden layers when compared with the "Weights" method (Beck, 2018). This method used the raw connection weights visualized by neural interpretation diagrams that were founded by Özesmi and Özesmi (1999) and is presented in the diagram shown in Fig. 3. First, we calculated the effect weight (E_{io}) of each input unit i that influenced the output unit o by summing the multiplication of connection weights from input unit i to hidden unit h (W_{ih}) and the connection weight from hidden unit h to output unit o (W_{ho}) (refer to Equation (3)). The sign of E_{io} implies an effect direction, in which the negative E_{io} value represents the inverted relationship between input i and output o and vice versa. The contribution of input i to the models (C_i) was then determined by summing the absolute value of E_{io} over the whole output (refer to Equation (4)).

$$E_{io} = \sum_{h=1}^{H} W_{ih} W_{ho} \tag{3}$$

$$C_i = \sum_{o=1}^{O} |E_{io}| \tag{4}$$

In the case of RF, there are two basic approaches used to determine variable importance called "mean decrease accuracy" (MDA) and "mean decrease Gini" (MDG). MDA using the variation of error rate before and after randomly permuted the value of the variables in the out-of-bag set; a higher change in errors indicated a stronger relationship between the variables and the outcomes. MDG is based on the Gini purity criterion when the data of the parent node was split into subnodes using variables from *mtry*. The more homogeneous the sub-data (high purity) was, a greater contribution of the variables was made on the response. Although MDG is preferred by many researchers (Archer and Kimes, 2008; Cheng et al., 2019), Breiman dropped MDA in his manual when applying RF (Breiman, 2007). Notably, MDG still retains its limitations of bias. Strobl et al. (2007) claimed that both original methods are influenced by the structure of predictor variables. The results from MDA and MDG could be misleading when being applied to large data types or the number of levels changed over the factor variables. Considering the data structure of the present research study, the variable set did not meet the criterion of Strobl; moreover, there was no missing data. Another effect on determining variable importance was implied by Hapfelmeier and Ulm (2014) who supported the use of MDA. The reading of relevant published research studies is recommended (Breiman, 2001a, 2001b; Han et al., 2017) in order to gain a clearer understanding of the above approaches. Note that at the present time we found no method for examining the effect trend of the explanatory variables on the dependent variables using RF. This is a notable weakness of this algorithm and this would need to be addressed in future research. Consequently, this study only considered reaching the stated objectives using the MNL and NN models.

2.2.4. Evaluation of predicting vehicle ownership pattern performance

With regard to the third research question, a necessary work was to examine the relevant vehicle ownership structure, in which the ability to correctly predict the portion of the household in each class is very important. As the era of motorization continues to expand in developing countries, the attitudes of people toward possessing modern motorized vehicles like cars is rising along with a decrease in the number of households possessing non-motorized transport. Although the proportion of these household types in the city of Phnom Penh remains at a low level, we propose to identify them by their attributes.

For classification models, to assess the performance, there are several methods can be applied. Here we refer to the three popular measurements namely: the overall accuracy, sensitivity, and specificity. The overall accuracy measures the proportion of correct prediction. The specificity and the sensitivity compute the correction of prediction at the true positive and true negative case. Besides that, the first measurement does not reveal the kind of errors made by the predictor. The two methods later do not demonstrate the overall performance score. With regard to this problem, we propose to apply the Kappa analysis which can satisfy two requirements. The detail of this method is described below.

Cohen introduced the Kappa in 1960 as an agreement index of two raters observing one problem. Imagine that rater one is the predicting model and that rater two assesses the actual data, the Kappa indicates the agreement between the raters and reflects how the model performs as correct for the actual data. The concept of this idea is presented in Equation (5) by Fleiss et al. (2003) using the relevant proportion, and in Equation (6) by Ben-David (2008) using the count. Either Equation (5) or (6) can be executed using the confusion matrix.

$$K = \frac{P_o - P_e}{1 - P_e} = \frac{\sum_{i=1}^{k} P_{ii}}{\sum_{i=1}^{k} P_{i.} P_{.i}} \tag{5}$$

$$K = \frac{N \times \sum_{i=1}^{k} x_{ii} - \sum_{i=1}^{k} x_{i.} x_{.i}}{N^2 - \sum_{i=1}^{k} x_{i.} x_{.i}} \tag{6}$$

Notably, P_o and P_e represent the observed weighted proportions of agreement and the change in the excepted weighted proportion of agreement. Additionally, $P_{.i}$ and $P_{i.}$ represent the total columns and rows proportions in Equation (5); x_{ii} is the count of cases in the main diagonal, k is the number of the outcomes, N is the number of examples,

and $x_{.i}$ and $x_{i.}$ are the total columns and rows counts, respectively as is presented in Equation (6). Kappa revealed a ranking from -1 (extreme differences between the model's predictions and the actual data) to +1 (model predictions perfectly fit the actual data).

When the confusion matrixes of various models have the same diagonal values but are not homogeneous with the other values, the Kappa values could be equal and the weighted Kappa could be applied to evaluate the cost-error. The weighted Kappa (K_w) can be obtained using the same form of Equation (5), but the values of P_o and P_e would need to be replaced by $P_{o(w)}$ and $P_{e(w)}$ in sequence. These would be calculated using the following Equations based on the concept of J. Fleiss (Fleiss et al., 2003).

$$P_{o(w)} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}P_{ij} \tag{7}$$

$$P_{e(w)} = \sum_{i=1}^{k}\sum_{j=1}^{k} w_{ij}P_{i.}P_{.j} \tag{8}$$

Here, w_{ij} represents agreement weights having a value of $0 \le w_{ij} \le 1$ and can be obtained from Equation (9) – the quadratic weight and Equation (10) – the linear weight as displayed below.

$$w_{ij} = 1 - \frac{(i-j)^2}{(k-1)^2} \tag{9}$$

$$w_{ij} = 1 - \frac{|i-j|}{k-1} \tag{10}$$

The illustration of Kappa of each classifier performance in all scenarios will be presented in Section 3.

3. Results

3.1. Models training

With regard to finding the best weight decay for the MNL models, the results are presented in Fig. 4. Notably, there were some differences found between the two groups of scenarios.

In the “4 outcomes” group, the accuracy of the urban and suburban models reached peak points when the weight decay was measured at about 0.0178 and 0.0422, respectively. After that, by the increased weight decay, the model's performance significantly decreased to the lowest values. However, the mixed model's accuracy fluctuated when the weight decay was within a small range value (from 0 to 0.01). Subsequently, the model performance increased quickly and archived the highest point when the “weight decay” was at the maximum value. In the “3 outcomes” scenario, the trends of the accuracies changed. While the mixed and suburban model's accuracy increased gradually and reached peak values when the weight decay was at a value of 0.1; however, the urban model accuracy dropped when the weight decay value increased to over 0.01.

With NN models, the weight decays were held with ten values for each scenario and the accuracy was found across the number of hidden units to increase from 1 to 19 (in an even sequence). With the “4 outcomes” groups, the accuracies displayed a similar pattern in that they increased and then became stable with an increase in the number of hidden units (refer to Fig. 5). Consequently, they reached stability at a later point. By contrast, the accuracy value trends in the “3 outcomes” groups were complicated, with the exception of the cases involving the suburban models. While the trends of the mixed models and suburban models mainly decreased (suburban models ran faster and were more stable) along with an increase in the hidden unit number, the urban models did increase in terms of the level of accuracy. Nevertheless, the final best tune for each model in this group had the same weight decay value of 0.0422 (refer to Table 3).

Training process results for RF are demonstrated in Fig. 6. As the “*mtry*” increased from a value of 2 to a value of 10, the accuracy of the model decreased. In the “4 outcomes” scenario, the models in the suburban area showed a slight increase at the maximum of the “*mtry*” value. However, the added accuracy of the suburban model still left a gap for its highest value. In the “3 outcomes” scenario, when the “*mtry*” values were 6 and 8, the performance of the suburban and urban models improved respectively when compared with the previous number of variables. Finally, the best tune of the RF models met the same value of “*mtry*” of 2. Table 3 lists all the tunes and their values in each scenario. This information will be used to evaluate the model's performance.

3.2. Feature-ranking in predicting vehicle ownership patterns

3.2.1. Groups of variables having high and low effects on classification

Figs. 7, 8 and 9 present the rankings of the features' attribution in the MNL, NN, and RF models, respectively. Besides the overall agreement between the models about the level of features' contribution, there are still remain some differences in ranking several individual cases. The results show that the population density mainly stayed at the low and medium important level in MNL and NN models, but in RF it appeared with higher impact except for the urban area with 4 outcomes.

