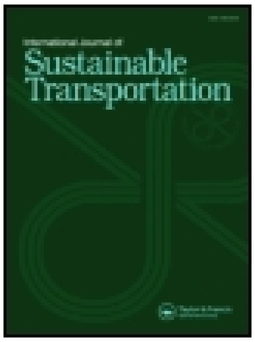

Taylor & Francis
Taylor & Francis Group

International Journal of Sustainable Transportation

ISSN: 1556-8318 (Print) 1556-8334 (Online) Journal homepage: https://www.tandfonline.com/loi/ujst20

Household travel mode choice estimation with large-scale data—an empirical analysis based on mobility data in Milan

Leilei Liang, Meng Xu, Susan Grant-Muller & Lorenzo Mussone

INTERNATIONAL JOURNAL OF SUSTAINABLE TRANSPORTATION
https://doi.org/10.1080/15568318.2019.1686782

Household travel mode choice estimation with large-scale data—an empirical analysis based on mobility data in Milan

Leilei Liang[a], Meng Xu[a], Susan Grant-Muller[b], and Lorenzo Mussone[c]

[a]State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China; [b]Institute for Transport Studies, University of Leeds, Leeds, UK; [c]Department of Architecture, Built Environment and Civil Construction, Politecnico di Milano, Milano, Italy

ABSTRACT

Data analysis plays a key role in supporting the development of sustainable transportation. Using the large-scale household mobility survey data collected in Milan, Italy during 2005–2006, we study whether the large-scale data contribute to improving accuracy in estimating household travel modes. This paper presents three machine learning methods including multinomial logit (MNL) model, random forest (RF) and support vector machine (SVM) to estimate the household travel mode. Their model accuracies are 70.41%, 71.89%, 72.74% respectively under the full sample size. It is found that the accuracies of these three methods fluctuate fiercely when the sample size is less than 20,000 and then stabilize gradually with continuous increasing it. After stabilization occurs, accuracies with these three methods do not significantly increase as the sample size continues to increase. We also study the travel characteristics derived from the large-scale survey data, which is fundamental for developing a sustainable transportation system. The collected data items include five explanatory variables, i.e., household size (HS), vehicle ownership, household income (HI), travel distance, travel time and one response variable (i.e., household travel mode), which includes public transport (PT), private car, usage of PT and private car simultaneously and the others travel modes (e.g., walk). We further investigate the importance of explanatory variables in terms of estimating household travel mode choice with the MNL model. It is found that vehicle ownership is the most critical factor influencing household travel mode choice, followed by travel distance, travel time, HS and HI. The ranking result is consistent with the RF approach.

ARTICLE HISTORY
Received 12 July 2019
Accepted 27 October 2019

KEYWORDS
Estimation; household mobility; large-scale data; machine learning; sustainable transportation system; travel mode choice

1. Introduction

1.1. Background

The emergence of big data makes a multitude of societal problems improved, such as enhancing public safety,[1] cutting down cost,[2] providing better customer service,[3] and also, supporting transportation sustainable development. The application of big data in transportation field has also become a hot topic and attracts much attention. For example, existing studies have used sundry data to estimate/predict individual travel behavior (Liu et al., 2013), to evaluate the relationship between undergraduate education and sustainable transport attitudes (Kim et al., 2016), to develop smart transportation system (Zhang et al., 2011), etc. Basically, the premise of carrying out these studies is the presence of authentic, accurate and sufficient data. However, it is difficult to answer exactly how much data they need, which brings challenge for data collection preparation. The blind request for larger amounts of data exacerbates the severity of data deluge (Baraniuk, 2011). To avoid huge resources waste with aimless data collection, it is important to study the significance of sample size in data analysis. This is also important to develop sustainable transportation system. To verify whether the larger amount of data is conductive to improving the accuracy of analysis, this study analyzes the effect of different sample sizes on the accuracy of estimating household travel mode choice by using the large-scale household mobility survey data collected by Milan municipality, Italy, during 2005–2006. Although the mobility survey data is a little old, it is well suited to investigate the influence of sample size on the accuracy of estimation due to the large scale and quality of the data collected. In addition, data items include both travel characteristics data and socio-economic data, which are important

CONTACT Meng Xu mengxu@bjtu.edu.cn State Key Laboratory of Rail Traffic Control and Safety, Beijing Jiaotong University, Beijing, China.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/ujst.

[1] Memphis Police's Blue CRUSH plan, the data-driven initiative uses information to determine local crime hotspots, which helps law enforcement determine where they need to deploy more officers, and therefore reduced crime by more than 30%. https://www.forbes.com/sites/gregsatell/2013/12/03/yes-big-data-can-solve-real-world-problems/#516e17f38896.

[2] UPS used to replace important parts every few years to ensure that its vehicles stayed in good working order. Now, they collect data from hundreds of sensors in each vehicle. Then algorithms analyze that data from thousands of trucks to predict when a part is likely to break down, allowing UPS to save millions in maintenance costs. https://www.cnet.com/news/ups-turns-data-analysis-into-big-savings/.

[3] Semantria worked with Schwan's frozen foods to evaluate thousands of responses and understand what their customers really thought about them. https://www.forbes.com/sites/gregsatell/2013/12/03/yes-big-data-can-solve-real-world-problems/#516e17f38896.

contributors in analyzing travel mode choice. Three state-of-the-art machine learning methods are applied in this study.[4] This paper aims to verify the efficiency and effectiveness of three methods (i.e., MNL, RF, and SVM) in estimating travel mode choice, as well as to reveal some empirical findings regarding data size and new substantive insights.

There are some fundamental questions in large scale analysis, e.g., given a data set, how much data is enough for achieving reliable estimation? Do all three methods yield the same pattern? Which factor is the most for influencing travel mode choice? When the factors change, what will happen to the household's travel mode choice? To approach these questions, we use a large-scale household mobility survey data collected by Milan municipality Moreover, to promote sustainable development of big data in transportation field, recommendations regarding suited sample size will be given in this study. This case study attempts to alleviate or reduce the burden of mass data on data processing and analysis, and raises the attention on sample size when using data. Moreover, this study could provide Milan municipality with information relating to household travel characteristics for better travel demand management with the gathered large-scale data analysis.

1.2. Literature review

With the rapid growth of data scale, researchers have raised concern about this in empirical studies. Lack of consideration for the sample size could cause substantial bias in empirical study and triggers serious problems in terms of data collection, storage and processing. Firstly, it takes considerable time and resource to collect data, especially for those researches supported by big data. Big data represents the information assets characterized by such a high volume, velocity and variety (Mauro et al., 2016), diversity and scale of big data burden the acquisition of data. Moreover, it is still appear to collect data in the conventional pen-and-paper form, but the painstakingly collected data are not efficiently utilized. The utilization gap (difference between amount of data collection and amount of data usage) is further widened for lacking elaborate plan for data size.[5] Secondly, data storage becomes a growing problem with the advent of big data era.[6] The cost of operation and overall management or integration of big data is time-consuming and costly. The cost of a complete analytic platform to analyze and store the exponential growth data seems prohibitive for some small-scale companies. Besides, research on data storage technology of *DRTDB* is conducted (Yan & Long, 2016), which demonstrates that such strategy has a huge advantage in satisfying the needs of data services. Thirdly, the increasing amount of data puts more pressure on data processing. Data mining has been regarded as an efficient strategy to relieve the pressure. Considerable efforts are devoted to extracting useful transportation data (Cottrill et al., 2013; Zhao et al., 2015) from large datasets by combining statistics and machine learning methods (Manyika et al., 2011). The main work includes the usage of data mining and innovative modeling of data mining. For instance, the spatial and temporal characteristics of travel patterns with large-scale data (Faroqi et al., 2017; Kim & Mahmassani, 2015). Lee et al. (2011) used the raw data of location-based services to discover urban network spatiotemporal traffic bottlenecks. Feng et al. (Feng et al., 2017) proposed the data mining model, which can capture the dynamics of traffic loads to optimize the traffic load balance. Further applications of data mining to transportation problems are reviewed including traffic management, monitoring drowsy drivers and road accidents analysis, etc., as we can refer to literature (Kumarbarai, 2003). Fourthly, data size planning plays a prominent role on the accuracy in parameter estimation approach. Kelley (2007) proposed the approach to validate the estimated coefficients for a variety of scenarios with different sample sizes. Moreover, Kenny and Judd (2019) manifests that variation in the estimated effect is attributed to sampling error alone for multiple studies. Therefore, an in-depth analysis is necessary to verify effect of sample size on improving the accuracy of estimation.

The data used in this paper is derived from the large-scale household mobility survey investigated by Milan municipality during 2005–2006. We apply three machine learning methods to study the effect of sample size on the accuracy of estimating travel mode choice. The literature review will focus on the two facets, i.e., travel mode choice and machine learning.

1.2.1. Travel mode choice

Identification of travel mode choice plays a predominant role on the sustainable development of travel demand management. For instance, prediction of share ratio of travel modes facilitates effective infrastructure investment in eco-friendly travel mode (Pye & Daly, 2015). In general, existing studies mainly focused on the following aspects: (i) Establishment of travel mode choice models. Logit model is the most widely used travel mode choice model, which was introduced by Berkson in 1944 (Berkson, 1944). Several modified logit models emerge in an endless stream and are extensively applied, e.g., conditional logit model, mixed logit model, multinomial logit model (Bhat, 2001; Hensher & Greene, 2003; Mcfadden, 1974; Mohanty & Blanchard, 2016); (ii) Studies regarding the analysis of influencing factors of travel mode choice. For instance, observed and unobserved heterogeneity (individual's intrinsic mode preference) is incorporated in the decision-making of residents' travel mode choice (Bhat, 2000; Gönül & Srinivasan, 1993). Moreover, some objective or subjective determinants are also considered to affect travel mode choice, as indicated in the literature (Frank et al., 2007; Scheiner & Holz-Rau, 2007). It is acknowledged that numerous vital studies make

[4] According to the findings of Hagenauer and Helbich (2017), they compared seven machine learning classifiers for travel mode choice analysis, and we select the best-performing (RF), medium-performing (SVM) and worst-performing (MNL) methods to study the travel characteristics in Milan. Meanwhile, the other purpose of this paper is to study the impact of sample size on the estimation accuracy. Three methods are used to find an objective rule regarding the impact of sample size on the estimation accuracy.

[5] https://www.hottopics.ht/19980/when-is-excess-data-in-marketing-a-bad-thing/.

[6] https://www.gooddata.com/blog/whats-true-cost-big-data.

tremendous contributions to the advance of research on travel mode choice. However, research on travel mode choice supported by large-scale real and reliable data remains great potential development.

As shown in Table 1, we summarize the related studies with respect to country, sample size, explanatory variables, available travel mode, methods and key findings. Compared with the sample size in this study, it is found that the samples sizes are relatively small in the studies listed in Table 1. We consider using machine learning methods to analyze and process the data.

1.2.2. Machine learning

Machine learning can efficiently process fast-growing data (Zang et al., 2014). Innovative methods have proposed to recognize the individual travel mode choice from large-scale mobility survey data, in which several explanatory variables (e.g., travel characteristics) are appreciable to estimate the choices among the travel modes (Omrani, 2015). Several machine learning methods have been relatively successfully applied to solve this kind of problem, e.g., artificial neural network, decision tree, support vector machine, random forest, multinomial logit (Ferri-Garcia et al., 2019; Potoglou & Kanaroglou, 2008; Rasouli & Timmermans, 2014; Shafique & Hato, 2015; Wang & Elhag, 2007; Xie & Lu, 2003; Yamamoto et al., 2002; Yang et al., 1993), etc. Zhang and Xie (2008) claimed that support vector machine (SVM) outperforms multinomial logit (MNL) in terms of prediction of travel mode choice. Besides, Hagenauer and Helbich (2017) compared the predictive performance of seven machine learning classifiers for travel mode choice analysis and make recommendations for model selection. Among machine learning classifiers in these comparative studies, random forest (RF) approach performs best to estimate the travel mode choice. The existing studies put the emphasis on the performance of machine learning methods under the full sample size, however, the reflection on the effect of different sample sizes is lacking.

It is acknowledged that there are many methods that can be used as an estimation tool for the key parameters of demand models, for example a travel mode choice model. These methods could be further categorized as model-based and data-driven approaches (Cottrill et al., 2013; Zhao et al., 2015). For model-based approaches, a key goal is often to show that the improvement of estimation with proposed models. However data-driven approaches are often expected to extract useful information from large-scale data. Their built-in models have also been modified for certain purposes, such as improving operation efficiency or higher estimation accuracy. Overall, both of these methods have merits and could be regarded as effective estimation tools. In this paper, we study not only the accuracy of estimating travel mode choice under the full sample size, but also the impact of the sample size. In terms of research methods, we choose three well-performed machine learning methods, i.e., MNL, SVM and RF, to process large-scale data.

1.3. Contributions

Extending the existing methods on the estimation of travel mode choice and addressing the limitations of previous studies at the same time, this study uses the large-scale household mobility survey data, coupled with machine learning methods to explore the influence of sample size on the estimation of household travel mode choice and the analysis of residents' travel characteristics in Milan. This study aims not only to explore the impact of sample sizes on the accuracy of estimating travel mode choice, but also tends to study travel characteristics in Milan by analyzing the gathered large-scale mobility survey data, which is important to ease the traffic pressure in Milan municipality, as can be found from the 2017 report of INRIX company, where Milan ranked at the tenth of the worst traffic hotspots in Europe with €3.8 billion in travel cost.[7]

The main contributions of this study can be summarized as follows.

1. Using the large-scale mobility survey data to verify whether larger-scale data perform better. The result is helpful to make efficient use of big data and to rationalize the collection and value of large-scale data.
2. Model parameters estimation with three machine learning methods in R software, coupled with statistical verification.
3. The households based travel characteristics investigation in Milan, which is based on the collected large-scale mobility survey data. This is important to relevant policies and strategies implementation in Milan.
4. Key factors influencing household travel mode will be identified and an in-depth sensitivity analysis will be presented along with the use of machine learning algorithms.

The rest of this paper is organized as follows. Section 2 presents an introduction to the three machine learning methods. Section 3 will present processing and analysis of large-scale collected mobility survey data, and the estimation results are shown in Section 4. Further discussion will be presented in Section 5, which includes the relationship between sample size and the accuracy of estimating household travel mode choice using different machine learning methods, the importance ranking of the factors influencing household travel mode choice, and the sensitivity analysis. Conclusions and further study will be shown in Section 6.

2. Methodology

With the proposed problem, this section will introduce the chosen three machine learning methods, i.e., multinomial logit model (MNL), random forest (RF) and support vector machine (SVM). Note that there are five type explanatory

[7] INRIX Reveals Congestion at Germany's Worst Traffic Hotspots to Cost Drivers €48 Billion Over the Next Decade. http://inrix.com/press-releases/inrix-reveals-congestion-at-germanys-worst-traffic-hotspots-to-cost-drivers-e48-billion-over-the-next-decade/.

Table 1. Comparative summary of travel mode choice studies.

Reference	Country	Sample size	Explanatory variables	Number of travel modes	Methods	Key findings
Bhat (2000)	U. S.	520 individuals	Preference Responsiveness	5	Multinomial logit (MNL) Deterministic-coefficients logit (DCL) Random-coefficients logit (RCL)	This paper formulates a multinomial logit based model of travel mode choice that accommodates observed and unobserved taste variations. In the comparison of data fit, they found that the RCL model provides a dramatic improvement in fit relative to the DCL model. The DCL model, in turn, rejects the MNL model based on a likelihood ratio test. The empirical results emphasize the need to accommodate observed and unobserved heterogeneity across individuals in urban mode choice modeling.
	Canada	3,511 households 5,684 individuals	Travel time Travel cost … (17 variables)	3	Activity-based model Tour-based model Chain-based model Trip-based model	The model presented in this paper is a “hybrid”, in which rules are combined with a “classical” random utility maximization decision criterion within an explicit microsimulation framework to model tour-level mode choices. Travel mode is well predicted with prediction success rates in the order of 95%, 75% and 70% for the auto-drive, transit and walk modes, respectively.
Scheiner and Holz-Rau (2007)	Germany	2,691 households	Life situation Lifestyle Residential location Urban form	5	Structural equation modeling	Lifestyles influence mode choice, although just slightly, even when life situation is controlled for. The influence of life situation on mode choice exceeds the influence of lifestyle. The influence of lifestyle and life situation on mode choice is primarily mediated by specific location attitudes and location decisions. Objective spatial conditions as well as subjective location attitudes are important to mode choice.
Frank et al. (2007)	U.S.	14,487 individuals	Travel time Costs Land use patterns	8	Nested logit model	Urban form at residential and employment locations, and travel time and cost were significant predictors of travel choice. Travel time was the strongest predictor of mode choice while urban form the strongest predictor of the number of stops within a tour. Reductions in highway travel time are associated with less transit use and walking.
	Australia	10,500 households	Travel cost In-vehicle time Wait time Walk time	3	Nested logit model	The results show that weekend travel is characterized by a high joint household activity participation rate while weekday travel is distinguished by more intra-household shared ride arrangements. The arrangements of joint household travel are highly associated with travel purpose, social and mobility constraints and household resources. On weekends, public transport is mainly used by captive users and its share is about half of that on weekdays. The value of travel time savings are found to be higher on weekends than on weekdays. This paper highlights the importance of studying joint household travel and using different transport management measures.
Mohanty and Blanchard (2016)	U.S.	58 participants	Sex Age … (10 variables)	4	Multinomial logit (MNL) model Nested-logit model Random taste mixed-logit model	A random taste mixed logit model successfully incorporates heterogeneity among various types of individuals regarding bicycle and pedestrian infrastructure. The results reveal that among the pedestrian infrastructure variables, sidewalk width and presence of pedestrian crossings significantly affected transit use.
	Iran	3,272 observations	Gender Age … (23 variables)	4	Multinomial logit model Regression model	High and low income families, are expected to respectively show 106% and 67% increase in the probability of walking, if their safety concerns are eliminated. They find that the probability of walking is reduced by 0.85% due to a 1% increase in travel distance, it propels parents to select non-active modes, particularly school bus. This study also demonstrates how addressing parental concerns about travel safety could double the propensity to walk to school.

variables (i.e., household size, vehicle ownership, household income, travel distance and travel time) and a response variable (i.e., travel mode choice with respect to public transport, private car, usage of public transport and private car simultaneously and other travel mode) in this study. Instead of general introduction of these methods, we will present the methods with respect to details regarding this study. The complicated operation process of these three methods with the large-scale mobility survey data is conducted in R software and their results will present in Section 4.

2.1. Multinomial logit (MNL) model

The MNL model is based on random utility maximization (RUM) assumption (Srinivasan & Mahmassani, 2005), i.e., decision-makers choose the best alternative for them. In this study, the utility of travel mode is composed by the linear combination of coefficients and explanatory variables. The utility of the ith household with the kth travel mode is formulated by the following two equations.

$$U(k,\ i) = V(k,i) + \varepsilon(k,i),\ \ i = 1,\dots,\ \ N;\ \ k \in K \tag{1}$$

$$V(k,\ i) = \beta_k \cdot {x_i}^T,\ \ i = 1,\dots,N;\ \ k \in K \tag{2}$$

where $U(k,\ i)$ is the total utility of the ith household with the kth travel mode, we assume that $V(k,i)$ is the fixed utility term and $\varepsilon(k,\ i)$ is the error utility term. Further, we assume N is the number of households. The travel mode $k \in K (K = 1,\ 2,\ 3,\ 4)$, where K is the set of travel modes. The values of five explanatory variables of the ith household are $x_i = (x_{1i},\ x_{2i},\ x_{3i},\ x_{4i},\ x_{5i})$, which consists of household size, vehicle ownership, household income, travel distance and travel time. The coefficients of five explanatory variables associated with the kth travel mode are $\beta_k = (\beta_{k1},\ \beta_{k2},\ \beta_{k3},\ \beta_{k4},\ \beta_{k5})$. According to the RUM assumption, the probability of choosing the kth travel mode is given by Equation (3).

$$P(k,i) = Prob\big(U(k,i) > U(k',i), \forall k'(\neq k) \in K\big) \tag{3}$$

where $k'(\neq k) \in K$ represents the other travel mode exclude travel mode k.