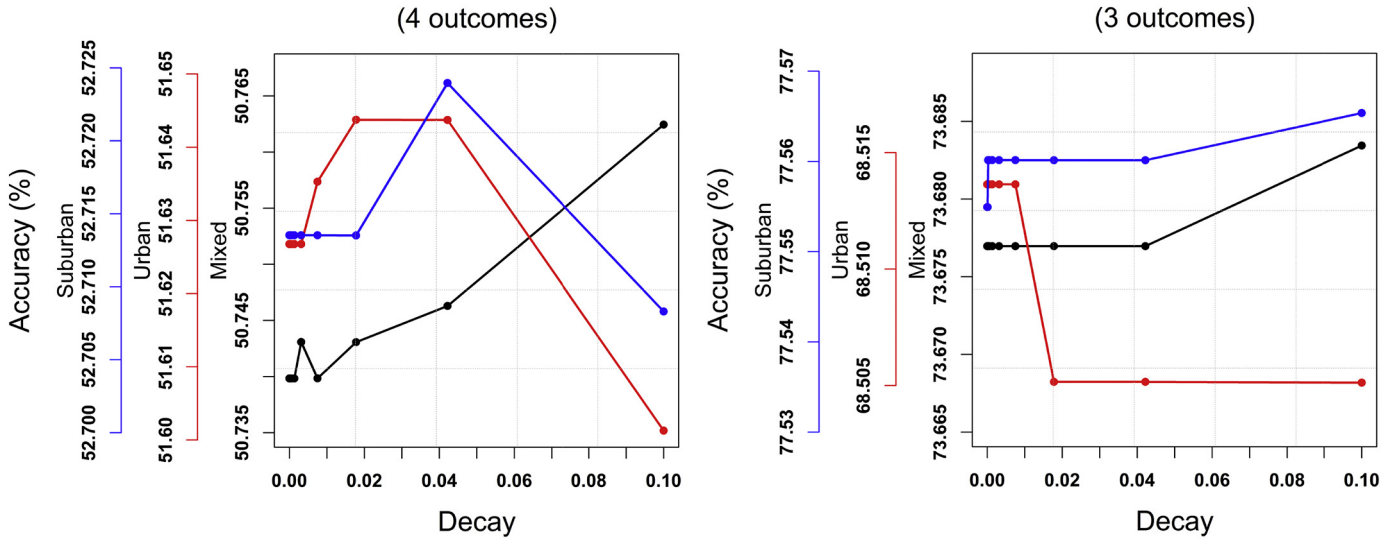

Fig. 4. Best weight decay for MNL model.

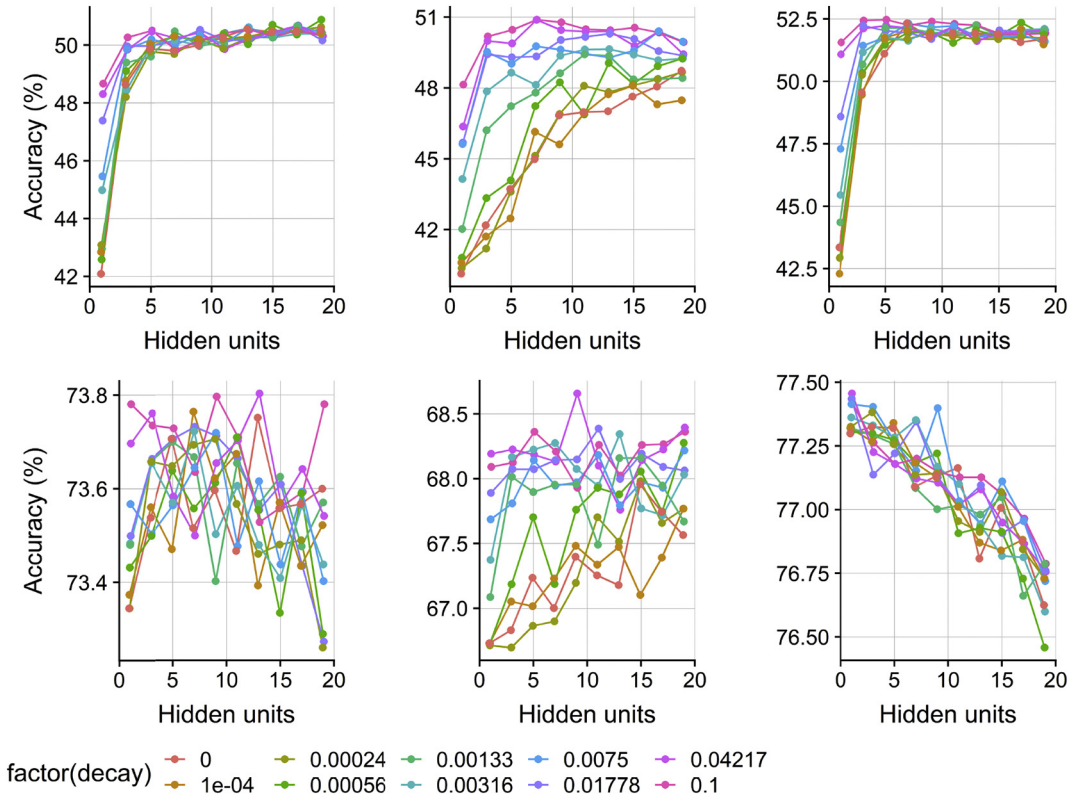

Fig. 5. Best hidden units for NN model. (Top row - 4 outcomes, bottom row - 3 outcomes; Left column – Mixed area, middle column – Urban area, right column – Suburban area.)

Table 3
Best tune results of each model by "10-fold" cross-validation.

Model	Tune	Value for 4 outcomes			Value for 3 outcomes		
		Mixed	Urban	Suburban	Mixed	Urban	Suburban
MNL	Decay	0.1	0.0178	0.0422	0.1	0.0075	0.1
NN	Size	19	7	5	13	9	1
	Decay	0.0006	0.1	0.1	0.0422	0.0422	0.0422
RF	Mtry	2	2	2	2	2	2

The contrary case was "O.16". While this feature appeared frequently in the top important group in MNL and NN model, it was stated as the weak explanatory variable in the RF two times, in the 3 outcomes of Mixed and Suburban models. It is suggested that the different algorithms produce the dissimilar outcome in several cases.

In order to more easily interpret the results, we grouped these features into three groups for each model. Group 1 presents the top three high effective features, while Group 3 presents the top three ineffective features. The vast values were put into the second group. Table 4 summarizes the appearance times of the features in the first and third groups in order from the first to the third in terms of the strength and weakness impacts, respectively.

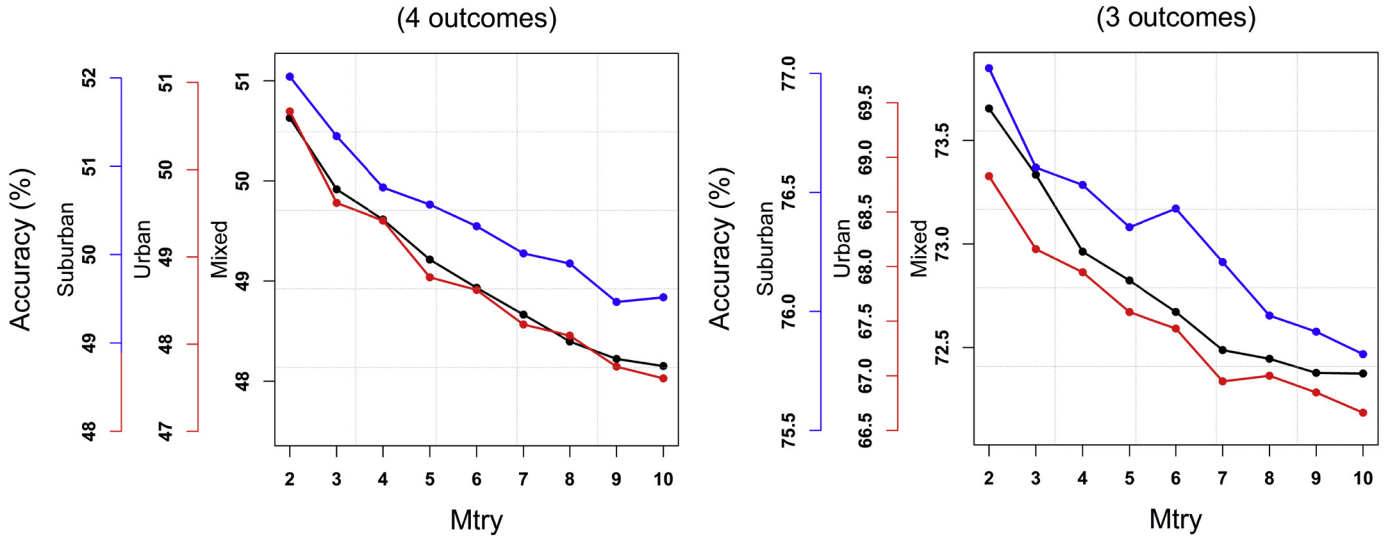

Fig. 6. Best "mtry" for RF model.