The probability of four travel modes chosen by the ith household can be formulated as one of these Equations (4)–(7):

$$P(k = 1 \mid i) \triangleq \pi_1(i) = \frac{e^{U(1,i)}}{1 + e^{U(1,i)} + e^{U(2,i)} + e^{U(3,i)} + e^{U(4,i)}} \tag{4}$$

$$P(k = 2 \mid i) \triangleq \pi_2(i) = \frac{e^{U(2,i)}}{1 + e^{U(1,i)} + e^{U(2,i)} + e^{U(3,i)} + e^{U(4,i)}} \tag{5}$$

$$P(k = 3 \mid i) \triangleq \pi_3(i) = \frac{e^{U(3,i)}}{1 + e^{U(1,i)} + e^{U(2,i)} + e^{U(3,i)} + e^{U(4,i)}} \tag{6}$$

$$P(k = 4 \mid i) \triangleq \pi_4(i) = 1 - \sum_{k=1}^{3} P(k \mid i) = \frac{1}{1 + e^{U(1,i)} + e^{U(2,i)} + e^{U(3,i)} + e^{U(4,i)}} \tag{7}$$

$$\pi_1(i) + \pi_2(i) + \pi_3(i) + \pi_4(i) = 1 \tag{8}$$

where $\pi_k(i)$ represents the probability of the ith household choosing the kth travel mode. The Equation (8) ensures that the sum of probabilities of choosing each travel mode for the ith household is 1. We define that z_k is a binary variable, $z_k = 1$ means the kth travel mode is chosen, otherwise $z_k = 0$. Therefore, the likelihood function of the ith household and all the households can be constructed as Equations (9) and (10).

$$l_i = \pi_1(i)^{z_1} \pi_2(i)^{z_2} \pi_3(i)^{z_3} \pi_4(i)^{z_4} \tag{9}$$

$$l(\beta_k) = l_1 \cdot l_2 \cdot \dots \cdot l_N = \prod_{i=1}^{N} \pi_1(i)^{z_1} \pi_2(i)^{z_2} \pi_3(i)^{z_3} \pi_4(i)^{z_4} \tag{10}$$

In this way, the parameters i.e., $\beta_k = (\beta_{k1}\ \beta_{k2}\ \beta_{k3}\ \beta_{k4}\ \beta_{k5})$ can be calculated by optimizing the likelihood function.

According to the Equations (4)–(8), we can define the following three odds ratio functions $odds_w(i), w = 1,\ 2,\ 3$ as given in Equations (11)–(13).

$$odds_1(i) = \frac{P(k = 1 \mid i)}{P(k = 4 \mid i)} = \frac{\pi_1(i)}{\pi_4(i)} = e^{U(1,i)} \tag{11}$$

$$odds_2(i) = \frac{P(k = 2 \mid i)}{P(k = 4 \mid i)} = \frac{\pi_2(i)}{\pi_4(i)} = e^{U(2,i)} \tag{12}$$

$$odds_3(i) = \frac{P(k = 3 \mid i)}{P(k = 4 \mid i)} = \frac{\pi_3(i)}{\pi_4(i)} = e^{U(3,i)} \tag{13}$$

Take the Equation (11) as example: the first odds ratio function represents the ratio of choosing the first travel mode over choosing the fourth travel mode for the i the household. Equations (11)–(13) present the relationships between the probabilities of the four travel modes. The log odds ratio functions are further formulated as follows:

$$\ln\ odds_1(i) = \ln\ e^{U(1,i)} = U(1,i) = \beta_1 \cdot {x_i}^T + \varepsilon(1,i) \tag{14}$$

$$\ln\ odds_2(i) = \ln\ e^{U(2,i)} = U(2,i) = \beta_2 \cdot {x_i}^T + \varepsilon(2,i) \tag{15}$$

$$\ln\ odds_3(i) = \ln\ e^{U(3,i)} = U(3,i) = \beta_3 \cdot {x_i}^T + \varepsilon(3,i) \tag{16}$$

The parameters of MNL model i.e., β_k and $\varepsilon(k,i)$, $k = 1,2,3$ can be yielded by the "*multinom()*" function in R software and as shown in Section 4 (see Table 8). Moreover, with the "*predict()*" function in R software, we get the estimated travel mode choice for each household and the accuracy of estimating household travel mode choice with MNL model is calculated as Equation (17) and as shown in Section 5:

Table 2. Mobility survey data of ten households.

ID	Household size (HS)	Vehicle ownership	Household income (HI)	Travel distance	Travel time	Travel mode
1	1	1	h_2	s_1	t_2	m_1
2	2	2	h_4	s_4	t_4	m_3
3	4	2	h_3	s_4	t_4	m_3
4	5	1	h_3	s_4	t_4	m_3
5	5	1	h_3	s_1	t_1	m_1
6	2	2	h_4	s_3	t_3	m_3
7	4	2	h_4	s_2	t_2	m_2
8	3	1	h_4	s_1	t_2	m_3
9	1	1	h_2	s_1	t_1	m_1
10	1	0	h_2	s_0	t_0	m_4

Notes:
(i) The sample data of ten households are derived from the original data.
(ii) The values of five explanatory variables in Table 2 represent different household characteristics, which can refer to Table 4 for specific information.

$$Accuracy_{MNL} = \frac{Number\ \ of\ \ households\ \ with\ \ correctly\ \ estimated\ \ travel\ \ mode\ \ choice}{Number\ \ of\ \ all\ \ households} \tag{17}$$

2.2. Random forest (RF)

Problems with respect to the identification of household travel mode choices can be also regarded as classification problem, which can be solved by random forest approach. Random forest operates to build a multitude of decision trees at training time first and then chooses the optimal classification (i.e., household travel mode in this study) amongst all the decision trees (DT) as the final result. We first illustrate the DT considering its role in RF, and take the mobility survey data of ten households as example to illustrate the process of building a decision tree. The mobility survey data of ten households are shown in Table 2.

According to household characteristics and travel characteristics as shown in Table 2, we build a decision tree as shown in Figure 1.

Given the decision tree in Figure 1, the root nodes represent ten households. Each branch represents household characteristics or classification rules, which is associated with a particular class label (e.g., HS $\geq$ 3), the final outcome of a decision path with respect to household travel mode is found at the end node. The accuracy of this decision tree in estimating household travel mode is marked as $Accuracy_{DT}$ according to Equation (17).

RF consists of several decision trees, and the operational process of random forest is shown in Figure 2.

The main steps to build the random forest can be summarized as following:

- Generate random subset for each DT with the bootstrap sampling method (i.e., random sampling with replacement).
- Build each DT based on each random subset. Each DT is used to produce the response variable (i.e., travel mode choice in this study) when given the explanatory variables (i.e., household size, vehicle ownership, household income, travel distance and travel time).
- Identify the optimal classification result. Final result of household travel mode choice depends on the optimal classification result among all the DTs (the total amount of DT is n).

The accuracy of estimating household travel mode choice with RF approach is given as Equation (18). Note that in RF, each DT is built using a different bootstrap sample from the original data. About one-third of the data are left out of the bootstrap sample and not used in the construction of the decision tree.[8] Put the data left out in the construction of the DT to get a classification, and the minimum estimation error of these partial data is called Out-of-bag (OOB) error.

$$Auucracy_{RF} = \max_{n} Accuracy_{DT_n} = 100\% - \min_{n} OOB\ \ error_n \tag{18}$$

Results including number of DTs i.e., n, confusion matrix and accuracy of estimating household travel mode choice are presented in Section 4 (see Table 9). Note that confusion matrix is a special kind of contingency table, with two dimensions (“original” and “estimated”), it lists the numbers of households whose travel modes to be correctly estimated as original travel modes and incorrectly estimated as the other three travel modes.

2.3. Support vector machine (SVM)

SVM is a typical high-performing machine learning algorithms since it was developed in the 1990s (Cortes & Vapnik, 1995). For the complex classification problems (i.e., multi-class classification problem in high-dimensional space), SVM performs classification tasks by constructing optimal hyperplanes that separate different classes in a high-dimensional space (Boser, 1992). However, for the simple classification problem (e.g., binary classification problem in one-dimensional or two-dimensional space), SVM attempts to find an optimal line to separate the classes in the low-dimensional space.

As shown in Table 3, we use $X_i = (X_{i1},\ X_{i2})$ to located the ith household in the two-dimensional space, where X_{i1} and X_{i2} are vehicle ownership and household size respectively for the ith household. Moreover, we use squares and triangles to represent PT and private car separately. Therefore, the ith household can be represented in a two-dimensional space, i.e., $X_i = (X_{i1},\ X_{i2})$, as shown in Figure 3(a). To separate two travel modes, it is found that several lines can separate them as we can refer to Figure 3(b). The optimal line is as far as

[8] https://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm.

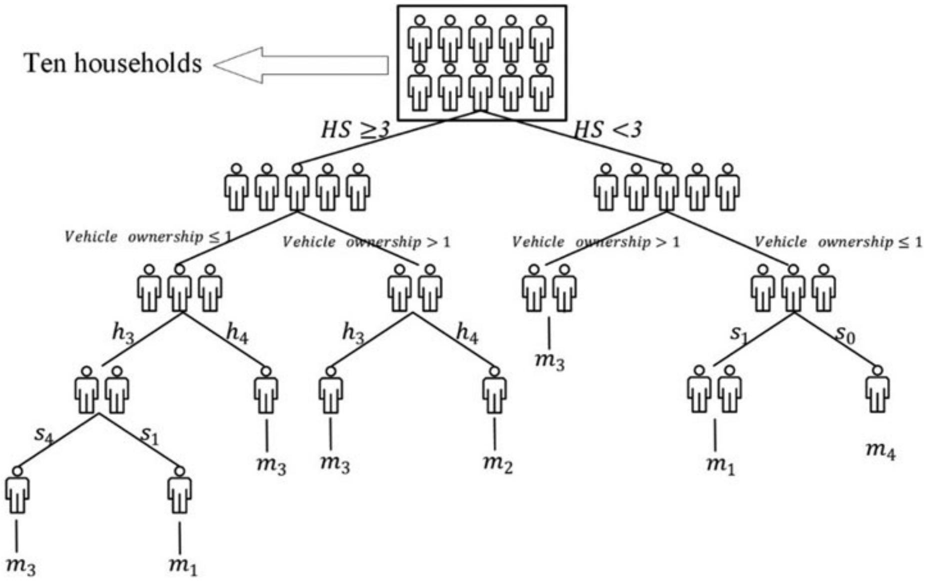

Figure 1. DT development process w.r.t. the mobility survey data of ten households.

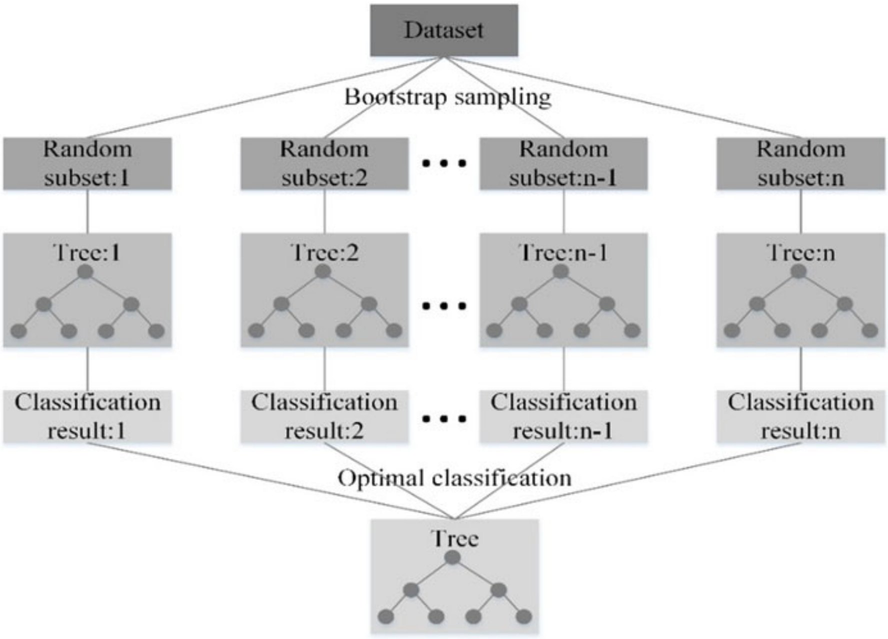

Figure 2. The operational process of random forest.

Table 3. The sample data of ten households.

ID	Vehicle ownership	Household size (HS)	Travel mode
1	1	5	PT
2	2	4	Private car
3	3	1	Private car
4	1	4	PT
5	1	1	Private car
6	0	1	PT
7	2	1	Private car
8	1	3	PT
9	1	2	Private car
10	0	2	PT
11	1	1	PT
12	2	3	Private car

Note: The simplified data is derived from the original data.

possible away from the data point of each travel mode like the solid line shown in Figure 3(c).

The line in the two-dimensional space can be defined as $\omega_1 X_1 + \omega_2 X_2 + b = 0$. Therefore, we can calculate the distance between the line and the closest data point of each travel mode. The double of that distance is defined as the margin. The larger the margin, the better the line classifies the data. The maximum margin can be calculated as Equation (19),

$$maxmargin = 2\frac{\omega_1 X_{i1} + \omega_2 X_{i2} + b}{\sqrt{\omega_1^2 + \omega_2^2}} \tag{19}$$

Therefore, the line whose margin is the largest is the optimal line. According to sample data of first six households, we can derive the optimal line i.e., $-6X_1 + 2X_2 + 1 = 0$

However, for the irregular and messy data point (e.g., the sample data of the last six households in Table 3) shown in

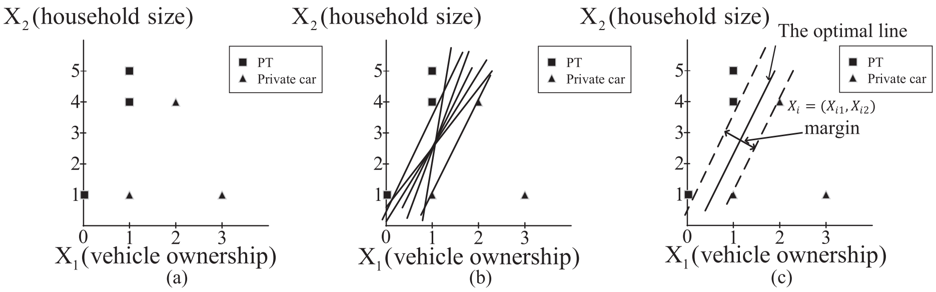

Figure 3. Illustration of classifying the first six data by SVM approach.

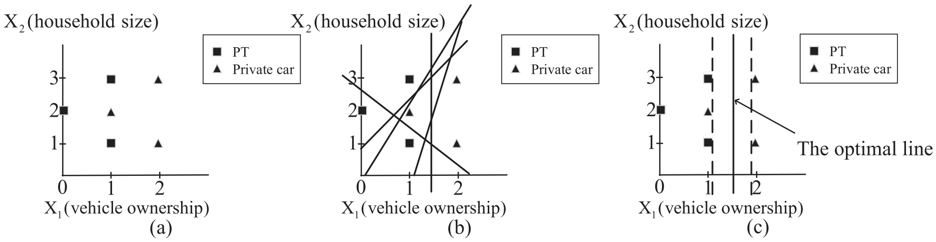

Figure 4. Illustration of classifying the last six data by SVM approach.

Figure 4(a), it is impossible to find a line which completely separates the two travel modes, as we can refer to Figure 4(b). Hence, we introduce a slack variable ξ_i. That is, we tolerate some misclassified data points. Therefore, objective function is revised as Equation (20). Amongst all the lines, the line which tolerates one misclassified data point is the optimal one and is shown as Figure 4(c).

$$max\ 2\frac{\omega_1 X_{i1} + \omega_2 X_{i2} + b}{\sqrt{\omega_1^2 + \omega_2^2}} + \mathsf{C}\sum_{i=1}^{6}\xi_i \tag{20}$$

where, C is the parameter in SVM and represents the tolerance of error.

A key feature of SVM is that it can map the data points into a new space using a kernel function. In doing so, a nonlinear classification problem is turned linear. The radial basis function (RBF) performs well on many types of data and is thought to be reasonable for many classification tasks. RBF is defined as Equation (21),

$$\begin{aligned} K(X_i,\ X_j) &= \exp\left(-\gamma\|X_i - X_j\|^2\right) \\ &= \exp\left(-\gamma\|X_i\|^2 - \gamma\|X_j\|^2 + 2\gamma X_i{}^T X_j\right) \end{aligned} \tag{21}$$

where $\|X_i\|^2$ is the squared Euclidean distance,[9] $\|X_i - X_j\|^2$ is the squared Euclidean distance between two data points X_i and X_j. With RBF, the original data are mapped into a new space. The parameter γ controls the number of support vectors (the nearest data points to the lines separating different travel modes are called support vectors). We choose four different values of γ to estimate household travel modes and the result is shown in Section 4.3 (see Table 10).

The classification problem in this study is complex because it is a multi-class classification problem in a five-dimensional space. For the multi-class characteristic, we transform the multi-class classification problem into a series of binary classification problems. Each binary classification problem is to separate one travel mode from the three other travel modes. Hence, a 4-class classification problem equals four binary classification problems. For the high-dimensional characteristic, we use $X_i = (X_{i1},\ X_{i2},\ X_{i3},\ X_{i4},\ X_{i5})$ to locate the ith household in the five-dimensional space. The classification task is to find the optimal hyperplane which maximizes the margin, as shown in Equation (22),

$$max\ 2\frac{\omega_1 X_{i1} + \omega_2 X_{i2} + \omega_3 X_{i3} + \omega_4 X_{i4} + \omega_5 X_{i5} + b}{\sqrt{\omega_1^2 + \omega_2^2 + \omega_3^2 + \omega_4^2 + \omega_5^2}} + \mathsf{C}\sum_{i=1}^{N}\xi_i \tag{22}$$

With the "*e1071*" package and "*svm()*" function in R software, the large-scale household travel modes data can be classified. The accuracy of estimating household travel mode with SVM is also calculated according to Equation (17).

3. Data

The sample data is almost 10% of the population (about 2.4 million) living in Milan metropolitan area, which was divided into 610 zones with respect to population density. The research area is shown in Figure 5. The survey was based on household (about 134,000 interviews with about 240,000 persons over 11 years interviewed), which involved vehicle drivers and users of public transport (PT) crossing the orbital highway road and railway around the city of Milan (quite close to Milan municipality boundary). The household survey collected all trips made by all household members in a typical weekday. Data are stored in a relational database of 20 archives linked together by keys. Tables contain data of households and of every household member, all trip chains and the start and end time, scope, mode, origin and destination of trips. Based on the mobility survey data,

[9] Euclidean distance: $\|X_i\| = \sqrt{X_{i1}{}^2 + X_{i2}{}^2}$.

we select data items related to the mobility of household, including household size (HS), vehicle ownership, household income (HI), daily travel distance (km) and travel time (minutes) by public transport and private car.