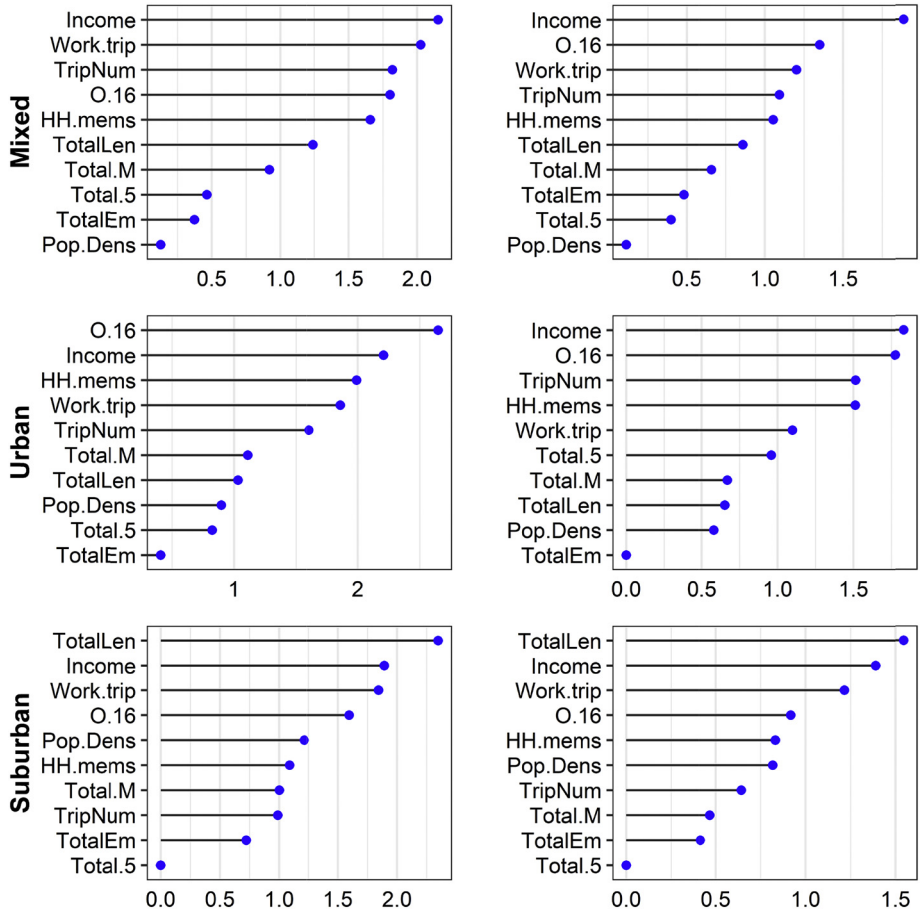

Fig. 7. Variable importance rank of MNL models (Left column – 4 outcomes; Right column – 3 outcomes.)

As the results indicate, the three most important features included "Income", "O.16" and "Work.trip". For these, "Income" was found to be the highest impact factor. It appeared 17 times in Group 1, which was 12 times as much as the first rank, 5 times as much as the second rank and did not appear in Group 3. The "O.16" variable stayed right after "Income", while this figure reached 11 times the value of appearance in Group 1 and remained mainly at the second rank (7 times). The "Work.trip" was slightly weaker than the "O.16" feature with 10 times the value emerging in Group 1 and 4 times in Group 3 (more than "O.16" 2 times). It should be noted that even "O.16" and "Work.trip" appeared in Group 3 at several times, but the value was considered not too low when compared with the average importance value in the model.

It is evident that the greatest weakness feature was "Total.5", which was found to be 13 times the amount in Group 3 and none in Group 1. The second feature was recorded as population density at 9 times in the low effective group, in which 6 times were at the lowest point and this was the same as with the number of children. There was a slight difference in contribution among the three features, namely "TripNum", "TotalLen" and "TotalEm". When the same amounts of time were used in Group 3, the total employment figure seemed to be weaker than the total trip-length by the difference occurring in Group 1 at 3 times and 5 times, respectively. Even as it was stated 7 times in Group 3, the family labor variable seemed to provide a greater advantage than "TripNum" when higher scores were archived in most cases.

The two features of "HH.mems" and "Total.M" were identified as good explanatory variables. Although their appearance was infrequent in Group 1 (at only 3 times for "HH.mems"), they still did not lose the capability of explaining the outcomes at 1 and 5 times in Group 3 with medium scores for household size and the number of men, respectively.

3.2.2. *Change of features' impact over areas*

By inspecting the features' importance ranking and the scores over areas, we can see that the effect level varied between the urban and suburban areas. Household income was not clearly different between the various areas because of its strength in most scenarios, but "O.16" seemed to have a higher influence in urban areas. The same situations were found for "HH.mems", "Total.M", "TripNum", and "Total.5", even if the presence of infants was ranked lower in urban areas of the NN models. However, higher scores were recorded in these categories when comparing urban areas with suburban areas.

Conversely, the categories of work-trips, "TotalLen", "TotalEm" and "Pop.Dens" were found to have a stronger effect in suburban areas than in urban areas. However, a weak variable such as population density showed its clear contribution in suburban areas as was seen in the MNL and RF models. The total trip length was found to have a lower rank in NN and RF, but it was superior by score in RF. Notably, it archived the highest rank in the MNL model. For the "TotalEm", it was found to have a greater influence in suburban areas in the MNL and RF models, but it was a little lower in the NN model. With regard to the last feature, the "Work.trip" was ranked higher in all scenarios.

3.3. *Effect trend of features on vehicle ownership*

In this section, the features effect trend will be revealed from the MNL and NN models (the RF was not available here as has been discussed in the previous sections). First, we will interpret the MNL models' results as a base-model and then we will look at the consequences of NN in comparison with MNL.

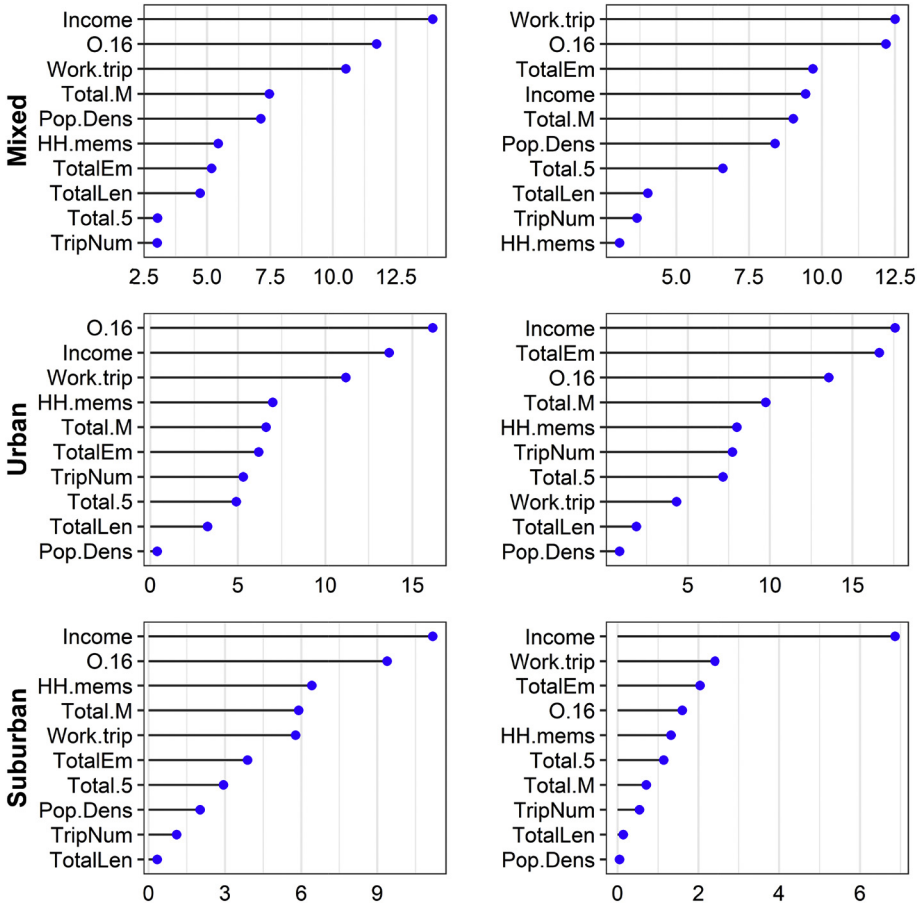

Fig. 8. Variable importance rank of NN models. (Left column – 4 outcomes; Right column – 3 outcomes.)

3.3.1. MNL models

Table 6 introduces detailed information of the MNL models in which households that had no motorized vehicles represented the reference outcomes. This means that the coefficients' sign can indicate a rise and fall in the motorization trend. By ignoring the variables that did not appear to be significant, we can divide the factors in the model into three groups – positive, negative and both negative and positive groups.

The positive group is the largest group as it contains six variables, namely "Total.M", "Total.5", "O.16", "Income", "TripNum" and "TotalLen". All features in this group (except for the presence of infants) were indicated as good explanatory variables by their level of significance in most of the cases. It is notable that the insignificant results appeared mainly in explaining the proportion of households that owned only one motorbike. The "Total.5" feature, which was not applicable in the suburban area, revealed significance for other areas especially in urban areas.

The negative group consists of three variables including "HH.mems", "TotalEm" and "Work.trip". Similar to the "Total.M", the "Work.trip" was found to be significant in all models and across all outcomes, but in a negative way. Even though there was no evidence that it affected households with own only one motorbike in urban and suburban areas with 4 outcomes, the household size was found to be significant in the rest of the cases. These variables are indicated as being contrary to our hypothesis. Eventually, the "TotalEm" feature showed a mostly significant effect on car ownership within that household.