3.1. Data processing

The data are processed first to obtain the sample for this study, and some raw data are initially removed:

- Households without daily travel time but with daily travel distance. The minimum unit of the travel time is 1 minute, so we don't take the households whose travel time is less than 1 minute into account. Similarly, the households whose daily travel distance is zero, but with daily travel time, are removed due to the minimum unit of the travel distance is 0.01km. Hence, 27 such items are removed from data set.
- Data with private car travel distance or travel time but without vehicle ownership are removed. There are 500 records removed from the data set. These data could happen when vehicles are rented/borrowed. The rationale for removing this part of data is that we do not consider the renting/borrowing vehicle behavior in this study.

3.2. Data analysis

The distribution of variables in the sample data is presented in Table 4.

As can be seen from Table 4, the majority of household size is within 1–5 persons and a small percentage (only 0.73%) has household size over 5 persons. The largest HS is 10 persons. For the convenience of data processing, we adopt a single category for HS > 5 in this study, the vehicle ownership is classified in a similar way. We then sort the sample into four categories as follows according to their income level: temporary group is defined with zero HI, whilst the HI for a low income group, middle income group and high income group are set as [€944.6, €27011.4], (€27011.4, €61856.3] and (€61856.3, €342066.0] separately, which is approximately consistent with the income level in Europe.[10] According to the distribution of travel distance and time, we divide the sample into 5 groups. When it comes to the household travel mode choice, about 35% of households are on public transport (PT), only 18.57% households travel by private car, and 33.25% of households use the two travel modes simultaneously. Obviously, there are some households who traveled by others travel modes (e.g., walk).

3.3. Description of variables

Some of the explanatory variables are categorical variables, such as HI, travel distance and travel time, the symbolic specifications of the variables are presented in Table 5.

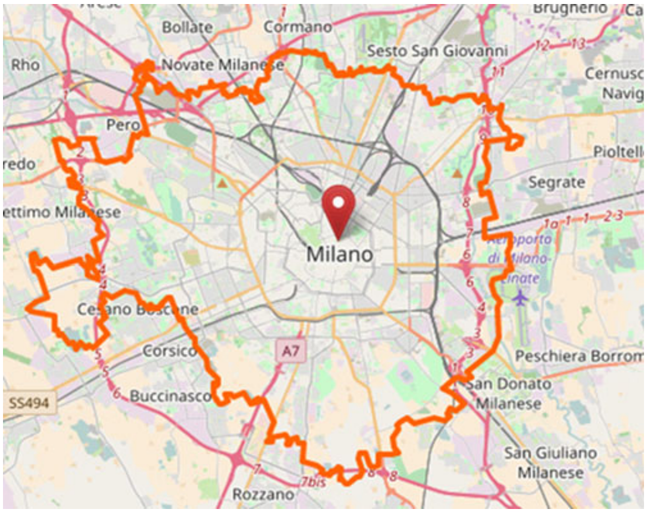

Figure 5. Mobility survey area in Milan (The orange area).

3.4. Data structure

The data framework of the investigation sample is listed in Table 6, which includes sample size, variable name, variable type, sequence of ordinal variables, and some sample data.

The data processing is based on R software (3.4.1) with computer system: Intel CoreTM i7-6700 CPU of 3.4 GHz and 8 GB RAM. There are totally 101,053 households with 6 different variables, five of them are ordinal variables and travel mode is the only unordered categorical variable. The response variable is travel mode, and the explanatory variables are HS, vehicle ownership, HI, travel distance and travel time.

3.5. Correlation test

We first study the correlation between the five explanatory variables. If the correlation between two variables is too strong, we need to transform the two correlated variables into one variable by dimensionality reduction. Table 7 displays the correlations of explanatory variables as calculated in R software.

Taylor (1990) introduced the threshold value of correlation coefficients in detail. That is, the correlation coefficients which are lower than 0.35 are generally considered to represent low or weak correlations, the correlation coefficients range from 0.36 to 0.67 represent the modest or moderate correlations, and 0.68 to 1.0 are strong or high correlations, higher than 0.90 are very high correlations. According to Table 7, the correlation between explanatory variables is weak with the correlation coefficients ranging from 0.2301 to 0.5728. The correlation between household income and other variables seems not strong (correlation coefficients are all lower than 0.68, whilst the correlation between HI and travel distance is the lowest). In contrast to some existing studies (e.g., Mallett, 2001), which shows people with higher income

[10] https://en.wikipedia.org/wiki/List_of_European_countries_by_average_wage.

Table 4. Statistics of household characteristics.

Household characteristics	Proportion	Travel characteristics	Proportion
Household size (HS)		**Travel distance (km)**	
1	28.63%	0	13.16%
2	33.90%	(0,10]	25.94%
3	20.49%	(10,20]	19.95%
4	13.61%	(20,30]	13.81%
5	2.64%	>30	27.14%
> =6	0.73%		
Vehicle ownership (cars and motorcycles)		**Travel time (minute)**	
0	26.75%	0	13.16%
1	41.74%	(0,60]	22.66%
2	23.22%	(60,120]	23.50%
3	6.36%	(120,180]	16.41%
4	1.52%	>180	24.26%
> =5	0.41%		
Household income (HI)		**Travel mode**	
Temporary group	6.88%	Public transport (PT)	35.03%
Low income group	31.95%	Private car	18.57%
Middle income group	40.34%	Usage of PT and private car simultaneously	33.25%
High income group	20.83%	Other	13.15%

Table 5. Variables description.

Variables			Description of the variables
Explanatory variables	Household characteristics	HS (person)	1: 1 person 2: 2 persons 3: 3 persons 4: 4 persons 5: 5 persons 6: no less than 6 persons
		Vehicle ownership (vehicle)	0: no vehicle 1: 1 vehicle 2: 2 vehicles 3: 3 vehicles 4: 4 vehicles 5: no less than 5 vehicles
		HI (€)	h_1: temporary group (0) h_2: low income group [944.6, 27011.4] h_3: middle income group (27011.4, 61856.3] h_4: high income group (61856.3, 342066.0]
Response variable	Travel characteristics	Travel distance (km)	s_0: 0 s_1: (0,10] s_2: (10,20] s_3: (20,30] s_4: over 30
		Travel time (minute)	t_0: 0 t_1: (0,60] t_2: (60,120] t_3: (120,180] t_4: over 180
		Travel mode	m_1: PT m_2: private car m_3: usage of PT and private car simultaneously m_4: the others travel modes such as walk

Notes: According to the data set, travel distance, travel time and household travel mode are the daily travel characteristics of each household, rather than in a single journey. For the third travel mode m$_3$, it refers to the travel mode choice of households who does not only use PT, but also use private car to travel in a day (i.e., usage of PT and private car simultaneously).

Table 6. Data structure in software R.

101053 households of 6 variables:
HS (integer): "1"<"2"<"3"<"4"< …: 1 2 4 5 5 2 4 …
Vehicle ownership (integer): "0"<"1"<"2"<"3"< …: 1 2 2 1 1 2 2 …
HI (4-level factor): "h_1"<"h_2"<"h_3"<"h_4": h_1 h_3 h_2 h_2 h_2 h_3 h_3 …
Travel distance (5-level factor): "s_1"<"s_2"<"s_3"<"s_4"< …: s_3 s_4 s_4 s_4 s_1 s_3 s_2 …
Travel time (5-level factor): "t_1"<"t_2"<"t_3"<"t_4"< …: t_2 t_4 t_4 t_4 t_1 t_3 t_2 …
Travel mode (4-level factor): "m_1", "m_2", "m_3", "m_4"

travel longer distance and time, this case study shows that HI has little influence on travel characteristics in Milan. Therefore, explanatory variables generally can be considered as weakly, modestly or moderately correlated in this study, we will retain these five variables in the following analyses.

4. Estimated results

4.1. Multinomial logit (MNL) model

The fourth household travel mode choice ("the others travel modes") is assumed as the benchmark mode. We then get the coefficients with standard errors, residual deviance and the Akaike information criterion (AIC) in MNL model by the multinomial package and statement "*summary()*" in R software, as given in Table 8.

As we can see from the Table 8, the coefficient of each explanatory variable represents the effect of explanatory variable on response variable. For instance, with the per-unit increase of HS, the log ratio of choosing PT over choosing the fourth travel mode approximately decreases by 0.457 (the black shadow figure in Table 8) under the same other variables condition. Moreover, the log ratio of choosing private car over choosing the fourth travel mode approximately increases by 0.158 (the red shadow figure in Table 8) when the household income changes from temporary group to low income group.

Table 7. Correlation coefficients between explanatory variables.

	HS	Vehicle ownership	HI	Travel distance	Travel time
HS	1.0000				
Vehicle ownership	0.5266	1.0000			
HI	0.4789	0.4912	1.0000		
Travel distance	0.2728	0.2956	0.2301	1.0000	
Travel time	0.5010	0.4084	0.3781	0.5728	1.0000

The standard errors of each coefficient are also shown in Table 8. It ranges from 0.0061 to 0.03. The index of residual deviance is a quality-of-fit statistic, which is achieved by the square of log likelihood (the log likelihood of the estimation result in this case is about 441). AIC is the index of relative quality in MNL models. Both the values of residual deviance and AIC are smaller, goodness of fit of proposed model is better. In general, the overall error is relatively acceptable considering that the sample size is more than 100,000.

4.2. Random forest (RF) approach

With the "*randomForest*" package and statement "*fit.forest*" in R software, the estimated results with RF approach can be shown in Table 9.