The last group contained only one variable – population density. The unique aspect of this variable was that it changed the direction of the effect over all models. As the table data indicates, the effect stated was negative in the urban models and positive in the suburban models.

3.3.2. NN models

Instead of keeping the no-vehicle households as a reference, the NN model returned the connection weight values with a magnitude and sign for all of the alternatives. In comparison, we used a fraction in which the numerator and denominator represent the number of effect weights of the NN model that have the same sign in the MNL model, and also represents the total significant coefficients of the MNL model in certain specific fields (features, outcomes or overall) respectively, as can be seen in the agreement index.

The statistical comparison of Table 5 and Table 6 indicates a level of about 66% of agreement between the NN and MNL models in an overall effect trend. Thus, the agreement of the positive, negative and for both signs would account for about 69%, 76%, and 36%, respectively. In the positive group, the concurrence of three features was highlighted; namely "Total.M", "O.16" and "TripNum", for which the measurement index was in the sequence of 11/15, 10/11 and 10/14. The "TotalLen" came as the fourth agreement followed by "Income" and "Total.5" with the same sign of about 50%. The negative group revealed total agreement in terms of the "TotalEm" features followed by "Work.trip" and "HH.mems" with agreement index values of 6/6, 11/15 and 9/13, respectively. The greatest disagreement appeared in the feature of population density, which was the unique variable in the last group that had an agreement index value of 4/11.

Interestingly, we found a strong agreement between MNL and NN in expressing the effect trend of features on non-motorized vehicle owners and car-owning households. In the MNL model, when positive features increased in value, members of the household would be prevented from owning vehicles. Simultaneously, in the NN model, this change led to a rise in households with non-motorized transport. In the latter case, the agreement percentage of the features' effect trend in explaining

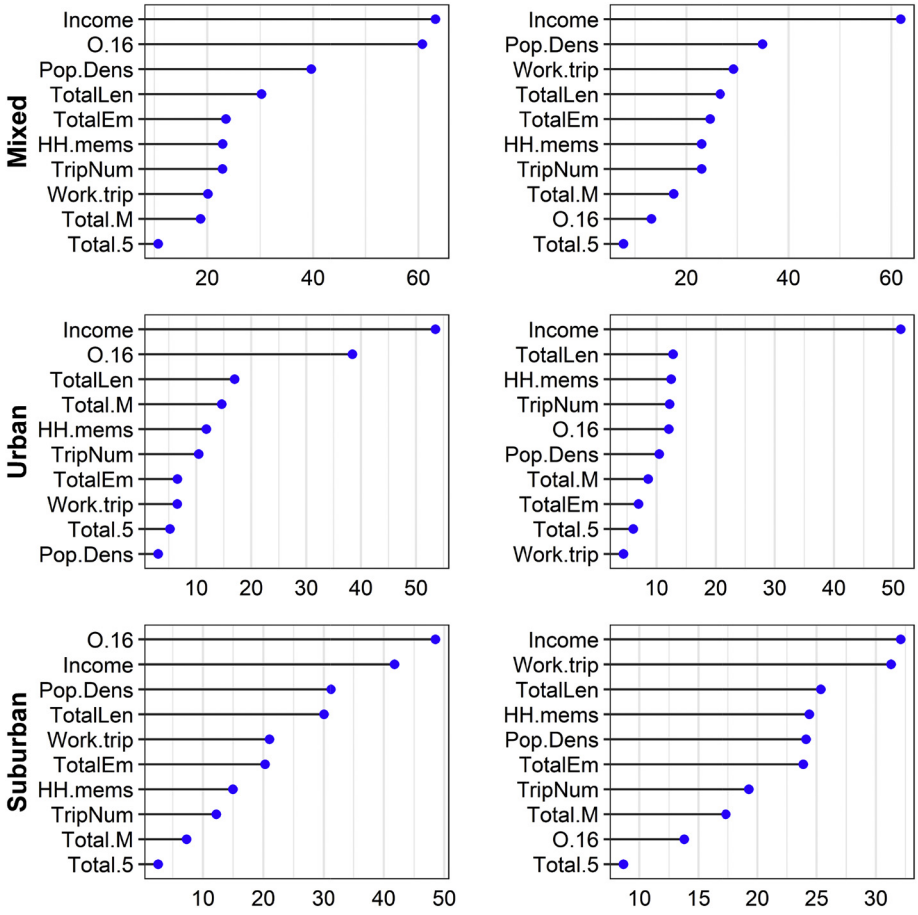

Fig. 9. Variable importance rank of RF models. (Left column – 4 outcomes; Right column – 3 outcomes.)

Table 4
Summary of variable importance ranking.

Variable	High impact level rank				Low impact level rank			
	1st	2nd	3rd	Total	1st	2nd	3rd	Total
HH.mems			3	3	1			1
Total.M				0		2	3	5
Total.5				0	6	5	2	13
O.16	3	7	1	11		2		2
Income	12	5		17				0
Pop.Dens		1	2	3	6	1	2	9
TripNum			2	2	1	2	3	6
TotalLen	2	1	2	5	1	3	3	7
TotalEm		1	2	3	2	3	2	7
Work.trip	1	3	6	10	1		3	4
Total	18	18	18	54	18	18	18	54

households that owned cars was about 82%. This was particularly true in the suburban area where it was completely the same in both the NN and MNL models.

3.4. Prediction capability on vehicle ownership patterns

As was introduced in the Methodologies Section, the training data (70% of original data) in this section served as the fitting data that went along with the second part as a way of predicting data to evaluate the models' performance. Table 7 displays the results of the indicator by models across all scenarios. Outlines of this process are described below.

The models tended to lose accuracy in predicting data, except in three cases involving the MNL and NN models in mixed areas with 4 outcomes and the MNL model in mixed areas with 3 outcomes. Notably, there was no clear discrimination among the methods in each scenario. The RFs were marked as having the highest fall accuracy performance, while in the fitting data accuracy scores in a range of about 96% to 100% were reached, which then dropped to around 48% to 78%. In addition, the amplitude of accuracy was just about 2%.

It was also stated that the overall accuracy varied over locations and outcome scenarios. If the accurate prediction values of the “4 outcomes” model were about 49% to 54%, the values of the “3 outcomes” model reached about 68% to 78%. The results indicate that the models displayed the lowest accuracy value in urban areas with an average of 48.93% for the “4 outcomes” model and 68.21% for the “3 outcomes” model. Whereas, the most accurate prediction for the “4 outcomes” model was in the mixed area and for the “3 outcomes” model, it was in the suburban area with average correction rates of 54.26% and 77.45%, respectively.

The sensitivity values indicate that the classification of non-motorized households had a very low correction rate, while the maximum sensitivity of this category was 10.61% in urban areas with 3 outcomes. Additionally, the cases involving car-owning households did not meet the high accuracy standard. The best prediction value of households with cars was 44.12% in urban areas with 4 outcomes and the poorest prediction value occurred in suburban areas using either the “4 outcomes” model or the “3 outcomes” model with average correction values of 8.26% and 6.38%, respectively.

When examining the Kappa value results in Table 7, it is feasible to conclude that the Kappa statistics were not so high. The average of the unweighted Kappa was not more than 0.3. Notably, at only one time in the NN models with 4 outcomes did this score reach a value of 0.306. It is also notable that in two cases in urban areas, MNL models were

Table 5
Variables affecting weights of NN models.