The results include number of decision trees, OOB error and confusion matrix. OOB error describes the estimation error of random forest. Confusion matrix shows the estimation of each household travel mode choice, e.g., For public

Table 8. Estimated results with multinomial logit (MNL) model.

	Travel mode (m_4)	Travel mode (m_1)	Travel mode (m_2)	Travel mode (m_3)
Intercept/standard error	Reference category	−2.3763001/0.05486624***	−3.6053644/0.05407460***	0.6751118/0.04778925***
HS/standard error		**−0.4571329**/0.01298290***	0.3706445/0.01001634***	−0.5713096/0.01746211***
Vehicle ownership/standard error		2.4975662/0.01950788***	1.9834522/0.01765356***	−0.4096975/0.02213585***
HI (h_1)/standard error			Reference category	
HI (h_2)/standard error		0.1218320/0.04430483***	0.1584522/0.04473498***	0.4669645/0.03793123***
HI (h_3)/standard error		0.06551422/0.04359021***	0.46340516/0.04277421***	0.16466768/0.04171280***
HI (h_4)/standard error		−0.4625822/0.05035913***	0.4479926/0.04645392***	−0.4298684/0.06638551***
Travel distance (s_0)/standard error			Reference category	
Travel distance (s_1)/standard error		−0.4592964/0.03058874***	−0.2917884/0.02748171***	−0.7480374/0.02997038***
Travel distance (s_2)/standard error		−0.15303493/0.02708626***	−0.02931298/0.02425415***	−0.57298628/0.02664858***
Travel distance (s_3)/standard error		0.1291548/0.03029620***	0.1420425/0.02727281***	−0.3537290/0.03080436***
Travel distance (s_4)/standard error		0.4438310/0.02907700***	0.3633251/0.02636008***	−0.1961039/0.02982816***
Travel time (t_0)/standard error			Reference category	
Travel time (t_1)/standard error		0.3147622/0.03172176***	0.1574518/0.02895096***	−0.3713059/0.03183055***
Travel time (t_2)/standard error		0.08121504/0.02567323***	0.09487610/0.02333411***	−0.41654750/0.02523419***
Travel time (t_3)/standard error		−0.127517160/0.02904063***	0.005513296/0.02598298***	−0.513332950/0.02901352***
Travel time (t_4)/standard error		−0.30780564/0.03025729***	−0.07357492/0.02687856***	−0.56967020/0.03070211***
Residual deviance			194760.6	
AIC			194838.6	

Note:
(i) We calculate p-value using Wald test (here z-test).
(ii) We use p-value to determine whether the estimated parameters are statistically significant. The smaller p-value means the estimated parameter is more statistically significant.
(iii) The p-value less than 0.001 is represented by three asterisks (***).
(iv) As can be seen from the Table 8, all of the estimated parameters are statistically significant.

Table 9. Estimated results with RF approach.

Number of DTs (n): 500					
Out-of-bag (OOB) error: 28.11%					
Confusion matrix:	Estimated travel mode				
Original travel mode	Travel mode (m_1)	Travel mode (m_2)	Travel mode (m_3)	Travel mode (m_4)	Accuracy
Travel mode (m_1)	24466	2508	6408	2005	69.14%
Travel mode (m_2)	3476	10747	1916	2620	57.29%
Travel mode (m_3)	1584	1698	29256	1071	87.05%
Travel mode (m_4)	1098	2236	1789	8175	61.48%

Table 10. Estimated results with support vector machine (SVM).

Kernel function	Parameters	Confusion matrix					Accuracy
Radial basis function (RBF)			Estimated				
		Original	m_1	m_2	m_3	m_4	
	$\gamma = 0.01$	m_1	25294	4034	2820	0	71.87%
		m_2	2296	4921	1672	0	
		m_3	7797	9804	29117	0	
		m_4	0	0	0	13298	
			Estimated				
		Original	m_1	m_2	m_3	m_4	
	$\gamma = 0.1$	m_1	24590	3669	1966	0	72.06%
		m_2	2217	4835	1548	0	
		m_3	8580	10255	30095	0	
		m_4	0	0	0	13298	
			Estimated				
		Original	m_1	m_2	m_3	m_4	
	$\gamma = 1$	m_1	24742	3475	2190	0	72.74%
		m_2	2555	6382	2330	0	
		m_3	8090	8902	29089	0	
		m_4	0	0	0	13298	
			Estimated				
		Original	m_1	m_2	m_3	m_4	
	$\gamma = 10$	m_1	24869	3609	2222	0	72.74%
		m_2	2387	6151	2194	0	
		m_3	8131	8999	29193	0	
		m_4	0	0	0	13298	

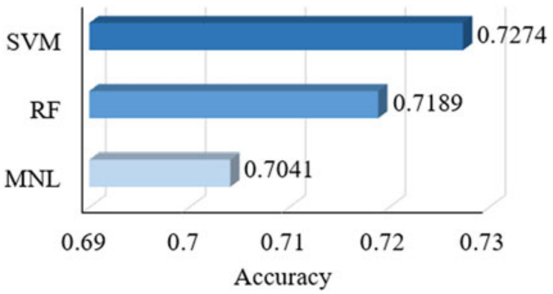

Figure 6. Accuracy in estimating household travel mode choice with MNL, RF and SVM.

transport (PT, m_1), the number of correct estimation is 24466, the number of misestimating PT for private car is 2508, and 6408 is the number of misestimating PT for the third travel mode choice (m_3), therefore, the error rate is 0.3086.

4.3. Support vector machine (SVM)

The two important parameters in SVM algorithm are C and γ. The larger value of C means the low tolerance of error, it could lead to overfitting. However, if the value of C is too small, it means high tolerance of error, it could lead to low accuracy of classification. Therefore, we choose the relative modest value (i.e., $C = 10$).

In this study, we choose the radial basis function (RBF), which is a better choice because it is a nonlinear kernel function and is good at dealing with the nonlinear relationship between respond variable and explanatory variables. Moreover, γ influences the number of support vectors. The larger value of γ means the more number of support vectors. Therefore, we select four values of γ to estimate the household travel modes. The estimated results are shown in Table 10.

According to Table 10, the confusion matrix changes with γ. The accuracy of estimating household travel mode choice increases until γ increases to 1. The maximum accuracy of estimating travel mode choice with SVM is 72.74% when γ equals to 1 or 10.

5. Discussions

5.1. Comparison

We compare the accuracies of these three methods in estimating household travel mode choice, which are based on the mobility survey data in Milan. Under the full sample size, the accuracies with three methods are shown in Figure 6. Whilst there isn't a high degree of difference in accuracy between models, it does show that the accuracy of SVM is clearly the best (72.74%) among the three methods, followed by RF (71.89%) and MNL (70.41%). Further, we investigate the effect of these three methods under various sample size, and nine experiments have been implemented. The accuracies in estimating household travel mode choice under different sample size with these three methods are shown in Figure 7. The sequence of sample data in each experiment is random.

It can be found that the estimation accuracies with these three methods fluctuate in the beginning when the sample size is approximately less than 20,000. That is, the estimation of household travel mode is inaccurate when the sample size is small. Model accuracies gradually stabilize with the increase of the sample size. However, there is no obvious improvement of the model accuracies when we continue to increase the sample size.

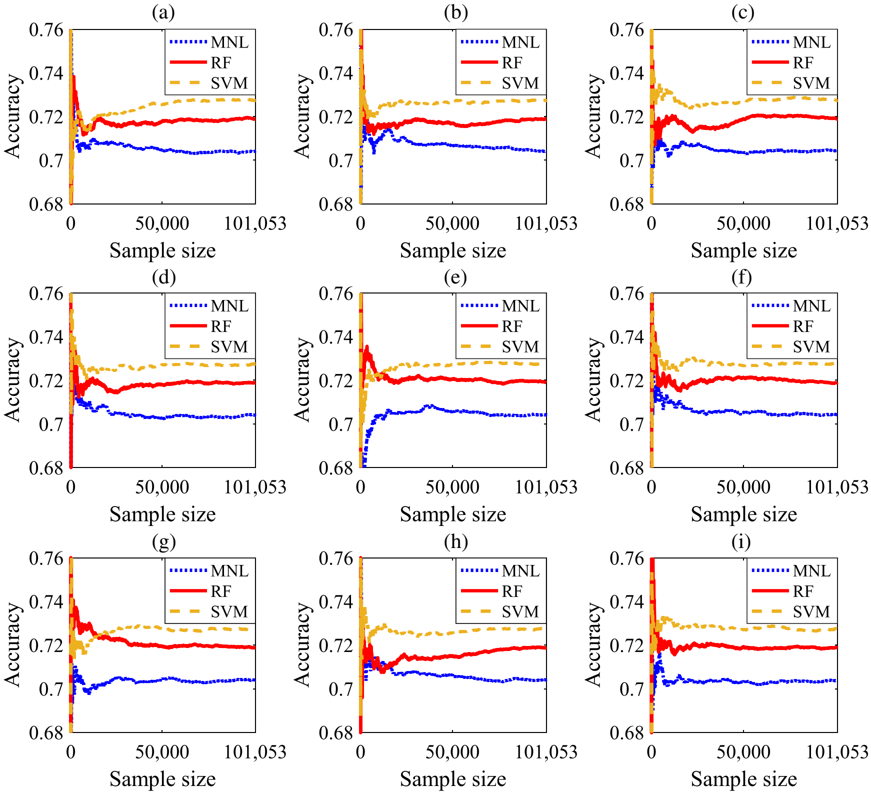

Figure 7. Accuracies of the three methods in estimating household travel mode choice with variant sample size.