		HH.mems	Total.M	Total.5	O.16	Income	Pop.Dens	TripNum	TotalLen	TotalEm	Work.trip
M4	NoVeh	**2.99**	-3.68	-1.48	-2.90	-3.78	-1.83	**1.21**	-1.24	**2.32**	**4.80**
	X1Bike	**1.02**	**0.22**	**0.62**	-3.42	-3.87	**1.31**	**0.82**	**2.00**	**0.92**	**2.14**
	X2Bikes	-0.46	**2.42**	-0.11	**3.76**	-0.36	**1.29**	-0.12	-0.15	**0.68**	-1.50
	Car.Bike	-0.97	**1.15**	-0.83	**1.66**	**5.94**	-2.69	**0.87**	-1.33	-1.26	-2.08
U4	NoVeh	**2.04**	-2.77	-1.56	-4.80	-4.76	**0.03**	-2.15	-1.64	**2.13**	**5.00**
	X1Bike	**1.46**	-0.54	-0.43	-3.30	-2.06	**0.14**	-0.51	**0.06**	-0.34	**0.60**
	X2Bikes	-3.06	**1.14**	**2.45**	**5.55**	**0.76**	**0.02**	**1.29**	**0.48**	**0.95**	-2.53
	Car.Bike	-0.43	**2.17**	-0.47	**2.52**	**6.10**	-0.20	**1.37**	**1.10**	-2.78	-3.07
S4	NoVeh	**3.02**	-1.84	-0.93	-3.00	-3.11	-0.17	-0.51	-0.12	**1.16**	**2.81**
	X1Bike	-0.04	-0.99	**0.68**	-1.62	-1.49	-0.28	**0.14**	-0.05	**0.85**	**0.01**
	X2Bikes	-1.62	**0.57**	-0.44	**4.32**	-0.94	-0.54	**0.23**	**0.13**	-0.28	-2.48
	Car.Bike	-1.74	**2.50**	**0.89**	**0.45**	**5.65**	**1.05**	**0.23**	**0.04**	-1.59	-0.47
M3	NoVeh	-0.55	-4.47	**1.39**	-5.01	-3.11	-4.47	**1.26**	**1.96**	**3.98**	**4.98**
	Bike	-1.95	**1.31**	**1.19**	**2.03**	-2.36	**0.71**	**0.48**	**0.88**	**0.58**	-1.37
	Car.Bike	**0.55**	**3.24**	-4.02	**5.16**	**3.98**	**3.21**	-1.91	-1.18	-5.13	-6.16
U3	NoVeh	**1.91**	-2.57	-3.61	-6.79	-4.69	**0.31**	-3.27	**0.66**	**4.28**	**2.17**
	Bike	-3.97	**4.83**	**3.26**	**0.43**	-4.04	**0.11**	-0.57	**0.27**	**4.01**	-1.62
	Car.Bike	**2.13**	-2.36	**0.29**	**6.34**	**8.86**	-0.44	**3.90**	-0.96	-8.34	-0.53
S3	NoVeh	**0.59**	-0.32	-0.51	-0.72	-3.08	-0.02	-0.24	-0.06	**0.91**	**1.08**
	Bike	**0.07**	-0.04	-0.06	-0.08	-0.35	0.00	-0.03	-0.01	**0.10**	**0.12**
	Car.Bike	-0.66	**0.35**	**0.57**	**0.80**	**3.44**	**0.02**	**0.27**	**0.07**	-1.02	-1.21

Note: M, U, and S refer to Mixed area, Urban area and Suburban area.
3, 4 prefixes refer to 3 outcomes and 4 outcomes.
Bold characters indicate the positive weights.

superior for both indicators in terms of the unweighted Kappa and accuracy values, but when weighted Kappa was utilized (including linear and quadric functions) the NN models achieved the best results. In summary, NN outperformed by four times the highest weighted Kappa value, followed by RF at two times, whereas MNL performed the worst by displaying the lowest weighted Kappa value in all scenarios. For the last notation, in contrary to the overall corrected predictions, the increasing level of accuracy from the "4 outcomes" model to the "3 outcomes" model resulted in a decrease in the Kappa values.

4. Conclusion and discussions

4.1. Phnom Penh vehicle ownership determinants

This study applied statistical and ML algorithms in examining vehicle ownership determinants in the city of Phnom Penh through the use of household attributes and individual trip survey data. The results demonstrate that household income was the most important feature followed by the number of adults from 16 years old and the number of to work trips. Although it was indicated as weak explanatory variables, the presence of infant and population density still exhibited significance in predicting vehicle ownership in the city. The other features namely household size, number of male family members, total trip-length, household workers and total trips made were stated as good variables in sequence of the impact rank. There was evidence that the vehicle ownership determinants varied across areas by their magnitude and trend. While the figures related to work-trips, commuting distance, household labor force and population density were more likely to have an effect in suburban areas, the remaining variables (except household income) were stronger explanators in urban areas than in those suburban areas.

Even though expenditure was indicated as the preeminent variable when compared with income in previous research studies (Choudhary and Vasudevan, 2017; Dash et al., 2013), household income monthly still served as a reliable explanator in this study. Expenditure, in large scale contexts like countries or regions has its advantages where income is difficult to collect or in cases when monthly income fluctuates by season (agriculture sector, etc.) and expenditures are reasonable. Nevertheless, if we use expenditure as an explanator following the concept of Dash et al. (2013), the relevant data on food, energy, miscellaneous, clothing and footwear would need to be collected. These data may not be available for all areas or may be difficult to obtain, so household income is more suitable as an explanator.

Work-trips, total employment and household size were the factors that restrained motorized vehicle ownership rates in Phnom Penh. Even though the proportion of employment in the family did not appear to have a significant impact in all cases, it was still a good explanator for households that owned cars. This is contrary to our own hypothesis and those of some other studies (Choudhary and Vasudevan, 2017; Potoglou and Susilo, 2008; Soltani, 2017; Yamamoto, 2009). These features are related to travel demands of transportation. The increase in family members and labor caused a rise in daily trips. It is supposed that there is a causal relationship between lifting travel demands and an increase in the vehicle ownership rate. However, this was not necessarily true in the city of Phnom Penh. It could be suggested that even though the travel demands in households increased, it did not result in an effort to buy more vehicles by those residents. Consequently, when public transport was not attractive or accessible, shared vehicles would be used.

The addition of the personal trip attributes was a significant variable, although attributes like these are rarely used to explain vehicle ownership patterns. The trips number and travel distances are representative of transportation's capacity and mobility. An increase in these fields requires an increased demand by the respondents of transportation services. In the case of the city of Phnom Penh, the public transport system was extremely deficient and insufficient. Additionally, the popularity of the motorbike at an affordable ownership price and with the advantage of mobility, could result in it being seen as a more attractive option for members of households. By contrast, a study by Shen et al. (2016) showed the negative effects of commuting distance relevant to car ownership in big cities in China. Additionally, even though the lengths of the trips in this study were not actually traced, they still served as a good explanator. Nowadays, with the development of mobile devices and the widening applications of the geographic positioning system (GPS), this variable could be determined exactly and would provide better result in the models.

The "Total.M" and "O.16" reflected the overall travel behavior in the city of Phnom Penh. The age of sixteen is the age threshold for which a person can operate and own a motorized vehicle like a motorbike in ASEAN countries, and this would lead to the possibility for

Table 6
Multinomial logit model parameters.

	Variables	(Intercept)	HH.mems	Total.M	Total.5	O.16	Income	Pop.Dens	TripNum	TotalLen	TotalEm	Work.trip
M4	X1Bike	0.9564	-0.209	0.1885	0.134	0.0387	0.1916	-0.0006	0.0933	0.0141	0.0478	-0.2885
		(0.2532)***	(0.086)*	(0.0623)**	**(0.1103)**	**(0.0682)**	(0.0724)**	(0.0003)*	(0.0246)***	(0.0036)***	**(0.0767)**	(0.0577)***
	X2Bikes	-1.5666	-0.5692	0.3693	0.3488	0.7644	0.5939	-0.0001	0.1799	0.0198	-0.073	-0.4648
		(0.2836)***	(0.0965)***	(0.0693)***	(0.1235)**	(0.0765)***	(0.0751)***	**(0.0003)**	(0.0268)***	(0.0039)***	**(0.0845)**	(0.0632)***
	Car.Bike	-3.2196	-0.4612	0.2843	0.3216	0.6225	1.1764	0.0001	0.2088	0.0207	-0.3314	-0.5353
		(-6.0779)***	(-6.4573)***	(5.8417)***	(3.0967)**	(10.9038)***	(8.2505)***	**(-0.2447)**	(7.2249)***	(5.4527)***	(-0.9494)***	(-7.9867)***
Residual Deviance: 13713.05; AIC: 13779.05												
U4	X1Bike	0.8184	-0.1992	0.2244	0.2641	0.1976	0.1814	-0.001	0.0804	0.0195	-0.0745	-0.2701
		(0.1824)***	**(0.138)**	(0.1137)*	**(0.1969)**	**(0.1322)**	**(0.1078)**	(0.0003)***	**(0.0443)**	**(0.0112)**	**(0.1386)**	(0.1097)*
	X2Bikes	-1.6659	-0.7979	0.4175	0.624	1.1015	0.5484	-0.001	0.2104	0.0359	-0.0581	-0.4932
		(0.1462)***	(0.1428)***	(0.115)***	(0.2012)**	(0.1366)***	(0.1057)***	(0.0003)***	(0.0445)***	(0.0111)**	**(0.1378)**	(0.1091)***
	Car.Bike	-3.0884	-0.5666	0.3402	0.4928	0.9038	1.0786	-0.0013	0.2037	0.0386	-0.3608	-0.4629
		(0.1639)***	(0.1492)***	(0.1201)**	(0.2091)*	(0.1418)***	(0.1078)***	(0.0003)***	(0.0463)***	(0.0114)***	(0.1442)*	(0.1132)***
Residual Deviance: 5088.84; AIC: 5154.84												
S4	X1Bike	0.783	-0.1684	0.2008	-0.0294	-0.005	0.198	0.0022	0.0673	0.0192	0.0091	-0.2327
		(0.3286)*	**(0.1079)**	(0.0765)**	**(0.1297)**	**(0.086)**	(0.0998)*	**(0.0012)**	(0.0295)*	(0.0044)***	**(0.0965)**	(0.0672)***
	X2Bikes	-1.9566	-0.5216	0.4125	0.2048	0.71	0.6645	0.0081	0.0934	0.0353	-0.1981	-0.3633
		(0.3422)***	(0.1124)***	(0.0794)***	**(0.1343)**	(0.0896)***	(0.1021)***	(0.0013)***	(0.0305)**	(0.0045)***	(0.0987)*	(0.0695)***
	Car.Bike	-4.302	-0.3586	0.3525	0.1741	0.5509	1.3447	0.0124	0.1072	0.0413	-0.4489	-0.5677
		(0.3994)***	(0.1279)**	(0.0916)***	**(0.1542)**	(0.1012)***	(0.1119)***	(0.0014)***	(0.0347)**	(0.0049)***	(0.1133)***	(0.0809)***
Residual Deviance: 8232.154; AIC: 8298.154												
M3	Bike	0.2922	-0.3779	0.3087	0.2587	0.4333	0.4855	-0.0005	0.1204	0.017	-0.096	-0.3161
		(0.2454)	(0.0851)***	(0.0603)***	(0.1083)*	(0.066)***	(0.0702)***	(0.0002)*	(0.0236)***	(0.0035)***	**(0.0745)**	(0.0557)***
	Car.Bike	-3.0369	-0.4253	0.3025	0.3167	0.6093	1.2225	-0.0002	0.1697	0.0208	-0.4271	-0.4491
		(0.283)***	(0.0987)***	(0.0696)***	(0.1254)*	(0.0764)***	(0.0757)***	**(0.0003)**	(0.0266)***	(0.0039)***	(0.0854)***	(0.0634)***
Residual Deviance: 8185.653; AIC: 8229.653												
U3	Bike	0.0903	-0.4784	0.3432	0.5606	0.6026	0.3586	-0.0009	0.1532	0.0231	-0.0328	-0.3026
		(0.4095)	(0.1505)**	(0.1085)**	(0.204)**	(0.122)***	(0.1042)***	(0.0003)***	(0.043)***	(0.0108)*	**(0.1276)**	(0.1039)**
	Car.Bike	-2.7479	-0.567	0.2544	0.7593	0.747	1.0002	-0.0013	0.2353	0.0251	-0.2408	-0.4008
		(0.4505)***	(0.1662)***	(0.1185)*	(0.222)***	(0.1346)***	(0.1089)***	(0.0003)***	(0.0465)***	(0.0114)*	**(0.1383)**	(0.1112)***
Residual Deviance: 3317.91; AIC: 3361.91												
S3	Bike	0.5053	-0.3224	0.2157	0.1632	0.3188	0.3963	0.0037	0.0654	0.025	-0.0324	-0.2772
		(0.3077)	(0.1052)**	(0.0725)**	**(0.1296)**	(0.0799)***	(0.0953)***	(0.0012)**	(0.0286)*	(0.0043)***	**(0.0914)**	(0.0654)***
	Car.Bike	-3.4506	-0.3618	0.2309	0.2626	0.4097	1.2348	0.0101	0.1093	0.0378	-0.3721	-0.5011
		(0.3821)***	(0.1273)**	(0.0895)**	**(0.1589)**	(0.0964)***	(0.1098)***	(0.0014)***	(0.034)**	(0.0048)***	(0.1113)***	(0.08)***
Residual Deviance: 4722.313; AIC: 4766.313												