5.2. Importance ranking of explanatory variables

We further study the importance ranking of explanatory variables influencing household travel mode choice. Based on the sorting function of RF approach, we acquire the importance ranking of explanatory variables and the result is verified by MNL model. Although the MNL model doesn't have similar function of importance ranking, we compare the accuracy of each model formulated by reducing explanatory variables and then we acquire the order of variables' importance. In that case, the defect of MNL model in identifying the importance of explanatory variables can be remedied. In this study, there are five explanatory variables including HS, vehicle ownership, HI, travel distance and travel time, hence the accuracies of thirty-one MNL models considering different variables through the permutation and combination need to be evaluated.

Based on the function "*the importance()*" in "*randomForest*" package in R software, we acquire the importance ranking of explanatory variables influencing household travel mode choice, as shown in Figure 8.

The index of mean decrease accuracy (i.e., Mean Decrese Accuracy in Figure 8, MDA for short) describes the degree of reduced accuracy of RF approach when a variable becomes a random number, and the larger of MDA value indicates that the variable is more important. The index of mean decrease Gini (i.e., MeanDecreaseGini in Figure 8, MDG for short) measures the impurity of explanatory variables. The larger of MDG value means the purer of the variable. According to the new method proposed based on random forest (RF) to rank the variables using MDA and MDG (Han et al., 2017), the rank of explanatory variables is vehicle ownership, travel distance, travel time, HS, HI according to their importance.

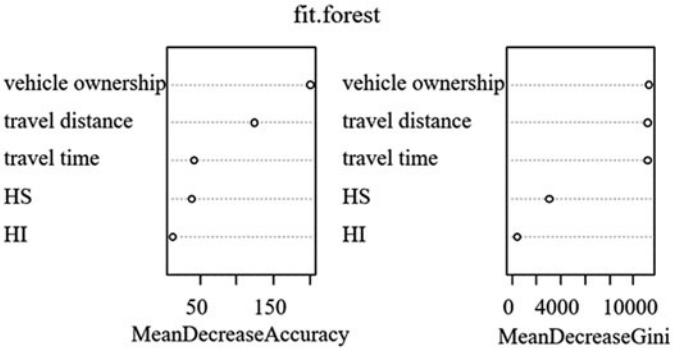

Figure 8. Variable importance used in the RF approach.

In terms of MNL model, there are five explanatory variables including HS, vehicle ownership, HI, travel distance and travel time, so we conduct thirty-one experiments to study the accuracies of these MNL models composed of different variables. The following graph shows the accuracies of each MNL model considering different variables. Note that the accuracy of MNL model with all explanatory variables is 70.41%, we just show the results of thirty MNL models in Figure 9.

As indicated in the Figure 9, each graph displays the accuracies of models with different explanatory variables. When we only take one explanatory variable into account, the accuracy of the MNL model considering vehicle ownership is the highest, when we consider two explanatory variables, the accuracy of the model composed by the vehicle ownership and travel distance is the highest. Furthermore, the accuracy of the model composed by vehicle ownership, travel distance and travel time is the highest. We find that the HI has a little influence on household travel mode choice by the model considering four factors.

To sum up, the importance ranking of explanatory variables by judging the accuracies of the MNL models composed by different explanatory variables are vehicle ownership, travel distance, travel time, HS and HI, which is consistent with the RF approach.

5.3. Influencing factors analysis

Based on the importance ranking of explanatory variables in Section 5.2, we identify that the vehicle ownership and travel distance are two paramount explanatory variables influencing household travel mode choice. In this section, we discuss the probability of the household travel mode choice influenced by these two explanatory variables in detail. A sensitivity analysis is conducted to illustrate the effect of these two explanatory variables influencing household travel mode choice.

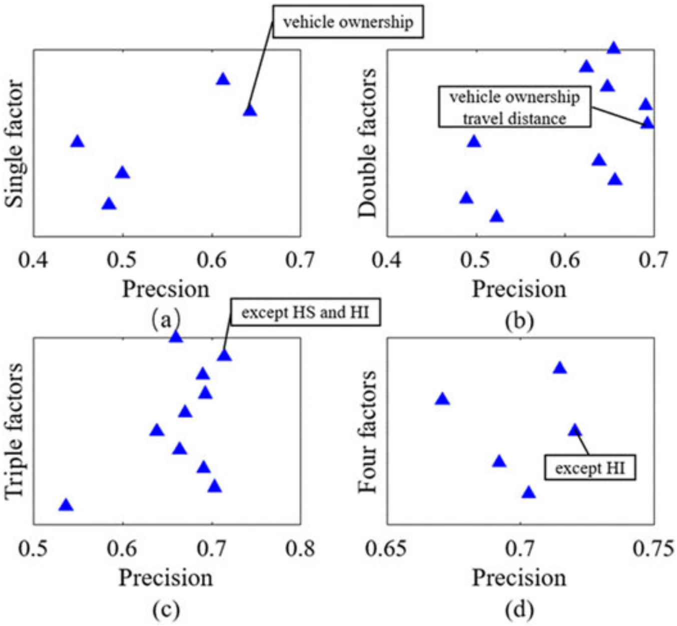

Figure 9. Accuracies of MNL models in estimating household travel mode choice considering different variables.

In this study, the vehicle ownership ranges from 0 to 4, and travel distance is divided into four groups (i.e., s_1, s_2, s_3, s_4, as we can refer to the variable description in Table 2). We focus on the discussion of the first three travel mode choices (i.e., PT, private car and usage of PT and private car simultaneously represented by m_1, m_2, m_3). The relationship between probability of household travel mode choice and these two explanatory variables i.e., vehicle ownership and travel distance is shown as Figure 10.

As shown in Figure 10, for the public transport-PT (m_1), which is displayed at the top of the figure, the probability of choosing this travel mode decreases with the increase of vehicle ownership. It implies that there is a transition from public transport to private car for households when they have more cars. For instance, when the households with short travel distance (s_1) changes from having no car to having one car, the probability of choosing public transport decreases, whereas the possibility of choosing private car and usage of the two travel modes increase. As a matter of fact, the household with different travel distance do the same with such household mentioned above. However, for the second travel mode-private car (m_2), the household with high travel distance (s_4) will increase the probability of using private car when the vehicle ownership increases to 3, but it will decrease even if they own more cars. That is because they may use the PT and private car simultaneously or the others travel modes, it can be found that the probability of the third travel mode increases. The sum of the probability of these three travel modes chosen by single household is 1.

To sum up, with the increase of vehicle ownership, the probability of choosing the public transport will decrease, most households will transit their travel mode to combined travel mode (i.e., usage of PT and private car simultaneously). Supporting that the vehicle ownership is more than 3, many households does not take the usage of private car for granted, they may decrease the probability of using private car solely and try to use public transport and private car together.

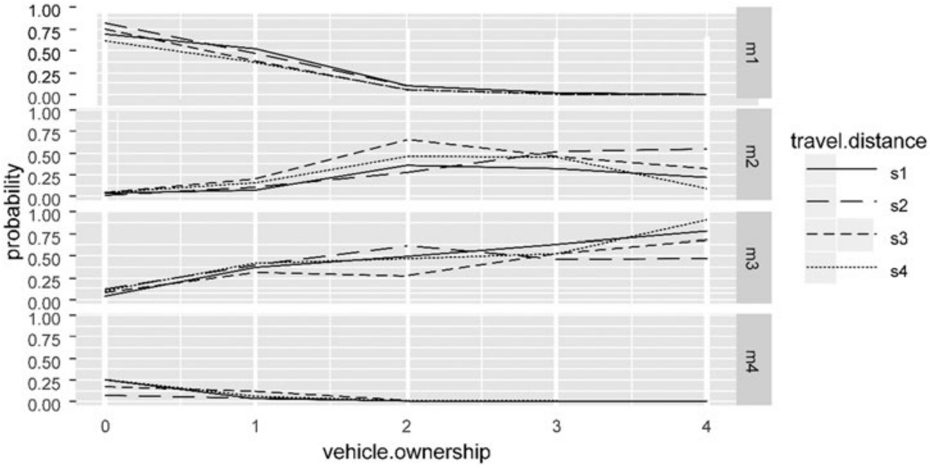

Figure 10. Effect of vehicle ownership and travel distance on household travel mode choice.

6. Conclusions

In the light of the rapid development of a big data agenda, this paper focuses on studying the effect of sample size on the accuracy of estimating household travel mode choice. As a distinction from some existing studies regarding travel mode choice, we use the large-scale household mobility survey data conducted by Milan metropolitan during 2005–2006 and put the emphasis on the study of sample size influencing the accuracy of estimation, coupled with the research on travel characteristics in Milan.

Firstly, the accuracies of these three methods (i.e., MNL, RF, SVM) fluctuate unsteadily with small sample and gradually stabilizes with the increase of sample size. However, when we continue to increase the sample size, the accuracies of these three methods are not increasing. With moderate sample size i.e., about 20,000, these methods can obtain the acceptable accuracies in estimating household travel mode choice.

Secondly, the accuracies of MNL, RF and SVM approaches in estimating household travel mode choice are 70.41%, 71.89% and 72.74% respectively under the full sample size, which shows that the SVM approach performs better.

Thirdly, according to the influence of deleted explanatory variables on the accuracies of MNL models, we get the importance ranking of explanatory variables, which replenishes the MNL model under the lacking the function of importance ranking. And the result is consistent with RF approach.

Fourthly, we identify the critical factor which influences household travel mode choice and rank the explanatory variables according to their importance. The result is that vehicle ownership is the most important factor which influences household travel mode choice, followed by travel distance, travel time, HS and HI.

Finally, we analyze the change of the probability of household travel mode choice with the change of explanatory variables including vehicle ownership and travel distance. The regularity of household travel mode choice is clear that the probability of the household choosing public transport decreases when the vehicle ownership increases, and the probability of household choosing private car or combined travel mode (i.e., usage of PT and private car simultaneously) is increasing when the vehicle ownership increases to 2. However, when the vehicle ownership increases from 3 to 4, the probability of household choosing private car decreases, they transform to use the third travel mode (i.e., usage of PT and private car simultaneously).