Note: M, U, and S refer to Mixed area, Urban area and Suburban area.
3, 4 prefixes refer to 3 outcomes and 4 outcomes.
*, **, *** refer to p-value at the three ranks of less than 0.001, 0.01 and 0.05, respectively.
Values in parenthesis represent standard errors.
Bold characters indicate the insignificant parameters.

these individuals to buy a motorbike for daily commuting purposes. In looking over the differences by gender, it appears that men in the households were more likely to use vehicles than woman. This can be explained by the priorities of the people who serve in the role of the main income earner of the family.

The differences in the urban and suburban vehicle ownership determinants were indicated by two items – changes in the features impact level and the effect direction. As has been seen in other research studies (Choudhary and Vasudevan, 2017; Guerra, 2015), it is suggested that an inhomogeneous influence of household attributes across all areas will reflect the variety of the features effect. Consider the cases where the "Total.5" are concerned. In the suburban area models, this feature was insignificant and turned to become significant with high coefficient values in urban areas. It is assumed that people living in the city's center would pay more attention to child care than those in suburban areas. Unlike the presence of infants, the population had a strong significant effect on urban and suburban areas but in a contradictory way. This was shown in the research of Yang et al. (2017), but there was no clear relationship in the numbers when we compare the center of the city and the surrounding areas. Another research study by Soltani (2017) did not find a relationship between population density and vehicle ownership, but Salon (2009) determined that this was a substantial variable.

4.2. Prediction capability of vehicle ownership patterns and limitations of the study

In predicting vehicle ownership patterns, the study results demonstrate the outperformance of ML relative to the statistical models, not only in terms of accuracy but also by the ability to control any unbalanced alternatives. While the better performance of ML was proved by the number of authors listed in Section 1, the higher values produced in Kappa have been expressed in the present paper. By using the confusion matrix introduced by Xie et al. (2003), we utilized the Kappa and weighted Kappa values for all models and found that the NN model was much better in comparison with the MNL model. In applying the data of Zhang and Xie (2008), it was noted that RF displayed the highest degree of accuracy, but was lower in Kappa values than the MNL model while the NN model came in as third in the testing data. This suggests that the method used for finding the best tune in their study was different in the present paper.

It has been highlighted that the method of predicting the four classes of household vehicle ownership revealed a lower accuracy value but was higher than the Kappa value in three classes. The phenomenon of changing the Kappa value when combining the outcomes was stated in a study of Warrens (2012). This research found that after merging a couple of categories, the Kappa can rise or fall and if the Kappa value increases, it may be difficult to discriminate between these two

Table 7
Accomplishment of vehicle ownership predicting.

	Area	Model	Sensitivity/accuracy/kappa															
			On fitting data								On predicting data							
			NoVeh	Bike*	Bikes	Car.Bike	Accuracy	Kappa	wKappa.L	wKappa.Q	NoVeh	Bike*	Bikes	Car.Bike	Accuracy	Kappa	wKappa.L	wKappa.Q
4 outcomes	Mixed	MNL	0.00	64.86	63.38	21.67	51.02	0.244	0.292	0.352	0.00	66.19	66.63	21.31	52.68	0.268	0.322	0.388
		NN	3.91	61.68	61.37	33.28	51.65	0.265	0.332	0.414	6.25	64.67	64.39	34.46	**54.26**	**0.306**	**0.376**	**0.459**
		RF	85.60	99.39	97.41	93.52	**96.43**	**0.948**	**0.952**	**0.958**	5.77	62.38	64.20	25.70	51.70	0.260	0.325	0.404
	Average		29.84	75.31	74.05	49.49	66.37	0.485	0.525	0.574	4.01	64.41	65.07	27.16	52.88	0.278	0.341	0.417
	Urban	MNL	1.94	57.12	67.73	37.17	52.37	0.285	0.345	0.418	0.00	53.70	67.74	31.99	**49.95**	**0.246**	0.302	0.370
		NN	0.00	57.28	61.15	46.30	52.12	0.288	0.357	0.437	0.00	53.33	58.06	44.12	49.26	0.245	**0.321**	**0.412**
		RF	96.77	99.53	100.00	99.69	**99.58**	**0.994**	**0.994**	**0.995**	7.58	46.67	61.29	37.87	47.58	0.217	0.282	0.364
	Average		32.90	71.31	76.29	61.05	68.02	0.522	0.565	0.617	2.53	51.24	62.37	37.99	48.93	0.236	0.302	0.382
	Suburban	MNL	0.00	70.51	61.06	13.78	52.86	0.242	0.288	0.346	0.00	68.89	60.55	9.13	51.40	0.218	0.246	0.281
		NN	0.00	71.90	63.68	6.33	53.36	0.245	0.292	0.352	0.00	69.81	65.54	4.35	**52.99**	**0.238**	**0.272**	**0.316**
		RF	88.86	99.67	96.76	92.55	**96.63**	**0.950**	**0.953**	**0.958**	1.42	67.18	61.03	11.30	51.34	0.224	0.263	0.315
	Average		29.62	80.69	73.83	37.55	67.62	0.479	0.511	0.552	0.47	68.63	62.37	8.26	51.91	0.227	0.260	0.304
3 outcomes	Mixed	MNL	0.00	96.67		15.96	73.80	0.120	0.131	0.151	0.00	97.22		15.34	74.10	0.123	0.130	0.143
		NN	7.41	97.44		12.29	74.25	0.126	0.140	0.166	3.37	96.29		10.76	72.81	0.082	0.093	0.116
		RF	78.19	100.00		90.61	**96.51**	**0.913**	**0.918**	**0.927**	4.81	96.60		17.13	**74.36**	**0.151**	**0.164**	**0.189**
	Average		28.53	98.04		39.62	81.52	0.387	0.396	0.415	2.72	96.70		14.41	73.76	0.119	0.129	0.149
	Urban	MNL	2.58	92.50		26.46	68.85	0.189	0.203	0.228	1.52	91.69		27.94	**68.68**	**0.192**	0.203	0.224
		NN	10.32	92.82		29.13	70.29	0.237	0.257	0.293	10.61	91.54		25.74	68.58	0.192	**0.211**	**0.247**
		RF	97.42	100.00		98.11	**99.32**	**0.986**	**0.987**	**0.988**	4.55	89.02		29.04	67.39	0.177	0.197	0.233
	Average		36.77	95.10		51.23	79.49	0.471	0.482	0.503	5.56	90.75		27.57	68.21	0.187	0.204	0.235
	Suburban	MNL	0.00	98.95		7.82	77.59	0.062	0.069	0.082	0.00	98.58		6.52	77.18	0.047	0.052	0.061
		NN	0.00	99.59		4.28	77.59	0.037	0.040	0.047	0.00	99.53		3.48	77.49	0.029	0.032	0.038
		RF	80.12	100.00		89.20	**96.76**	**0.909**	**0.919**	**0.927**	2.13	98.58		9.13	**77.67**	**0.081**	**0.096**	**0.110**
	Average		26.71	99.52		33.77	83.98	0.336	0.343	0.352	0.71	98.90		6.38	77.45	0.052	0.060	0.069

Note: W.Kappa.L – weight Kappa with linear function.
W.Kappa.Q – weight Kappa with quadric function.
"Bike*" in 3 outcome models representing households owning only motorbike(s).
Bold numbers are the optimal value in each scenario.

categories. Nevertheless, in the case of the city of Phnom Penh, the merging of two outcomes led to a decrease in Kappa values. Moreover, while the correction of the combined outcomes was much higher (over 90%) than the originals (about 60%), the proportion of incorrect predictions for car ownership increased. It appears that the results were not clearly distinguishable between the households that owned one motorbike and those that had more than one motorbike. Furthermore, the splitting of these two outcomes may support the model to recognize the households that owned cars. Eventually, further studies on this issue, on the one hand, need to focus more on building a classifying model and on the other hand, need to provide an alternative set of predictors as to whether the high predicting correction value is preferred for car-owning households or motorbike owning households.

4.3. Limitations and future work

Although this study has covered large features in order to attempt to explain vehicle ownership, the lack of some variables has limited the outcomes. As various researchers have pointed out, the price of vehicles, fuel (Dargay, 2002) and the cost of other relevant fees like vehicle insurance, maintenance fees (Whelan, 2007) and the availability of parking spaces (Guo, 2013) will affect the decision-making process of a person in terms of deciding whether to own a vehicle, especially with regard to purchasing a car. The other factor is the existence of a public transport system. In many cases, the quality of a public transport system will dissuade people from owning a motorized vehicle. This finding can be verified in the studies of Jiang et al. (2017); Yagi and Managi (2016); Yamamoto (2009) and Zegras (2010).

Determining the importance of variables in a classified model is still a very interesting field of study. This work has provided useful tools for researchers in choosing the most appropriate variables in the models that can deal with a huge amount of data and can cut down on computation costs. Moreover, with regard to the planners or policymakers, knowing the influencing strength of features could help them adjust and control the outcomes in the most effectual or fastest way, but with an affordable impact on the current situation. Taking this demand into account, many classifiers have been integrated into algorithms that are used in the methods employed to evaluate the strength of the associated variables. Even though a model may be followed according to various methods, it is advised to apply these methods with caution, especially while this matter is still in debate. This is true in cases using the NN model (Fischer, 2015; Olden et al., 2004) or the RF model (Archer and Kimes, 2008; Genuer et al., 2010; Hapfelmeier and Ulm, 2014; Janitza et al., 2016; Strobl et al., 2007). In another research study, (Hagenauer and Helbich, 2017), the permutation-based method (same concept of MDA) was applied to evaluate the variable effect levels in seven classifiers involving the MNL, NN, RF models as well as others. The results of these studies can serve as valuable reference resources for future studies. Lastly, M. Kuhn recommended using the intrinsic method of models because of its direct interconnection (Kuhn and Johnson, 2013). This research study followed this suggestion by considering the structure of the explanator features and we found a desirable upshot with the level of agreement among the models. However, the process of selecting a suitable method must still be of extreme importance in future studies.

Acknowledgments

The authors greatly thank the Japan International Cooperation Agency (JICA) for providing data for this study.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sector.

Declarations of interest

None.

References

Allahviranloo, M., Recker, W., 2013. Daily activity pattern recognition by using support vector machines with multiple classes. Transp. Res. Part B Methodol. 58, 16–43. https://doi.org/10.1016/j.trb.2013.09.008.

Archer, K.J., Kimes, R.V., 2008. Empirical characterization of random forest variable importance measures. Comput. Stat. Data Anal. 52, 2249–2260. https://doi.org/10.1016/j.csda.2007.08.015.

Beck, M.W., 2018. NeuralNetTools : visualization and analysis tools for neural networks. J. Stat. Softw. 85, 1–20. https://doi.org/10.18637/jss.v085.i11.

Ben-David, A., 2008. Comparison of classification accuracy using Cohen's Weighted Kappa. Expert Syst. Appl. 34, 825–832. https://doi.org/10.1016/j.eswa.2006.10.022.

Bhat, C.R., Pulugurta, V., 1998. A comparison of two alternative behavioral choice mechanisms for household auto ownership decisions. Transp. Res. Part B Methodol. 32, 61–75. https://doi.org/10.1016/S0191-2615(97)00014-3.

Borra, S., Di Ciaccio, A., 2010. Measuring the prediction error. A comparison of cross-validation, bootstrap and covariance penalty methods. Comput. Stat. Data Anal. 54, 2976–2989. https://doi.org/10.1016/j.csda.2010.03.004.

Breiman, L., 1996. Bagging predictors. Mach. Learn. 24, 123–140. https://doi.org/10.1007/BF00058655.

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

Brieman, L., 2001. Statistical Modeling: The Two Cultures. Stat. Sci. 16, 199–231.

Breiman, L., 2007. Manual Setting Up, Using, And Understanding Random Forests V4.0. https://www.stat.berkeley.edu/~breiman/Using_random_forests_v4.0.pdf 136. pp. 23–42.

Cantarella, G.E., de Luca, S., 2005. Multilayer feedforward networks for transportation mode choice analysis: an analysis and a comparison with random utility models. Transp. Res. Part C Emerg. Technol. 13, 121–155. https://doi.org/10.1016/j.trc.2005.04.002.

Cheng, L., Chen, X., De Vos, J., Lai, X., Witlox, F., 2019. Applying a random forest method approach to model travel mode choice behavior. Travel Behav. Soc. 14, 1–10. https://doi.org/10.1016/J.TBS.2018.09.002.

Choudhary, R., Vasudevan, V., 2017. Study of vehicle ownership for urban and rural households in India. J. Transp. Geogr. 58, 52–58. https://doi.org/10.1016/j.jtrangeo.2016.11.006.

Clark, B., Lyons, G., Chatterjee, K., 2016. Understanding the process that gives rise to household car ownership level changes. J. Transp. Geogr. 55, 110–120. https://doi.org/10.1016/j.jtrangeo.2016.07.009.

Dargay, J.M., 2002. Determinants of car ownership in rural and urban areas: a pseudo-panel analysis. Transp. Res. Part E Logist. Transp. Rev. 38, 351–366. https://doi.org/10.1016/S1366-5545(01)00019-9.

Dash, S., Vasudevan, V., Singh, S., 2013. Disaggregate model for vehicle ownership behavior of Indian households. Transp. Res. Rec. 55–62. https://doi.org/10.3141/2394-07.

De Dios Ortúzar, J., Willumsen, L.G., 2011. Modelling Transport, 4th ed. John Wiley & Sons, Ltd.

Fischer, A., 2015. Garson's method trumps Olden's method in every case - how to determine relative importance of input-variables in nonlinear regression with artificial neural networks. Ecol. Modell. 309–310, 60–63. https://doi.org/10.1016/j.ecolmodel.2015.04.015.

Fotheringham, A.S., 1988. Market Share Analysis Techniques: a Review and Illustration of Current U.S. Practice. In: Wrigley, N. (Ed.), Store Choice, Store Location and Market Analysis, 1st ed. Routledge, London, pp. 120–159. https://doi.org/10.4324/9781315736686.

Fleiss, J.L., Levin, B., Paik, M.C., 2003. Statistical Methods for Rates and Proportions, 3rd ed. Wiley & Sons, Inc, Hoboken, New Jersey.

Genuer, R., Poggi, J.M., Tuleau-Malot, C., 2010. Variable selection using random forests. Pattern Recognit. Lett. 31, 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014.

Gevrey, M., Dimopoulos, I., Lek, S., 2003. Review and comparison of methods to study the contribution of variables in artificial neural network models. Ecol. Modell. 160, 249–264. https://doi.org/10.1016/S0304-3800(02)00257-0.

Guerra, E., 2015. The geography of car ownership in Mexico City: a joint model of households' residential location and car ownership decisions. J. Transp. Geogr. 43, 171–180. https://doi.org/10.1016/j.jtrangeo.2015.01.014.

Guo, Z., 2013. Does residential parking supply affect household car ownership? The case of New York City. J. Transp. Geogr. 26, 18–28. https://doi.org/10.1016/j.jtrangeo.2012.08.006.

Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Syst. Appl. 78, 273–282. https://doi.org/10.1016/j.eswa.2017.01.057.