For future study, more explanatory variables will be considered to analyze the household travel mode choice. In this paper, the explanatory variables are limited, and an excess of data with the same explanatory variables, but with different travel mode choice, which reduces the accuracy of estimating household travel mode choice. Furthermore, the reason why larger volume of data will not increase the accuracy of estimating household travel mode choice should be studied in depth.

Acknowledgement

Thanks are due to the Mobility, Environment, Territory Agency of Milan (AMAT-MI) who provided mobility data for the research.

Funding

The work described in this paper was supported by the National Natural Science Foundation of China (71771015, 71890971 and 71961137005).

ORCID

Meng Xu http://orcid.org/0000-0003-4738-928X

References

Baraniuk, R. G. (2011). More is less: Signal processing and the data deluge. *Science*, *331*(6018), 717–719. doi:10.1126/science.1197448

Berkson, J. (1944). Application of the logistic function to bio-assay. *Journal of the American Statistical Association*, *39*(227), 357–365. doi:10.1080/01621459.1944.10500699

Bhat, C. R. (2000). Incorporating observed and unobserved heterogeneity in urban work travel mode choice modeling. *Transportation Science*, *34*(2), 228–238. doi:10.1287/trsc.34.2.228.12306

Bhat, C. R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model. *Transportation Research Part B: Methodological*, *35*(7), 677–693. doi:10.1016/S0191-2615(00)00014-X

Boser, B. E. (1992). A training algorithm for optimal margin classifiers. In The workshop on computational learning theory (pp. 144–152). ACM.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. doi:10.1007/BF00994018

Cottrill, C. D., Pereira, F. C., Zhao, F., Dias, I. F., Lim, H. B., Ben-Akiva, M. E., & Zegras, P. C. (2013). Future mobility survey: Experience in developing a smartphone-based travel survey in Singapore. *Transportation Research Record: Journal of the Transportation Research Board*, *2354*(1), 59–67. doi:10.3141/2354-07

Faroqi, H., Mesbah, M., & Kim, J. (2017). Spatial-temporal similarity correlation between public transit passengers using smart card data. *Journal of Advanced Transportation*, *2017*, 1–15. doi:10.1155/2017/1318945

Feng, Z., Li, X., & Zhang, Q. (2017). Proactive radio resource optimization with margin prediction: A data mining approach. *IEEE Transactions on Vehicular Technology*, *66*(10), 9050–9060.

Ferri-Garcia, R., Fernandez-Luna, J. M., Rodriguez-Lopez, C., & Chilon, P. (2019). Data mining techniques to analyze the factors influencing active commuting to school. *International Journal of Sustainable Transportation*, 1–16. doi:10.1080/15568318.2018.1547465

Frank, L., Bradley, M., Kavage, S., Chapman, J., & Lawton, T. K. (2007). Urban form, travel time, and cost relationships with tour complexity and mode choice. *Transportation*, *35*(1), 37–54. doi:10.1007/s11116-007-9136-6

Gönül, F., & Srinivasan, K. (1993). Modeling multiple sources of heterogeneity in multinomial logit models: Methodological and managerial issues. *Marketing Science*, *64*(4), 138–143. doi:10.1287/mksc.12.3.213

Hagenauer, J., & Helbich, M. (2017). A comparative study of machine learning classifiers for modeling travel mode choice. *Expert Systems with Applications*, *78*, 273–282. doi: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. In *IEEE International Conference on Software Engineering and Service Science* (pp. 219–224). IEEE.

Hensher, D. A., & Greene, W. H. (2003). The mixed logit model: The state of practice. *Transportation*, *30*(2), 133–176.

Kelley, K. (2007). Sample size planning for the coefficient of variation from the accuracy in parameter estimation approach. *Behavior Research Methods*, *39*(4), 755–766. doi:10.3758/BF03192966

Kenny, D. A., & Judd, C. M. (2019). Unappreciated heterogeneity of effect sizes: Implications for power, precision, planning of research, and replication. Retrieved from http://davidakenny.net/doc/KJ17R.pdf.

Kim, J., & Mahmassani, H. S. (2015). Spatial and temporal characterization of travel patterns in a traffic network using vehicle trajectories. *Transportation Research Part C: Emerging Technologies*, *59*, 375–390. doi:10.1016/j.trc.2015.07.010

Kim, J., Schmocker, J. D., & Fujii, S. (2016). Expolring the relationship between undergraduate education and sustainable transport attitudes. *International Journal of Sustainable Transportation*, *10*(4), 385–392. doi:10.1080/15568318.2014.961108

Kumarbarai, S. (2003). Data mining applications in transportation engineering. *Transport*, *18*(5), 216–223. doi:10.3846/16483840.2003.10414100

Lee, W.-H., Tseng, S.-S., Shieh, J.-L., & Chen, H.-H. (2011). Discovering traffic bottlenecks in an urban network by spatiotemporal data mining on location-based services. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1047–1056. doi:10.1109/TITS.2011.2144586

Liu, F., Janssens, D., Wets, G., & Cools, M. (2013). Annotating mobile phone location data with activity purposes using machine learning algorithms. *Expert Systems with Applications*, *40*(8), 3299–3311. doi:10.1016/j.eswa.2012.12.100

Mallett, W. J. (2001). Long-distance travel by low-income households. TRB Transportation Research Circular E-C026—Personal travel: The long and short of it, 169–177.

Manyika, J., Chui, M., & Brown, B. (2011). *Big data: The next frontier for innovation, competition, and productivity*. McKinsey Global Institute.

Mauro, A. D., Greco, M., & Grimaldi, M. (2016). A formal definition of big data based on its essential features. *Library Review*, *65*(3), 122–135. doi:10.1108/LR-06-2015-0061

Mcfadden, D. (1974). Conditional logit analysis of qualitative choice behavior. Academic Press, New York: *Frontiers in Econometrics*, 105–142.

Mohanty, S., & Blanchard, S. (2016). Complete transit: Evaluating walking and biking to transit using a mixed logit mode choice model. In *Transportation Research Board Annual Meeting*.

Omrani, H. (2015). Predicting travel mode of individuals by machine learning. *Transportation Research Procedia*, *10*, 840–849. doi:10.1016/j.trpro.2015.09.037

Potoglou, D., & Kanaroglou, P. S. (2008). Disaggregate demand analyses for conventional and alternative fueled automobiles: A review. *International Journal of Sustainable Transportation*, *2*(4), 234–259. doi:10.1080/15568310701230398

Pye, S., & Daly, H. (2015). Modelling sustainable urban travel in a whole systems energy model. *Applied Energy*, *159*, 97–107. doi:10.1016/j.apenergy.2015.08.127

Rasouli, S., & Timmermans, H. J. P. (2014). Using ensembles of decision trees to predict transport mode choice decisions: Effects on predictive success and uncertainty estimates. *European Journal of Transport & Infrastructure Research*, *14*(4), 412–424.

Scheiner, J., & Holz-Rau, C. (2007). Travel mode choice: Affected by objective or subjective determinants? *Transportation*, *34*(4), 487–511. doi:10.1007/s11116-007-9112-1

Shafique, M. A., & Hato, E. (2015). Use of acceleration data for transportation mode prediction. *Transportation*, *42*(1), 163–188. doi:10.1007/s11116-014-9541-6

Srinivasan, K. K., & Mahmassani, H. S. (2005). A dynamic kernel logit model for the analysis of longitudinal discrete choice data: Properties and computational assessment. *Transportation Science*, *39*(2), 160–181. doi:10.1287/trsc.1040.0093

Taylor, R. (1990). Interpretation of the correlation coefficient: A basic review. *Journal of Diagnostic Medical Sonography*, *6*(1), 35–39. doi:10.1177/875647939000600106

Wang, Y. M., & Elhag, T. M. S. (2007). A comparison of neural network, evidential reasoning and multiple regression analysis in modelling bridge risks. *Expert Systems with Applications*, *32*(2), 336–348. doi:10.1016/j.eswa.2005.11.029

Xie, C., & Lu, J. (2003). Work travel mode choice modeling with data mining decision trees and neural networks. *Journal of the Transportation Research Board*, *1854*(1), 50–61.

Yamamoto, T., Kitamura, R., & Fujii, J. (2002). Drivers' route choice behavior—Analysis by data mining algorithms. *Journal of the Transportation Research Board*, *1807*(1), 59–66.

Yan, H., & Long, D. (2016). Research on key technology for data storage in smart community based on big data. In *International Conference on Intelligent Transportation, Big Data and Smart City* (pp. 653–656). IEEE.

Yang, H., Kitamura, R., Jovanis, P. P., Vaughn, K. M., & Abdel-Aty, M. A. (1993). Exploration of route choice behavior with advanced traveler information using neural network concepts. *Transportation*, *20*(2), 199–223. doi:10.1007/BF01307059

Zang, W., Zhang, P., Zhou, C., & Guo, L. (2014). Comparative study between incremental and ensemble learning on data streams: Case study. *Journal of Big Data*, *1*(1), 1–16. doi:10.1186/2196-1115-1-5

Zhang, J., Wang, F.-Y., Wang, K., Lin, W.-H., Xu, X., & Chen, C. (2011). Data-driven intelligent transportation systems: A survey. *IEEE Transactions on Intelligent Transportation Systems*, *12*(4), 1624–1639. doi:10.1109/TITS.2011.2158001

Zhang, Y., & Xie, Y. (2008). Travel mode choice modeling with support vector machines. *Transportation Research Record: Journal of the Transportation Research Board*, *2076*(1), 141–150. doi:10.3141/2076-16

Zhao, F., Pereira, F. C., Ball, R., Kim, Y., Han, Y., Zegras, C., & Ben-Akiva, M. (2015). Exploratory analysis of a smartphone-based travel survey in Singapore. *Transportation Research Record: Journal of the Transportation Research Board*, *2494*(1), 45–56. doi:10.3141/2494-06