Han, H., Guo, X., Yu, H., 2017. Variable selection using Mean Decrease Accuracy and Mean Decrease Gini based on Random Forest. Proc. IEEE Int. Conf. Softw. Eng. Serv. Sci. ICSESS 219–224. https://doi.org/10.1109/ICSESS.2016.7883053.

Hapfelmeier, A., Ulm, K., 2014. Variable selection by Random Forests using data with missing values. Comput. Stat. Data Anal. 80, 129–139. https://doi.org/10.1016/j.csda.2014.06.017.

He, S.Y., Thogersen, J., 2017. The impact of attitudes and perceptions on travel mode choice and car ownership in a Chinese megacity: the case of Guangzhou. Res. Transp. Econ. 62, 57–67. https://doi.org/10.1016/j.retrec.2017.03.004.

Hensher, D.A., Ton, T.T., 2000. A comparison of the predictive potential of artificial neural networks and nested logit models for commuter mode choice. Transp. Res. Part E Logist. Transp. Rev. 36, 155–172. https://doi.org/10.1016/S1366-5545(99)

00030-7.
Janitza, S., Tutz, G., Boulesteix, A.L., 2016. Random forest for ordinal responses: Prediction and variable selection. Comput. Stat. Data Anal. 96, 57–73. https://doi.org/10.1016/j.csda.2015.10.005.
Jiang, Y., Gu, P., Chen, Y., He, D., Mao, Q., 2017. Influence of land use and street characteristics on car ownership and use: Evidence from Jinan, China. Transp. Res. Part D Transp. Environ. 52, 518–534. https://doi.org/10.1016/j.trd.2016.08.030.
Jou, R.C., Huang, W.H., Wu, Y.C., Chao, M.C., 2012. The asymmetric income effect on household vehicle ownership in Taiwan: A threshold cointegration approach. Transp. Res. Part A Policy Pract. 46, 696–706. https://doi.org/10.1016/j.tra.2012.01.001.
Karlaftis, M.G., Vlahogianni, E.I., 2011. Statistical methods versus neural networks in transportation research: differences, similarities and some insights. Transp. Res. Part C Emerg. Technol. 19, 387–399. https://doi.org/10.1016/j.trc.2010.10.004.
Kim, J.H., 2009. Estimating classification error rate: repeated cross-validation, repeated hold-out and bootstrap. Comput. Stat. Data Anal. 53, 3735–3745. https://doi.org/10.1016/j.csda.2009.04.009.
Kuhn, M., Johnson, K., 2013. Applied Predictive Modeling. Springerhttps://doi.org/10.1007/978-1-4614-6849-3.
Law, T.H., Hamid, H., Goh, C.N., 2015. The motorcycle to passenger car ownership ratio and economic growth: a cross-country analysis. J. Transp. Geogr. 46, 122–128. https://doi.org/10.1016/j.jtrangeo.2015.06.007.
Levine, J., 1998. Rethinking accessibility and jobs-housing balance. J. Am. Plan. Assoc. 64, 133–149. https://doi.org/10.1080/01944369808975972.
Maltha, Y., Kroesen, M., Van Wee, B., van Daalen, E., 2017. Changing influence of factors explaining household car ownership levels in the Netherlands. Transp. Res. Rec. J. Transp. Res. Board 103–111. https://doi.org/10.3141/2666-12.
Menard, S., 2004. Six approaches to calculating standardized logistic regression coefficients. Am. Stat. 58, 218–223. https://doi.org/10.1198/000313004X946.
Menard, S., 2011. Standards for standardized logistic regression coefficients. Soc. Forces 89, 1409–1428. https://doi.org/10.1093/sf/89.4.1409.
Mohammadian, A., Miller, E.J., 2002. Nested logit models and artificial neural networks for predicting household automobile choices comparison of performance. Transp. Res. Rec. 1807, 92–100. https://doi.org/10.3141/1807-12.
Oakil, A.T.M., Manting, D., Nijland, H., 2016. Determinants of car ownership among young households in the Netherlands: the role of urbanisation and demographic and economic characteristics. J. Transp. Geogr. 51, 229–235. https://doi.org/10.1016/j.jtrangeo.2016.01.010.
Olden, J.D., Jackson, D.A., 2002. Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. Ecol. Modell. 154, 135–150. https://doi.org/10.1016/S0304-3800(02)00064-9.
Olden, J.D., Joy, M.K., Death, R.G., 2004. An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data. Ecol. Modell. 178, 389–397. https://doi.org/10.1016/j.ecolmodel.2004.03.013.
Özesmi, S.L., Özesmi, U., 1999. An artificial neural network approach to spatial habitat modelling with interspecific interaction. Ecol. Modell. 116, 15–31. https://doi.org/10.1016/S0304-3800(98)00149-5.
Potoglou, D., Susilo, Y., 2008. Comparison of vehicle-ownership models. Transp. Res. Rec. J. Transp. Res. Board 2076, 97–105. https://doi.org/10.3141/2076-11.
Rahul, T.M., Verma, A., 2017. The influence of stratification by motor-vehicle ownership on the impact of built environment factors in Indian cities. J. Transp. Geogr. 58, 40–51. https://doi.org/10.1016/j.jtrangeo.2016.11.008.
Ripley, B.D., 2007. Pattern Recognition and Neural Networks. Cambridge University Press.
Ritter, N., Vance, C., 2013. Do fewer people mean fewer cars? Population decline and car ownership in Germany. Transp. Res. Part A Policy Pract. 50, 74–85. https://doi.org/10.1016/j.tra.2013.01.035.
Salon, D., 2009. Neighborhoods, cars, and commuting in New York City: a discrete choice approach. Transp. Res. Part A Policy Pract. 43, 180–196. https://doi.org/10.1016/j.tra.2008.10.002.
Schapire, R.E., Freund, Y., 2012. Boosting: Foundations and Algorithms. MIT Press. The MIT Press, Cambridge, Massachusetts.
Shen, Q., Chen, P., Pan, H., 2016. Factors affecting car ownership and mode choice in rail transit-supported suburbs of a large Chinese city. Transp. Res. Part A Policy Pract. 94, 31–44. https://doi.org/10.1016/j.tra.2016.08.027.
Sillaparcharn, P., 2007. Modeling of vehicle ownership. Transp. Res. Rec. J. Transp. Res. Board 2038, 98–104. https://doi.org/10.3141/2038-13.
Soltani, A., 2017. Social and urban form determinants of vehicle ownership; evidence from a developing country. Transp. Res. Part A Policy Pract. 96, 90–100. https://doi.org/10.1016/j.tra.2016.12.010.
Strobl, C., Boulesteix, A.L., Zeileis, A., Hothorn, T., 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinform. 8. https://doi.org/10.1186/1471-2105-8-25.
Sun, B., Park, B.B., 2017. Route choice modeling with support vector machine. Transp. Res. Procedia 25, 1811–1819. https://doi.org/10.1016/j.trpro.2017.05.151.
Tuan, V.A., 2011. Dynamic interactions between private passenger car and motorcycle ownership in asia: a cross-country analysis. J. East. Asia Soc. Transp. Stud. 9, 541–556. https://doi.org/10.11175/eastpro.2011.0.97.0.
Tuan, V.A., Shimizu, T., 2005. Modeling of household motorcycle ownership. J. East. Asia Soc. Transp. Stud. 6, 1751–1765.
Venables, W.N., Ripley, B.D., 2002. Modern Applied Statistics With S. Springer.
Warrens, M.J., 2012. A family of multi-rater kappas that can always be increased and decreased by combining categories. Stat. Methodol. 9, 330–340. https://doi.org/10.1016/j.stamet.2011.08.008.
Whelan, G., 2007. Modelling car ownership in Great Britain. Transp. Res. Part A Policy Pract. 41, 205–219. https://doi.org/10.1016/j.tra.2006.09.013.
Xie, C., Lu, J., Parkany, E., 2003. Work travel mode choice modeling using data mining: decision trees and neural networks. Transp. Res. Rec. 1854, 50–61. https://doi.org/10.3141/1854-06.
Yagi, M., Managi, S., 2016. Demographic determinants of car ownership in Japan. Transp. Policy 50, 37–53. https://doi.org/10.1016/j.tranpol.2016.05.011.
Yamamoto, T., 2009. Comparative analysis of household car, motorcycle and bicycle ownership between Osaka metropolitan area, Japan and Kuala Lumpur, Malaysia. Transportation (Amst.) 36, 351–366. https://doi.org/10.1007/s11116-009-9196-x.
Yang, Z., Jia, P., Liu, W., Yin, H., 2017. Car ownership and urban development in Chinese cities: a panel data analysis. J. Transp. Geogr. 58, 127–134. https://doi.org/10.1016/j.jtrangeo.2016.11.015.
Zegras, C., 2010. The built environment and motor vehicle ownership and use: evidence from Santiago de Chile. Urban Stud. 47, 1793–1817. https://doi.org/10.1177/0042098009356125.
Zhang, Y., Xie, Y., 2008. Travel mode choice modeling with support vector machines. Transp. Res. Rec. J. Transp. Res. Board 2076, 141–150. https://doi.org/10.3141/2076-16.